THREE ESSAYS ON EMPIRICAL FINANCE

by

Ruoyang Wang

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Business Administration

Charlotte

2013

Approved by:

____

Dr. Steven P. Clark

____

Dr. Christopher M. Kirby

____

Dr. I-Hsuan Ethan Chiang

____

Dr. Zongwu Cai

ABSTRACT

RUOYANG WANG. Three essays on empirical finance. (Under the direction of DR. STEVEN P. CLARK)

This dissertation includes 3 papers in empirical finance.

In chapter 1, since theory suggests a relationship between both volatility of volatility, variance risk premium, and the equity risk premium; we empirically investigate the relationship between volatility of volatility and the equity risk premium, and the relationship between the variance risk premium and the equity risk premium; we find that volatility of volatility alone explains 5 to 10% of the total variation of equity risk premium, and together with VIX data, it explains more than 20% of the total variation of equity premium; and we fail to find a significant relationship between volatility of volatility and the variance risk premium; we also use six measures of volatility of volatility based on non-parametric models, a GARCH model and VVIX data.

In chapter 2, we proposes a new way to measure the variance risk premium by applying a fractional cointegration relationship between implied variance and realized variance. To find the fractional cointegration coefficient between implied variance and realized variance, we develop a search method based on minimization of the score test statistic proposed by Robinson(1994). We use daily, weekly and monthly data of five stock market indexes (S&P500, S&P100, DJIA, NASDAQ100 and Russell2000) and their volatility indexes from the CBOE. We find our new measure improves the return prediction power of the variance risk premium both in-sample statically and out-of-

sample dynamically, and the result is robust for the monthly data among all five indexes.

In chapter 3, by using submortgage data, we found that investors are being charged with a significant risk premium over owner occupants; besides that, they are also facing a more restricted loan; with the market getting hotter, this risk premium and restrictions are getting even worse. Being treated like that, our findings show that investors were actually not more risky than owner occupants in terms of both prepayment and default. We suspect the reason for this puzzle is that when the market getting hotter, there are more speculative investors who commit occupancy fraud to get a more favorable loan. And these speculative investors were actually recorded as owner occupants on loan documents, which increased our estimation of the hazard of owner occupants group. And our information asymmetry test actually reaffirmed our suspect. Therefore, this paper, for the first time, give statistical evidence on occupancy fraud, and we also proposed a statistical scanning way to reduce to potential occupancy fraud.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: VOLATILITY OF VOLATILITY, EXPECTED STOCK RETURN AND VARIANCE RISK PREMIUM

Theory suggests a relationship between both volatility of volatility, variance risk premium, and the equity risk premium. We empirically investigate the relationship between volatility of volatility and the equity risk premium, and the relationship between the variance risk premium and the equity risk premium. We find that volatility of volatility alone explains 5 to 10% of the total variation of equity risk premium, and together with VIX data, it explains more than 20% of the total variation of equity premium. We fail to find a significant relationship between volatility of volatility and the variance risk premium. We use six measures of volatility of volatility based on non-parametric models, a GARCH model and VVIX data.

1.1 Introduction

There are always gaps between the expectations of market participants and subsequent reality. What insights can be gleaned from studying the differences between expected volatility implied by option prices and *ex post* realized volatility of the time series of prices of the underlying security? (See, Christensen and Prabhala (1998) and Demeterfi et al (1998).) When the implied volatility is higher than the realized volatility, it means traders anticipated certain risks and usually these risks could be hedged by trading financial derivatives. However, when the implied volatility is smaller than the realized volatility , it means on the market there are some risk that has not been expected by investors and that's when some extreme events happens. Usually, crash will happen when the implied volatility is smaller than the realized volatility.

However, due to the measurement error of implied volatility and realized volatility, the idea of variance risk premium is only theoretical, it's hard to quantify the variance risk premium to predict extreme cases(see, Chernov(2007) and Carr and Wu (2009)). However, from both the equilibrium model and probabilistic model, people can prove that this variance risk premium primarily (if not solely) depends on the volatility of volatility. Therefore, the measurement of volatility of volatility becomes a crucial variable to measure the risk of extreme cases. It's easier to understand this from a risk management perspective: for example, nowadays, Basel II require banks to calculate Value at Risk every day based on their historical volatility. However, this will only help to prevent some common risk. To prevent some extreme cases, people not only need to know the historical volatility, but also the volatility of volatility, or the probability that this historical volatility itself is going to change. that's why volatility of volatility is a crucial variable to measure the potential extreme risk(see Jullizrd and Ghosh(2012) and Liu, Pan and Wang(2005)).

But so far, there is no literature measuring volatility of volatility, although lots of literature mentioned its existence(see, Jones(2003) and Bollerslev and Todorov (2011)). The reason why so far it's not in the literature might because in a equilibrium model, if one assume the consumption growth rate is constant, or the volatility of consumption growth rate is constant, then when solving the equilibrium model, there will be no volatility of volatility term. And also in a probabilistic model, if one assume the return is normal distributed, there will also be no volatility of volatility term. Only when we set our model to be more realistic, with a changing volatility of consumption growth rate, or allow the return follow a Lévy process, we will recognize the volatility of volatility term.

Usually when people measure volatility, especially in stochastic volatility model, people tend to think the volatility itself is unobservable, so usually people use state space model to measure it. However, due to the importance of volatility of volatility in risk management and derivative pricing, it will be useful if we have explicit measurements of volatility of volatility. This paper proposed 6 different measurements of volatility of volatility based on different ideas. One measurement is an almost non-parametric measurement of volatility of volatility; 3 measurements based on GARCH model: we proposed a new GARCH model for volatility of volatility, which we call nested GARCH model. We also proposed estimation method for this GARCH model. And from the empirical test, the measurement based on this nested GARCH model has the best performance when predicting future return. The other two measurements come from the trading price of options on S&P500 indexes and options on VIX index. Our empirical results show that volatility of volatility, together with VIX, could explain more than 20% of the variation of equity premium and it has some prediction power on future market returns.

Our paper is structured as follows. In Section 2, we discuss the theoretical motivation for this paper. In Section 3, we propose our six measures of volatility of volatility. We describe the data we use in Section 4 and we do the empirical tests in Section 5. Section 6 concludes.

1.2 Motivation

1.2.1 Variance Risk Premium and Volatility of Volatility

We begin by highlighting two models in the literature that imply that the variance risk premium depends on volatility of volatility .Bollerslev et al. (2009) and Drechsler and Yaron (2011) presented equilibrium models in which the variance risk premium depends on the volatility of volatility.

1.2.1.1 General Equilibrium Model

The general equilibrium model developed by Bansel and Yaron (2004), Bollerslev, Tauchen and Zhou (2009) and Dressler and Yaron (2011) all start from the geometric growth rate of consumption, which is assumed to be

$$g_{t+1} = \mu_g + \sigma_{g,t} z_{g,t+1}$$

The volatility of this growth rate is stochastic, following

$$\sigma^2_{g,t+1} = a_\sigma + \rho_\sigma \sigma^2_{g,t} + \sqrt{q_t} z_{\sigma,t+1}$$

$$q_{t+1} = a_q + \rho_q q_t + \phi_q \sqrt{q_t} z_{g,t+1}$$

where $a_\sigma > 0$, $a_q > 0$, $|\rho_\sigma| < 1$, $|\rho_q| < 1$, $\phi_q > 0$, $\{z_{g,t}\}$, $\{z_{\sigma,t}\}$ and $\{z_{q,t}\}$ are i.i.d. series with mean zero and unit variance.

They assume the representative agent has recursive preferences as described in Epstein and Zin (1989) and Weil (1989),

$$V_t = [(1-\delta) C_t^{\frac{1-\gamma}{\theta}} + \delta (E_t[V_{t+1}^{1-\gamma}])^{\frac{1}{\theta}}]^{\frac{\theta}{1-\gamma}}$$

then using the approximation by Campbell and Shiller (1988), solve this model for an equilibrium,

$$r_{t+1} = \kappa_0 + \kappa_1 \omega_{t+1} + \omega_t + g_{t+1}$$

where r_{t+1} is the logarithm return of any consumption asset from time t to $t+1$, ω_t is the logarithm of the price-consumption ratio and κ_0 and κ_1 are coefficients for approximation.

By assuming the solution follows an affine form of these two state variables $\sigma_{g,t}$ and q_t,

$$\omega_t = A_0 + A_\sigma \sigma_{g,t}^2 + A_q q_t$$

by solving the Euler equation from the standard asset pricing condition $E_t[M_{t+1}R_{i,t+1}] = 1$ where the logarithm of the inter-temporal marginal rate of substitution should be the form

$$m_{t+1} = \theta \log \delta - \theta \psi^{-1} g_{t+1} + (\theta - 1) r_{t+1}$$

where δ refers to the subjective discount factor.

Since the Euler condition must hold for all values of the state variables, we can solve the 3 coefficients:

$$A_0 = \frac{\log \delta + (1 - \psi^{-1})\mu_g + \kappa_0 + \kappa_1 [A_\sigma a_\sigma + A_q a_q]}{1 - \kappa_1}$$

$$A_\sigma = \frac{(1-\gamma)^2}{2\theta(1 - \kappa_1 \rho_\sigma)}$$

$$A_q = \frac{1 - \kappa_1 \rho_q - \sqrt{(1 - \kappa_1 \rho_\sigma)^2 - \theta^2 \kappa_1^4 \phi_q^2 A_\sigma^2}}{\theta \kappa_1^2 \phi_q^2}$$

Based on the solution of this model [1], we will get the relationship between variance

[1] In the appendix

risk premium and volatility of volatility

$$VRP \equiv E_t^Q(\sigma_{r,t+1}^2) - E_t(\sigma_{r,t+1}^2) = (\theta - 1)\kappa_1[A_\sigma + A_q\kappa_1^2(A_\sigma^2 + A_q^2\phi_q^2)\phi_q^2]q_t$$

where

$$\theta \equiv (1-\gamma)(1-\psi^{-1})^{-1}$$

Here γ denotes the coefficient of risk aversion and ψ refers to the inter-temporal elasticity of substitution, usually both of them are bigger than 1, which makes $\theta < 0$. And the expression of coefficient A_σ and A_q is in the appendix. Since both of them are negative, we can see the variance risk premium is positively correlated with the volatility of volatility, and we will test this in Section 5.

1.2.1.2 Probabilistic Model

Barndorff-Nielsen and Veraart (2012) proposed the volatility modulated non-Gaussian Ornstein-Uhlenbeck (VMOU) processes and quantified the impact of the volatility of volatility on the variance risk premium. They assume the volatility process V_t follows

$$dV_t = -\lambda V_t dt + q_t dL_t$$

where L_t is a levy process with characteristic triplet $(\gamma, 0, q)$, which means the characteristic function of L_t $E(exp(i\theta L_t)) = exp(t\Psi_L(\theta))$ satisfy

$$\Psi_L(\theta) = i\theta\gamma + \int_0^\infty (e^{i\theta x} - 1)\nu(dx)$$

They proved that if the volatility of volatility q_t follows

$$dq_t = a(b - q_t)dt + g\sqrt{q_t}dW_t$$

where a, b and g are positive constants satisfying the Feller condition ($2ab > g^2$, $q_0 > 0$) and W is a standard Brownian motion, then the variance risk premium and the volatility of volatility satisfy

$$VRP_{t,t+h} = q_t F_1(h) + F_2(h)$$

where $F_1(h)$ and $F_2(h)$ are explicitly known deterministic functions.

By defining

$$G(h) = -\frac{a+1}{a} + h - \frac{a}{1-a}e^{-h} + \frac{1}{a(1-a)}e^{-ah}$$

and the risk-neutral measure for a and b are a^Q and b^Q, and $G^Q(h) = G(a^Q, h)$, according to Barndorff-Nielsen and Veraart (2012), we have:

$$F_1(h) \equiv (\kappa_1 - \kappa_1^Q)(1 - \frac{1-e^{-h}}{h}) - \frac{1}{h}(\kappa_1 G(h) - \kappa_1^Q G^Q(h))$$
$$F_2(h) \equiv \frac{1}{h}[\kappa_1 G(h) b - \kappa_1^Q G^Q(h) b^Q]$$

where κ_1 is the 1st cummulant of the Levy subordinator L_t, and κ_1^Q is the risk-neutral measure of κ_1.

It is difficult to determine the sign of $F_1(h)$ and $F_2(h)$, but at least we know that they are deterministic functions depends on the frequency of data. Also, we can see, if there is no volatility of volatility, the variance risk premium would be deterministic. Therefore, the probabilistic model proposed by Barndorff-Nielsen and Veraart (2012) also imply the variance risk premium depends on the volatility of volatility, and the coefficient depends on the frequency of data. We will empirically test this in Section 5.

1.2.2 Expected Stock Return and Volatility of Volatility

Returning to the Bollerslev, Tauchen and Zhou (2009) model, if we substitute the solution of ω_t into the approximation

$$r_{t+1} = \kappa_0 + \kappa_1 \omega_{t+1} + \omega_t + g_{t+1}$$

we have

$$r_{t+1} = -\log_{\delta} + \Psi^{-1}\mu_g - \frac{(1-\gamma)^2}{2\theta}\sigma_{g,t}^2 + (\kappa_1\rho_q - 1)A_q q_t + \sigma_{g,t} z_{g,t+1} + \kappa_1\sqrt{q_t}[A_\sigma z_{\sigma,t+1} + A_q\phi_q z_{q,t+1}]$$

So according to their model, the return from t to $t+1$ positively related to the the volatility $\sigma_{g,t}$ and the volatility of volatility q_t, compensating investors for the additional volatility risk.

Then, plug the solution into the logarithm of inter-temporal marginal rate of substitution,

$$m_{t+1} = \theta \log \delta - \theta\psi^{-1} g_{t+1} + (\theta - 1) r_{t+1}$$

one can get the equity premium $r_m - r_f$:

$$\pi_{r,t} \equiv -Cov_t(m_{t+1}, r_{t+1}) = \gamma\sigma_{g,t}^2 + (1-\theta)\kappa_1^2(A_q^2\phi_q^2 + A_\sigma^2)q_t$$

From the above expression, one can see the equity premium can be decomposed into two parts: the first comes from the volatility, which has been well studied by previous literature, and the second comes from the volatility of volatility. Moreover, since θ is smaller than zero, the equity premium is positively correlated with volatility of volatility, or, say, the shock to volatility. We will test this result in Section 5.

1.3 Measurement of Volatility of Volatility

We investigate six possible measures for volatility of volatility based on three different ideas. One is a direct method with minimal assumptions except that returns can be decomposed into a time varying volatility times a normal innovation. The second category is based on GARCH modeling and the third category is based on the option implied volatility. We list the basic idea of each of the measures in table 1, also plot each of them in Figures 2 and 3, and will explain them in this section one by one. From Figures 2 and 3, one can see although the detailed plot of each measurement is different, these six measurements almost peak at the same time around September 2008, this verified our statement about volatility of volatility is good at detecting extreme things.

1.3.1 A Semi parametric Way to Estimate Volatility of Volatility

According to Figure 1 in Anderson et al. (2001) , the returns scaled by realized standard deviations is approximated Gaussian, so we can model the stock return as

$$r = \sigma\epsilon$$

where $\sigma \geq 0$ and $\epsilon \sim N(0,1)$, if we further assume that σ and ϵ are uncorrelated, then we will have the expected return as:

$$E(r) = E(\sigma\epsilon) = E(\sigma)E(\epsilon)$$

and if σ^2 and ϵ^2 are also uncorrelated,

$$E(r^2) = E(\sigma^2)E(\epsilon^2) = E(\sigma^2)$$

and

$$E(|r|) = E(\sigma)E(|\epsilon|) = \sqrt{\frac{2}{\pi}}E(\sigma)$$

since we know that if $\epsilon \sim N(0,1)$, we have $E(\epsilon^2) = 1$, $E(|\epsilon|) = \sqrt{\frac{2}{\pi}}$. Therefore, to estimate $E(\sigma^2)$ and $E(\sigma)$, we can use:

$$E(\sigma^2) = \frac{1}{N}\sum_{i=1}^{N} r_i^2$$

and

$$E(\sigma) = \sqrt{\frac{\pi}{2}} \frac{1}{N} \sum_{i=1}^{N} |r_i|$$

Therefore, to estimate the realized variance of volatility, we use

$$Var(\sigma) = E(\sigma^2) - E^2(\sigma) = \frac{1}{N} \sum_{i=1}^{N} r_i^2 - \frac{\pi}{2} [\frac{1}{N} \sum_{i=1}^{N} |r_i|]^2 \quad (1)$$

Since this measurement cannot guarantee it's always positive, we take the absolute value of this measurement as the variance of volatility.

1.3.2 GARCH type of model to measure VolVol

Here, we proposed three ways to measure volatility of volatility by using GARCH model.

1.3.2.1 the realized volatility of the GARCH volatility

Here, if we assume the daily return follows a simple GARCH(1,1) process as following:

$$\begin{cases} r_t = \mu_r + h_t \epsilon_t \\ h_t^2 = \alpha_0 + \alpha_1 h_{t-1}^2 + \beta_1 u_{t-1}^2 \end{cases}$$

then we could get a estimated volatility $\hat{h}_t$ every day, the realized volatility of this estimated volatility based on GARCH model

$$RV_t(\hat{h}) \equiv \sum_{i=1}^{n} [\hat{h}_{t+\frac{i}{n}\Delta} - \hat{h}_{t+\frac{i-1}{n}\Delta}]^2$$

would be a very straight forward measurement of volatility of volatility. To get a monthly $RV_t(h)$, n is around 22 each month.

1.3.2.2 The volatility of realized volatility

According to Corsi et al. (2006), if we define

$$RV_t \equiv \sum_{i=1}^{n} [p_{t+\frac{i}{n}\Delta} - p_{t+\frac{i-1}{n}\Delta}]^2$$

then the logarithm of realized volatility actually follows a normal distribution with a time-varying variance.

$$\frac{\sqrt{RV_t} - \sqrt{\int_{t-1}^{t} \sigma^2(s)ds}}{\sqrt{\frac{Q_t^*}{2MRV_t}}} \xrightarrow{d} N(0,1)$$

where $\sqrt{\frac{Q_t^*}{2MRV_t}}$ is an approximation of the standard deviation of the realized volatility. And following Corsi et al. (2006), without loss of generality, we can assume that the logarithm of realized volatility actually follows a GARCH (1,1) process as

$$\begin{cases} y_t = \mu_y + \sqrt{h_t}\epsilon_t \\ h_t = \omega + \alpha_1 h_{t-1} + \beta_1 u_{t-1}^2 \end{cases}$$

where y is $\sqrt{RV}$ and $u_{t-1} = \sigma_{t-1}\epsilon_{t-1}$ and $\{\epsilon_t\}$ is a sequence of i.i.d. random variables with mean 0 and variance 1. Since here, we use the daily price to get the monthly realized volatility, then by using the GARCH model on the realized volatility, we actually get the monthly volatility of the realized volatility as $\hat{h}_t$.

1.3.2.3 A nested GARCH model

Here, we assume the return as following

$$r_t = \mu_r + \sigma_t \epsilon_{r,t} \tag{2}$$

where μ_r could be an ARMA process itself, but since we are focusing on volatility and volatility of volatility, for simplicity, we just put it as μ_r, and $\{\epsilon_{q,t}\}$ is a sequence of i.i.d. random variables with mean 0 and variance 1.

For the volatility, we assume part of it is deterministic as a GARCH process, but the other part is stochastic.

$$\sigma_t^2 = \alpha_0 + \alpha_1 \sigma_{t-1}^2 + \beta_1 u_{t-1}^2 + q_t \epsilon_{\sigma,t} \tag{3}$$

where $u_{t-1} = \sigma_{t-1}\epsilon_{r,t-1}$, $\{\epsilon_{\sigma,t}\}$ is a sequence of i.i.d. random variables with mean 0 and variance 1 as well, and it's independent from $\{\epsilon_{q,t}\}$. $\alpha_0 > 0$, $\alpha_1 \geq 0$, $\beta_1 \geq 0$ and $\alpha_1 + \beta_1 < 1$. Here, q_t is the volatility of volatility and we assume this part also follow a GARCH process

$$q_t^2 = \alpha_q + \rho_q q_{t-1}^2 + \phi_q \eta_{t-1}^2 \tag{4}$$

where $\eta_{t-1} = q_{t-1}\epsilon_{\sigma,t-1}$, $\alpha_q > 0$, $\rho_q \geq 0$, $\phi_q \geq 0$ and $\rho_q + \phi_q < 1$.

If we put them together, it would be

$$\begin{cases} r_t = \mu_r + \sigma_t \epsilon_{r,t} \\ \sigma_t^2 = \alpha_0 + \alpha_1 \sigma_{t-1}^2 + \beta_1 u_{t-1}^2 + q_t \epsilon_{\sigma,t} \\ q_t^2 = \alpha_q + \rho_q q_{t-1}^2 + \phi_q \eta_{t-1}^2 \end{cases}$$

The 3-step estimation method for this "nested" GARCH model is as following:

First, by ignoring any ARCH effects, we can estimate the mean equation μ_t of a return series by using MLE; denote the residual series by $\hat{u}_t$, we get

$$\hat{u}_t = r_t - \hat{\mu}_{r,t}$$

For the second step, we can treat $\{\hat{u}_t^2\}$ as an observed time series, and denote $\xi_t \equiv u_t^2 - \sigma_t^2$, and plug it into equation (2), we get

$$u_t^2 = \alpha_0 + (\alpha_1 + \beta_1)u_{t-1}^2 + \xi_t - \alpha_1 \xi_{t-1} + q_t \epsilon_{\sigma,t}$$

So we can treat $\{u_t^2\}$ as a ARMA process, say, the AR coefficient estimate is $\hat{\theta}_1$ and for MA is $\hat{\phi}_1$, then $\hat{\beta}_1 = \hat{\theta}_1 - \hat{\phi}_1$ and $\hat{\alpha}_1 = \hat{\phi}_1$. Since it's easy to check that $E(\xi_t) = 0$ and $cov(\xi_t, \xi_{t-s}) = 0$ for $s \geq 1$, so $\{\xi_t\}$ is a martingale difference series, therefore the estimation of α_0, α_1 and β_1 is unbiased.

The third step is very similar to the second step, again, we can denote the residual series by $\hat{\eta}_t$ and treat $\hat{\eta}_t^2$ as an observed time series, define $m_t \equiv \eta_t^2 - q_t^2$, plug it into equation (3), we get

$$\eta_t^2 = \alpha_q + (\rho_q + \phi_q)\eta_{t-1}^2 + m_t - \rho_q m_{t-1}$$

so we can treat η_t^2 as a ARMA process and get unbiased estimation of α_q, ρ_q and ϕ_q.

Although both ξ_t and m_t are martingale martingale difference series, so we can get unbiased estimation for our model, in general, they are not i.i.d. sequences. There we will inevitably lose some efficiency by this 3-step estimation.

The $\hat{q}_t$ would be a daily measurement of volatility of volatility, but since here we care about the monthly measurement of volatility of volatility, so we only use the $\hat{q}_t$ at the end of each month.

1.3.3 Volatility of VIX

Since VIX is considered to be a barometer of market volatility, the volatility of VIX naturally becomes a measure of volatility of volatility [2]. So there are two ways to measure the volatility of VIX: the realized volatility of VIX and the implied volatility of VIX.

[2]Chicago Board of Exchange explained the volatility of volatility by using VIX at `http://www.cboe.com/micro/vix/VIXoptionsFAQ.aspx\#9`

1.3.3.1 The realized volatility of VIX

Since VIX is an index for volatility, it is straight forward to use the realized volatility of VIX as a measures of volatility of volatility

$$RV_t(VIX) \equiv \sum_{i=1}^{n}[VIX_{t+\frac{i}{n}\Delta} - VIX_{t+\frac{i-1}{n}\Delta}]^2$$

1.3.3.2 The implied volatility of VIX

When CBOE first launched options written on VIX in 2006, people were worrying about the liquidity of these options, we plot the trading volume of options on VIX in Figure 2, the trading volume of the options on VIX has increased exponentially from several thousand contracts per day in 2006 to more than 1 million contract in 2013 [3]. With such a big volume of options trading everyday, the volatility of VIX itself is crucial in pricing these options. To give the market an opportunity to see the implied volatility from the trading of these options, CBOE launched VVIX in March 2012, the method to determine the option implied volatility of VIX is similar to what they used to determine VIX, the implied volatility of options on SP500. So the VVIX is a natural measure of volatility of volatility.

1.4 Data and Measurement

1.4.1 Realized Volatility

Barndorff-Nielsen and Shephard (2002) proved that when $n \to \infty$ and Δ is fixed, (using the same notation as Bollerslev et al. (2009),) the realized volatility measured

[3] see `http://ir.cboe.com/releasedetail.cfm?ReleaseID=756850`

by the following

$$E_t^P(\sigma_{r,t+1}^2) \equiv RV_t \equiv \sum_{i=1}^{n}[p_{t+\frac{i}{n}\Delta} - p_{t+\frac{i-1}{n}\Delta}]^2$$

will be a good approximation of the unobserved integrated volatility, where p_t is the logarithmic price.

Here, we used the daily data to get the monthly realized volatility, so Δ is one month, and n is around 22 for each month[4]. The data is from January 1st, 1990 to December 31st, 2012. The reason it start from January 1st, 1990 is because that's the earliest data we can get from CBOE for the VIX which we will use later.

We plot this monthly realized volatility time series on the second chart of Figure 4, comparing with Figure 2 in Bollerslev et al. (2009), we get very similar result from Jan 1990 to December 2007, but the financial crisis drove the realized volatility around 2009 much higher than before. We list the 10 months with the highest realized volatility from Jan 1990 to Dec 2012 in table 1, 7 out of these 10 highest realized volatility months is the 7 consecutive months from Sep 2008 to Mar 2009.

1.4.2 Implied volatility

Just like the "implied" interest rate can be extracted from bond price, implied volatility could also be extracted from option price. Historically, people used the famous Black-Scholes formula to get the implied volatility from option price, but it's based on the Gaussian distribution and only incorporated one strike price; to incorporate more information from option prices with different striking prices, Carr and Madan (1998) and Demeterfi et al. (1999) proposed a "model-free" way to incorpo-

[4]the reason we didn't use the 5 minute data is because according to the Table 1 in Drechsler and Yaron (2010), the "nontrivial auto-correlation" in the five-minute returns tend to drive the mean of realized volatility based on 5 minute data much smaller than that for the daily data.

rate prices of both call options and put options with different striking prices but the same expiration date, but because of the put-call parity and sometimes the illiquidity of put options, the following way, which is proved by Britten-Jones and Neuberger (2000) to be the risk neutral expectation of the integrated volatility, become a popular way in the market to calculate implied volatility.

$$E_t^Q(\sigma_{r,t+1}^2) \equiv IV \equiv 2\int_0^\infty \frac{C_t(t+1, \frac{K}{B(t,t+1)}) - C_t(t,K)}{K^2} dK$$

where $C_t(T,K)$ denote the price of a European call option at time t with the strike price at K and Maturity at T, $\frac{1}{B(t,t+1)}$ is the discount rate from t to $t+1$. And since 2003, CBOE start using this "model-free" measurement for the VIX index, which is based on S&P 500 index options with a 30 day maturity.

Here, to plot the monthly implied volatility, we used the last observation of each month of the VIX data, and plot it in the first chart of figure 4. We can see although the implied volatility was very high during the financial crisis, but it's not as high as the realized volatility during that time.

We also listed the highest 10 months with the highest implied volatility in the same table, from which we can see that there is a lead-lag relationship between realized volatility and implied volatility as documented by other literature, that if one month get a high realized volatility, then people will trade the options with a higher implied volatility the same month or the next month, which drive the implied volatility from the same month or the next month higher.

1.4.3 Variance Risk Premium

Following Bollerslev, Tauchen and Zhou (2009) and Drechsler and Yaron (2011), We define the variance Risk Premium as the difference between the Implied Variance and the Realized Variance:

$$VRP \equiv E_t^Q(\sigma^2_{r,t+1}) - E_t(\sigma^2_{r,t+1}) = IV - RV$$

Most of the time, options are traded with an implied volatilities higher than the realized volatilities, which means usually, the realized volatility has been expected, so the implied volatility subtracted from option price are usually higher than the realized volatility, so the variance risk premium is usually positive, and the higher the variance risk premium, the higher the implied volatility than the realized volatility, it means the more cautious people are, or the more risk averse people are during that period. However, when the realized volatility is bigger than the implied volatility, that's when there are uncovered, or unpredicted risk, then at these extreme cases, the variance risk premium become negative. We used our daily data to get the realized variance and the VIX data to get the implied variance; and then get the difference between these two as the variance risk premium, and plot it in Figure 1, and we get the similar results as Bollerslev, Tauchen and Zhou (2009) [5].

1.4.4 Data Description

We select the daily data of S&P500 index to do empirical test, there are three major reasons we chose S&P500 index, one is it has a good coverage of the entire market, which means it's a good representative for the market; the second reason is

[5] We appreciate that the authors of this paper shared the data:
`http://www.federalreserve.gov/econresdata/researchdata/feds200711.xls`

the options based on S&P500 has good liquidity, and the implied volatility based on these options are calculated everyday as VIX, and every the options based on VIX are also liquid now, so it's easy get data for us to verify our thoughts on volatility of volatility; the third reason is that S&P500 index data has been used by many scholars before us, so it's easy for us to compare our results with previous literature.

We have four different data sets here: Data Set 1 is from January 1st, 1990 to September 30th, 2012, that's the longest data we can get so far, but since the authors of Bollerslev, Tauchen and Zhou(2009) only updated their realized volatility measure by 5-min high-frequency data to December 2008, for data set 1, we used our own measurement of realized volatility, which is based on daily data. For robustness check, we have data set 2, which is from January 1990 to December 2008, with the Realized volatility data updated by the authors of Bollerslev, Tauchen and Zhou(2009). To compare our results with Bollerslev, Tauchen and Zhou(2009), we have data set 3, which is from January 1990 to December 2007, and also the other reason for data 3 is to see if we get rid of the peak data, will that affect our results. To incorporate Measurement 6 based on the VVIX data, which was launched January 1st, 2007, we use data set 4, which is from January 1st, 2007 to September 30th, 2012, the other reason for checking this data set is to see if our model works with the most recent data. To summarize our data sets, we have table 2.

Besides all 6 measurements of variance of volatility, realized variance, implied variance and the variance risk premium that we've discussed earlier in this section and the previous section, we also used the excess return, defined by the difference between the market return and the risk free rate, we used Fama-French data [6]. We also used other popular predicting variables from previous literature (Lamont(1998), Lettau and Ludvigson (2001), Ang and Bekaert (2007) and Bollerslev, Tauchen and Zhou(2009)): the price-earning ratio and the price-dividend ratio of S&P500 index comes from public data set [7]; the default spread (DFSP), defined as the difference between Moody's BAA and AAA Bond Yield Indices, the term spread (TMSP), defined as the difference between the ten year and three month treasury yields, and the stochastically detrended risk free rate (RREL), defined as the difference between the one month T-bill rate and its trailing twelve month averages, come from the Federal reserve website [8]. The consumption-wealth ratio CAY comes from the Lettau's website [9].

We listed basic summary statistics for the monthly returns and predictor variable for data set 1 in Table 3. For robustness check, we also listed the summary statistics for data set 2, 3, 4 in Table 4, 5 and 6.

1.5 Empirical Results

1.5.1 Expected Stock Return and Volatility of Volatility

With the equilibrium theory developed by Bansel and Yaron (2004), Bollerslev, Tauchen and Zhou (2009) and Dressler and Yaron (2011), at the existence of volatility

[6] We do appreciate their effort on updating the data `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/ftp/F-F_Research_Data_Factors.zip`

[7] `http://www.multpl.com/table?f=m` and `http://www.multpl.com/s-p-500-dividend-yield/table?f=m`

[8] `http://www.federalreserve.gov/releases/h15/data.htm`

[9] `http://faculty.haas.berkeley.edu/lettau/data/cay_q_12Q3.txt`

of volatility, the equity premium should follow

$$\pi_{r,t} \equiv -Cov_t(m_{t+1}, r_{t+1}) = \gamma\sigma_{g,t}^2 + (1-\theta)\kappa_1^2(A_q^2\phi_q^2 + A_\sigma^2)q_t$$

To test the relationship between the equity premium and the volatility of volatility, we run the simple regression for all the measurements of volatility of volatility first, and then also other predictor variables, such as the realized variance, implied variance, the variance risk premium, and other macro-economic predictor variables. We listed the result in Table 7, from which we can see that measurement 1, 2 and 5 perform really good in the simple regression context, with the R^2 being 9%, 5% and 11% respectively. In the multiple regression context, since the implied variance (IV) is significant in the simple regression context as well, we put IV as the additional explanatory variable. then the R^2 is more than 20%.

The coefficient before volatility of volatility is negative in the simple regression context, however, after we add the implied volatility as the additional explanatory variable, the coefficient before volatility of volatility becomes positive and the coefficient before IV is negative. Since we are testing the relationship between the equity premium and volatility of volatility, Implied Volatility is a good approximation of volatility itself. So we can see there is a positive variance risk premium due to the positive coefficient in front of volatility of volatility. However, the negative coefficient in front of implied volatility could more likely be the risk-return tradeoff.

For robustness check, we run the same regression on the other three data sets, and the results are in table 8, 9 and 10. From which we can see that although the performance varies a little bit with different data sets, for example, with measurement 4 of volatility of volatility and implied volatility, the R^2 could get to as high as 33% for data set 2, which means 33% of variation in equity premium could be explained by just these two variables, and for data set 4, these two variables together with the realized volatility could explain 42% of the total variation of equity premium. Although the performance for data set 3 is not as good as other data sets, the result is still quite robust for all these 4 data sets: measurement 1, 2 and 5 for volatility of volatility always perform good in the simple regression context, and the measurement 2 or 4 of volatility of volatility, together with the implied volatility and realized volatility, could explain more than 20% of the total variation of equity premium.

Here, the negative coefficient in front of the implied variance term and the positive coefficient in front of the variance of volatility term also reflect that people are aware of the volatility of the market: when the volatility is low, people tend to get into the market, which drives the return to be higher, and when the market volatility is high, people actually get scared by the high volatility of the market, so people tend to leave the market, which in turn drives the return lower. However, most people are not aware of the volatility of volatility, or the risk of the extreme cases, so even after considering the volatility, there are still some risk left uncovered for some extreme cases, and because people's unawareness of this risk, the returns are even higher when this uncovered risk is higher, when this effect accumulate to some extent, some extreme event, like crash, will happen.

1.5.2 Return forecasting

To see whether our measurements of volatility of volatility have some predicting power on stock market return, we run simple regression to predict return of different horizons from 1 month to 24 months by using 1-month lagged volatility of volatility measurement. We tried all of our 5 measurements, we find although the measurements 1, 2 and 5 perform good in explaining the equity premium at the same period, they don't really have a lot of predicting power in the long run. We put the predicting market return result by measurement 4 in table 15 and 16, respectively. We also plot the estimated slope coefficient and the 95% confidence band for the estimated slope coefficient, and also the adjusted R^2 in Figure 5.

From the results we can see that for measurement 3, with the increase of the predicting horizon, the predicting power is getting stronger from less than 1 percent in 5 months, to more than 4 percent in 2 years. For measurement 4 of volatility of volatility, the prediction get best when the predicting horizon is from 9 months to 16 months, the prediction power could get to as high as 6%, which is a good result even comparing with previous literature.

1.5.3 Variance Risk Premium and Volatility of Volatility

To test the relationship between variance risk premium and the volatility of volatility predicted by both the equilibrium model and probabilistic model in section 2 of this paper, we run regression between the variance risk premium and 5 different measurements of volatility of volatility for data sets 1, 2 and 3, and we put the results in table 17, 18 and 19, respectively.

We get very mixed results: for data set 1, the relationship is not significant at all, but for data set 2 and 3, the relationship is significant, especially for measurement 2, 3, 4 and 5; but the sign of those coefficients in front of volatility of volatility measures changed from one data set to another. We think the reason for this is because for data set 1, the variance risk premium is measured by daily data instead of high frequency data, but for data set 2 and 3, the variance risk premium is measured by the 5-min high frequency data, this might cause the different result between data set 1 and data set 2,3. The other reason is the data for the whole year 2008 is very different from previous data due to the 2008 financial crisis. So by including the data in 2008, data set 2 perform differently than data set 3, this could be the reason why the sign of coefficients changed from data set 2 to data set 3. To test this relationship, we need data with longer horizon and high frequency data to measure realized volatility.

1.6 Conclusion

From the empirical tests we can see that volatility of volatility itself could explain 5-10% variation of equity premium, together with VIX, these two variables could always explain more than 20% of variation of equity premium, this result is robust through all 4 of our data sets. The volatility of volatility also have some predicting power on the future returns, especially good at one year around. To test the relationship between the volatility of volatility and variance risk premium, we need high frequency data and longer time series.

All in all, with the 6 different measurements of volatility of volatility, we provide empirical evidence that volatility of volatility itself could be a good measure of the risk of extreme cases and it has predicting power on future returns and could explain a fair amount of equity premium in the same period.

Table 1: 6 ways to measure volatility of volatility

Category		Model	Measurement of Volatility of volatility
Nonparametric	1		$Var(\sigma) = E(\sigma^2) - E^2(\sigma) = \frac{1}{N}\sum_{i=1}^{N} r_i^2 - \frac{\pi}{2}[\frac{1}{N}\sum_{i=1}^{N}\lvert r_i\rvert]^2$
GARCH	2	$\begin{cases} r_t = \mu_r + h_t\epsilon_t \\ h_t^2 = \alpha_0 + \alpha_1 h_{t-1}^2 + \beta_1 u_{t-1}^2 \end{cases}$	$RV_t(\hat{h}) \equiv \sum_{i=1}^{n}[\hat{h}_{t+\frac{i}{n}\Delta} - \hat{h}_{t+\frac{i-1}{n}\Delta}]^2$
	3	$\begin{cases} \sqrt{RV_t} = \mu + \sqrt{h_t}\epsilon_t \\ h_t = \omega + \alpha_1 h_{t-1} + \beta_1 u_{t-1}^2 \end{cases}$	$\hat{h}_t$
	4	$\begin{cases} r_t = \mu_r + \sigma_t\epsilon_{r,t} \\ \sigma_t^2 = \alpha_0 + \alpha_1\sigma_{t-1}^2 + \beta_1 u_{t-1}^2 + q_t\epsilon_{\sigma,t} \\ q_t^2 = \alpha_q + \rho_q q_{t-1}^2 + \phi_q\eta_{t-1}^2 \end{cases}$	$\hat{q}_t$
VIX	5	Realized Volatility of VIX	$RV_t(VIX) \equiv \sum_{i=1}^{n}[VIX_{t+\frac{i}{n}\Delta} - VIX_{t+\frac{i-1}{n}\Delta}]^2$
	6	Implied Volatility of VIX	VVIX Index

Table 2: 4 different data set we are using

	From	To	Realized Volatility Measurement	Reason
Data Set 1	Jan, 1990	Sep, 2012	Daily Data	Longest Data Set we can get so far
Data Set 2	Jan, 1990	Dec, 2008	High Frequency Data	Used updated Bollerslev, Tauchen and Zhou(2009)'s data
Data Set 3	Jan, 1990	Dec, 2007	High Frequency Data	Compare results with Bollerslev, Tauchen and Zhou(2009)
Data Set 4	Jan, 2007	Sep, 2012	Daily Data	Measurement 6 VVIX start from Jan 2007

Table 3: Summary statistics for data set 1

	Mkt-RF	q1	q2	q3	q4	q5	IV	RV	VRP	logPEratio	logPDratio	DFSP	TMSP	RREL	CAY
Mean	0.532015	13.35434	28.9161	1.763095	64.26596	4.973536	38.83747	167.7407	10.82902	3.127702	3.911809	0.968645	1.885604	-0.18665	0.001996
Std.Dev	4.486457	28.34475	43.12624	2.057753	59.03446	10.72358	51.42978	393.374	28.85035	0.391093	0.309382	0.426656	1.165713	0.781711	0.023783
Skewness	-0.6396	6.499788	6.066847	6.167164	2.559036	6.123384	6.629029	4.00717	-5.002	1.916618	-0.14294	3.046807	-0.15971	-0.44842	-0.0616
Kurtosis	1.042247	58.69305	45.72772	49.62393	9.113538	48.59	58.15698	19.53079	68.90016	5.153592	-0.80994	12.09571	-1.15783	0.24503	-1.46293
AR(1)	0.095041	0.433066	0.777636	0.7431	0.890174	0.611938	0.72024	0.882322	-0.19963	0.977047	0.987555	0.963148	0.975875	0.971465	0.979051
Mkt-RF	1														
q1	-0.29432	1													
q2	-0.21631	0.712815	1												
q3	-0.12222	0.161128	0.14296	1											
q4	0.024316	0.413834	0.630111	0.097055	1										
q5	-0.33266	0.699462	0.864813	0.175389	0.421103	1									
IV	-0.39473	0.75938	0.883082	0.168775	0.497076	0.855173	1								
RV	-0.24188	0.575915	0.772185	0.148161	0.579939	0.677506	0.691924	1							
VRP	-0.15198	0.111969	-0.01386	0.014028	0.012671	-0.03225	0.300309	0.080996	1						
logPEratio	-0.00033	0.18499	0.353795	-0.02836	0.633918	0.148479	0.288306	0.309496	0.007407	1					
logPDratio	-0.0028	-0.02982	-0.0824	0.131782	-0.03611	-0.10405	-0.03729	-0.231	0.030422	0.178488	1				
DFSP	-0.08345	0.430123	0.699069	0.094222	0.710344	0.508584	0.530293	0.787917	-0.0875	0.503023	-0.27477	1			
TMSP	0.011038	0.035995	0.086189	-0.00322	0.203994	0.08102	0.039405	0.254456	-0.02863	0.218014	-0.38025	0.265835	1		
RREL	0.082284	-0.19678	-0.30584	-0.01939	-0.38354	-0.26042	-0.30082	-0.24761	-0.04592	-0.3752	0.16669	-0.40051	-0.28778	1	
CAY	-0.05474	-0.06233	-0.09348	-0.22716	-0.12566	-0.06303	0.012576	-0.33816	0.111942	-0.0704	-0.37282	-0.34617	-0.01127	-0.10793	1

Table 4: Summary statistics for data set 2

	Mkt-RF	q1	q2	q3	q4	q5	IV	RV	VRP	logPEratio	logPDratio	DFSP	TMSP	RREL	CAY
Mean	0.362544	12.828	26.86319	1.651564	55.78818	4.59027	36.63991	19.57285	17.06706	3.118406	3.925334	0.891096	1.719386	-0.19659	0.007853
Std.Dev.	4.324621	29.75476	44.02994	2.046761	53.40869	11.03702	33.80307	38.74569	19.98729	0.271638	0.327846	0.339252	1.16982	0.846024	0.021426
Skewness	-0.80041	6.626053	6.694027	7.133489	3.102167	6.516123	3.987865	8.515514	-3.33984	0.647723	-0.18379	3.730396	0.03593	-0.3763	-0.33676
Kurtosis	1.49916	58.14595	51.52954	61.51707	14.50341	52.22825	23.63669	91.95264	44.40786	0.167255	-1.03064	21.62484	-1.18985	-0.19835	-1.22646
AR(1)	0.097859	0.442045	0.72003	0.666691	0.777738	0.621648	0.776188	0.626021	0.291535	0.947194	0.983331	0.839737	0.974835	0.96622	0.979048
Mkt-RF	1														
q1	-0.33026	1													
q2	-0.24747	0.727903	1												
q3	-0.13557	0.725616	0.909408	1											
q4	-0.00312	0.491169	0.640267	0.713745	1										
q5	-0.33637	0.714086	0.901169	0.744189	0.498733	1									
IV	-0.44261	0.745659	0.851053	0.740418	0.634109	0.818073	1								
RV	-0.38718	0.678746	0.914176	0.702232	0.45862	0.899624	0.856816	1							
VRP	0.001999	-0.05468	-0.33282	-0.10907	0.183381	-0.36039	0.030276	-0.48944	1						
logPEratio	-0.09983	0.258521	0.372118	0.399294	0.585684	0.259346	0.438034	0.304251	0.15102	1					
logPDratio	0.007104	-0.00789	-0.01163	0.001837	0.116631	-0.093	0.110111	0.026514	0.134825	0.545138	1				
DFSP	-0.20266	0.551117	0.757474	0.745397	0.587046	0.667911	0.607147	0.645221	-0.22395	0.357708	-0.16276	1			
TMSP	-0.03157	0.032306	0.073863	0.075537	0.108847	0.067379	0.002481	0.062377	-0.11672	0.24045	-0.39733	0.234069	1		
RREL	0.084863	-0.19988	-0.29478	-0.29054	-0.40286	-0.27469	-0.35281	-0.26679	-0.07951	-0.45222	0.14522	-0.47316	-0.32262	1	
CAY	0.01389	-0.06662	-0.08333	-0.085	0.039664	-0.04657	-0.00409	-0.10895	0.204283	-0.12435	-0.51647	-0.25285	0.232218	-0.10116	1

Table 5: Summary statistics for data set 3

	Mkt-RF	q1	q2	q3	q4	q5	IV	RV	VRP	logPEratio	logPDratio	DFSP	TMSP	RREL	CAY
Mean	0.587639	9.595019	21.45801	1.421551	52.65037	3.360958	33.22895	14.92793	18.30102	3.10809	3.935449	0.839769	1.687037	-0.11733	0.009045
Std.Dev.	4.094514	17.19158	18.16166	0.899821	44.15916	5.095808	23.73355	15.24624	15.13228	0.268226	0.331523	0.204886	1.185668	0.788951	0.021288
Skewness	-0.62718	4.909254	2.625363	2.535238	2.067243	3.518439	2.029571	2.742258	2.1589	0.587132	-0.23879	0.904612	0.094635	-0.33616	-0.4443
Kurtosis	1.055201	31.23881	9.249309	7.924231	6.864252	14.38527	6.287061	10.2428	9.303583	-0.20997	-1.0451	0.318636	-1.21298	-0.12732	-1.11858
AR(1)	0.0943	0.219201	0.749879	0.744	0.878287	0.293329	0.787413	0.702179	0.494378	0.984331	0.989357	0.943115	0.978378	0.965374	0.976087
Mkt-RF	1														
q1	-0.23532	1													
q2	-0.02851	0.484091	1												
q3	0.055971	0.330921	0.857314	1											
q4	0.087187	0.199158	0.612247	0.575035	1										
q5	-0.18408	0.552801	0.600169	0.433583	0.348215	1									
IV	-0.33445	0.573632	0.770688	0.647739	0.597544	0.614215	1								
RV	-0.24211	0.560551	0.89077	0.657322	0.516803	0.635012	0.783125	1							
VRP	-0.28061	0.334915	0.311273	0.353645	0.416496	0.323543	0.779383	0.220727	1						
logPEratio	-0.07944	0.212715	0.458949	0.456389	0.572048	0.20331	0.479242	0.473982	0.274093	1					
logPDratio	-0.03914	0.20881	0.358772	0.356647	0.262348	0.109878	0.335587	0.424282	0.098859	0.621881	1				
DFSP	-0.03959	0.031252	0.335435	0.307244	0.382615	0.197888	0.262973	0.313568	0.096519	0.278949	-0.03579	1			
TMSP	0.007288	-0.13868	-0.11952	-0.08486	0.055214	-0.1063	-0.14344	-0.14015	-0.08378	0.222022	-0.38327	0.209528	1		
RREL	0.009546	-0.05014	-0.25561	-0.25545	-0.42536	-0.24639	-0.30504	-0.21423	-0.26259	-0.45271	0.107227	-0.41924	-0.30971	1	
CAY	-0.03258	0.016343	-0.00732	-0.02511	0.087597	0.045746	0.096796	-0.07502	0.227403	-0.10406	-0.56673	-0.22803	0.267089	-0.22939	1

Table 6: Summary statistics for data set 4

	Mkt-RF	q1	q2	q3	q4	q5	q6	IV	RV	VRP	logPEratio	logPDratio	DFSP	TMSP	RREL	CAY
Mean	0.286232	24.56235	50.85058	2.7659	91.97431	9.907942	76.40564	54.4409	602.1562	5.138232	3.153732	3.857614	1.364783	2.227246	-0.44698	-0.02518
Std.Dev.	5.358734	46.19009	75.95381	3.618239	83.69927	18.63079	21.23886	92.21309	601.0081	51.74585	0.62527	0.181786	0.620426	1.115233	0.681031	0.008814
Skewness	-0.56906	4.676329	3.564698	3.693627	1.979141	3.770404	0.744728	4.10196	2.014729	-3.20643	1.687118	-1.4395	1.952635	-0.79091	-1.40453	1.038288
Kurtosis	0.67194	27.12556	13.50293	15.45087	4.110063	16.12295	0.242103	18.9198	4.643572	24.82286	1.606866	1.427967	3.032597	-0.21451	1.153908	0.298623
AR(1)	0.229916	0.471773	0.751968	0.853583	0.880584	0.646126	0.388129	0.692286	0.790553	-0.31455	0.971113	0.952275	0.945598	0.931716	0.940503	0.853583
Mkt-RF	1															
q1	-0.38498	1														
q2	-0.37948	0.76173	1													
q3	-0.25411	0.753511	0.887495	1												
q4	-0.02619	0.482278	0.636463	0.758861	1											
q5	-0.50693	0.726264	0.905661	0.695997	0.401611	1										
q6	-0.06162	0.273134	0.254894	0.27293	0.121142	0.246588	1									
IV	-0.52456	0.810435	0.905423	0.703229	0.455084	0.902787	0.231824	1								
RV	-0.46033	0.658264	0.841507	0.776467	0.718202	0.739207	0.155774	0.820448	1							
VRP	-0.21884	0.215414	0.030258	0.000978	-0.0553	0.011638	0.032751	0.309328	0.231987	1						
logPEratio	0.085893	0.161785	0.340137	0.379836	0.716163	0.121231	-0.25779	0.230735	0.445553	-0.03513	1					
logPDratio	0.133394	-0.47354	-0.70098	-0.70266	-0.87994	-0.47095	-0.04146	-0.57919	-0.80685	-0.01562	-0.79091	1				
DFSP	-0.12634	0.50199	0.758302	0.793353	0.868773	0.519207	0.068777	0.595233	0.780864	-0.03841	0.698036	-0.94841	1			
TMSP	0.049433	0.158753	0.191329	0.178556	0.475696	0.175229	-0.15578	0.166779	0.411	0.05979	0.311037	-0.35661	0.235641	1		
RREL	0.288214	-0.35684	-0.44253	-0.43036	-0.21194	-0.31813	0.208387	-0.40377	-0.34275	0.018104	-0.33741	0.406224	-0.44483	0.033175	1	
CAY	-0.43603	0.542824	0.664485	0.595117	0.430554	0.585733	0.084981	0.648852	0.637952	-0.03358	0.276121	-0.59329	0.572164	-0.00479	-0.60757	1

Table 7: Monthly return simple regression for data set 1

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12	Model 13
(Intercept)	1.15^{***}	1.18^{***}	1.02^{**}	0.41	1.22^{***}	1.87^{***}	0.99^{***}	0.79^{**}	0.54	1.38^{*}	0.45	0.62^{*}	0.55^{*}
	(0.29)	(0.32)	(0.36)	(0.40)	(0.28)	(0.31)	(0.29)	(0.29)	(2.20)	(0.67)	(0.52)	(0.28)	(0.27)
VoV1	-0.05^{***}												
	(0.01)												
VoV2		-0.02^{***}											
		(0.01)											
VoV3			-0.27^{*}										
			(0.13)										
VoV4				0.00									
				(0.00)									
VoV5					-0.14^{***}								
					(0.02)								
IV						-0.03^{***}							
						(0.00)							
RV							0.00^{***}						
							(0.00)						
VRP								-0.02^{*}					
								(0.01)					
PE									0.00				
									(0.70)				
DFSP										-0.88			
										(0.64)			
TMSP											0.04		
											(0.23)		
RREL												0.47	
												(0.35)	
CAY													-10.33
													(11.44)
R^2	0.09	0.05	0.02	0.00	0.11	0.16	0.06	0.02	0.00	0.01	0.00	0.01	0.00
Adj. R^2	0.08	0.04	0.01	0.00	0.11	0.15	0.06	0.02	0.00	0.00	0.00	0.00	0.00
Num. obs.	273	273	273	273	273	273	273	273	273	273	273	273	273

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table 8: Monthly return simple regression for data set 2

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12	Model 13
(Intercept)	0.98^{**}	1.02^{**}	0.84^{*}	0.38	0.97^{**}	2.44^{***}	1.21^{***}	0.36	5.32	2.66^{***}	0.56	0.45	0.34
	(0.30)	(0.33)	(0.37)	(0.42)	(0.29)	(0.38)	(0.30)	(0.38)	(3.30)	(0.79)	(0.51)	(0.29)	(0.31)
VoV1	-0.05^{***}												
	(0.01)												
VoV2		-0.02^{***}											
		(0.01)											
VoV3			-0.29^{*}										
			(0.14)										
VoV4				0.00									
				(0.01)									
VoV5					-0.13^{***}								
					(0.02)								
IV						-0.06^{***}							
						(0.01)							
RV							-0.04^{***}						
							(0.01)						
VRP								0.00					
								(0.01)					
PE									-1.59				
									(1.05)				
DFSP										-2.58^{**}			
										(0.83)			
TMSP											-0.12		
											(0.25)		
RREL												0.43	
												(0.34)	
CAY													2.80
													(13.43)
R^2	0.11	0.06	0.02	0.00	0.11	0.20	0.15	0.00	0.01	0.04	0.00	0.01	0.00
Adj. R^2	0.11	0.06	0.01	0.00	0.11	0.19	0.15	0.00	0.01	0.04	0.00	0.00	0.00
Num. obs.	228	228	228	228	228	228	228	228	228	228	228	228	228

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table 9: Monthly return simple regression for data set 3

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12	Model 13
(Intercept)	1.13^{***}	$0.73^{\cdot}$	0.23	0.16	1.08^{**}	2.50^{***}	1.56^{***}	1.98^{***}	4.36	1.25	0.55	0.59^{*}	0.64^{*}
	(0.31)	(0.43)	(0.52)	(0.43)	(0.33)	(0.45)	(0.38)	(0.42)	(3.25)	(1.18)	(0.49)	(0.28)	(0.30)
VoV1	-0.06^{***}												
	(0.02)												
VoV2		-0.01											
		(0.02)											
VoV3			0.25										
			(0.31)										
VoV4				0.01									
				(0.01)									
VoV5					-0.15^{**}								
					(0.05)								
IV						-0.06^{***}							
						(0.01)							
RV							-0.07^{***}						
							(0.02)						
VRP								-0.08^{***}					
								(0.02)					
PE									-1.21				
									(1.04)				
DFSP										-0.79			
										(1.37)			
TMSP											0.03		
											(0.24)		
RREL												0.05	
												(0.35)	
CAY													-6.27
													(13.14)
R^2	0.06	0.00	0.00	0.01	0.03	0.11	0.06	0.08	0.01	0.00	0.00	0.00	0.00
Adj. R^2	0.05	0.00	0.00	0.00	0.03	0.11	0.05	0.07	0.00	0.00	0.00	0.00	0.00
Num. obs.	216	216	216	216	216	216	216	216	216	216	216	216	216

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table 10: Monthly return simple regression for data set 4

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12	Model 13
(Intercept)	1.38^{*}	1.65^{*}	$1.33^{\cdot}$	0.44	1.73^{**}	1.95^{**}	2.76^{**}	0.40	-2.04	1.78	-0.24	$1.30^{\cdot}$	-6.39^{***}
	(0.68)	(0.73)	(0.79)	(0.97)	(0.64)	(0.64)	(0.82)	(0.64)	(3.35)	(1.57)	(1.46)	(0.75)	(1.78)
VoV1	-0.04^{**}												
	(0.01)												
VoV2		-0.03^{**}											
		(0.01)											
VoV3			-0.38^{*}										
			(0.17)										
VoV4				0.00									
				(0.01)									
VoV5					-0.15^{***}								
					(0.03)								
IV						-0.03^{***}							
						(0.01)							
RV							0.00^{***}						
							(0.00)						
VRP								$-0.02^{\cdot}$					
								(0.01)					
PE									0.74				
									(1.04)				
DFSP										-1.09			
										(1.05)			
TMSP											0.24		
											(0.59)		
RREL												2.27^{*}	
												(0.92)	
CAY													-265.08^{***}
													(66.84)
R^2	0.15	0.14	0.06	0.00	0.26	0.28	0.21	0.05	0.01	0.02	0.00	0.08	0.19
Adj. R^2	0.14	0.13	0.05	-0.01	0.25	0.26	0.20	0.03	-0.01	0.00	-0.01	0.07	0.18
Num. obs.	69	69	69	69	69	69	69	69	69	69	69	69	69

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table 11: Monthly return multiple regression for data set 1

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
(Intercept)	1.86^{***}	1.79^{***}	1.43^{***}	1.33^{***}	0.93^{*}	0.81^{*}
	(0.30)	(0.30)	(0.32)	(0.34)	(0.36)	(0.38)
VoV2	0.06^{***}	0.07^{***}				
	(0.01)	(0.01)				
IV	-0.08^{***}	-0.08^{***}	-0.05^{***}	-0.05^{***}	-0.05^{***}	-0.04^{***}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
RV		$0.00^{\cdot}$		0.00		0.00
		(0.00)		(0.00)		(0.00)
VoV3			0.66^{***}	0.75^{***}		
			(0.17)	(0.19)		
VoV4					0.02^{***}	0.02^{***}
					(0.00)	(0.01)
R^2	0.24	0.25	0.20	0.21	0.22	0.22
Adj. R^2	0.23	0.24	0.20	0.20	0.21	0.21
Num. obs.	273	273	273	273	273	273

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table 12: Monthly return multiple regression for data set 2

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
(Intercept)	3.07^{***}	2.45^{***}	2.43^{***}	2.17^{***}	1.72^{***}	1.92^{***}
	(0.39)	(0.40)	(0.36)	(0.41)	(0.37)	(0.41)
VoV2	0.05^{***}	0.09^{***}				
	(0.01)	(0.01)				
IV	-0.11^{***}	-0.08^{***}	-0.10^{***}	-0.08^{***}	-0.09^{***}	-0.11^{***}
	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)	(0.02)
RV		-0.07^{***}		-0.02		0.01
		(0.02)		(0.01)		(0.01)
VoV3			0.90^{***}	0.94^{***}		
			(0.18)	(0.18)		
VoV4					0.04^{***}	0.04^{***}
					(0.01)	(0.01)
R^2	0.26	0.32	0.28	0.28	0.32	0.33
Adj. R^2	0.25	0.31	0.27	0.27	0.32	0.32
Num. obs.	228	228	228	228	228	228

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table 13: Monthly return multiple regression for data set 3

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
(Intercept)	2.27^{***}	1.67^{***}	1.21^{*}	1.04^{*}	1.85^{***}	1.85^{***}
	(0.42)	(0.40)	(0.47)	(0.49)	(0.43)	(0.44)
VoV2	0.13^{***}	0.25^{***}				
	(0.02)	(0.03)				
IV	-0.13^{***}	-0.10^{***}	-0.11^{***}	-0.09^{***}	-0.10^{***}	-0.10^{***}
	(0.02)	(0.02)	(0.01)	(0.02)	(0.01)	(0.02)
RV		-0.21^{***}		-0.04		0.00
		(0.03)		(0.03)		(0.03)
VoV3			2.14^{***}	2.31^{***}		
			(0.36)	(0.38)		
VoV4					0.04^{***}	0.04^{***}
					(0.01)	(0.01)
R^2	0.24	0.35	0.24	0.25	0.24	0.24
Adj. R^2	0.23	0.34	0.23	0.24	0.23	0.23
Num. obs.	216	216	216	216	216	216

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table 14: Monthly return multiple regression for data set 4

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
(Intercept)	1.56^{*}	2.28^{**}	1.52^{*}	2.14^{**}	0.75	$1.33^{\cdot}$
	(0.65)	(0.79)	(0.69)	(0.80)	(0.80)	(0.78)
VoV2	0.04^{*}	0.05^{**}				
	(0.02)	(0.02)				
IV	-0.06^{***}	-0.05^{***}	-0.04^{***}	-0.03^{**}	-0.04^{***}	-0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
RV		0.00		0.00		-0.01^{**}
		(0.00)		(0.00)		(0.00)
VoV3			0.34	0.51^{*}		
			(0.21)	(0.24)		
VoV4					0.02^{*}	0.04^{***}
					(0.01)	(0.01)
R^2	0.33	0.35	0.30	0.32	0.33	0.42
Adj. R^2	0.31	0.32	0.28	0.29	0.31	0.39
Num. obs.	69	69	69	69	69	69

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table 15: Predicting market return for horizon 1 to 12 by measurement 4

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12
(Intercept)	0.24	0.31	0.27	0.18	0.14	0.11	0.11	0.11	0.14	0.18	0.21	$0.22^{\cdot}$
	(0.40)	(0.30)	(0.24)	(0.21)	(0.19)	(0.17)	(0.16)	(0.15)	(0.14)	(0.13)	(0.13)	(0.12)
q1	0.00	0.00	0.00	0.01^{*}	0.01^{**}	0.01^{***}	0.01^{***}	0.01^{***}	0.01^{***}	0.01^{***}	0.01^{***}	0.01^{***}
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
R^2	0.00	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.06	0.06	0.06
Adj. R^2	0.00	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.06	0.06	0.06	0.06
Num. obs.	273	273	273	273	273	273	273	273	273	273	273	273

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table 16: Predicting market return for horizon 13 to 24 by measurement 4

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12
(Intercept)	$0.23^{\cdot}$	0.25^{*}	0.26^{*}	0.29^{**}	0.31^{**}	0.34^{**}	0.36^{***}	0.37^{***}	0.38^{***}	0.39^{***}	0.40^{***}	0.42^{***}
	(0.12)	(0.11)	(0.11)	(0.11)	(0.11)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.09)	(0.09)
q1	0.01^{***}	0.01^{***}	0.01^{***}	0.01^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{**}	0.00^{**}	0.00^{**}
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
R^2	0.07	0.07	0.06	0.06	0.05	0.05	0.04	0.04	0.04	0.04	0.04	0.03
Adj. R^2	0.06	0.06	0.06	0.06	0.05	0.04	0.04	0.04	0.04	0.04	0.03	0.03
Num. obs.	273	273	273	273	273	273	273	273	273	273	273	273

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table 17: Variance risk premium and volatility of volatility for data set 1

	Model 1	Model 2	Model 3	Model 4	Model 5
(Intercept)	9.31^{***}	11.10^{***}	10.84^{***}	10.43^{***}	11.26^{***}
	(1.92)	(2.11)	(2.31)	(2.59)	(1.93)
VoV1	$0.11^{\cdot}$				
	(0.06)				
VoV2		-0.01			
		(0.04)			
VoV3			-0.01		
			(0.85)		
VoV4				0.01	
				(0.03)	
VoV5					-0.09
					(0.16)
R^2	0.01	0.00	0.00	0.00	0.00
Adj. R^2	0.01	0.00	0.00	0.00	0.00
Num. obs.	273	273	273	273	273

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table 18: Variance risk premium and volatility of volatility for data set 2

	Model 1	Model 2	Model 3	Model 4	Model 5
(Intercept)	17.54^{***}	21.13^{***}	18.83^{***}	13.24^{***}	20.06^{***}
	(1.44)	(1.47)	(1.70)	(1.89)	(1.34)
VoV1	-0.04				
	(0.04)				
VoV2		-0.15^{***}			
		(0.03)			
VoV3			-1.07		
			(0.65)		
VoV4				0.07^{**}	
				(0.02)	
VoV5					-0.65^{***}
					(0.11)
R^2	0.00	0.11	0.01	0.03	0.13
Adj. R^2	0.00	0.11	0.01	0.03	0.13
Num. obs.	228	228	228	228	228

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table 19: Variance risk premium and volatility of volatility for data set 3

	Model 1	Model 2	Model 3	Model 4	Model 5
(Intercept)	15.47^{***}	12.74^{***}	9.85^{***}	10.79^{***}	15.07^{***}
	(1.11)	(1.52)	(1.81)	(1.46)	(1.17)
VoV1	0.29^{***}				
	(0.06)				
VoV2		0.26^{***}			
		(0.05)			
VoV3			5.95^{***}		
			(1.08)		
VoV4				0.14^{***}	
				(0.02)	
VoV5					0.96^{***}
					(0.19)
R^2	0.11	0.10	0.13	0.17	0.10
Adj. R^2	0.11	0.09	0.12	0.17	0.10
Num. obs.	216	216	216	216	216

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table 20: Variance risk premium and volatility of volatility for data set 4

	Model 1	Model 2	Model 3	Model 4	Model 5
(Intercept)	-0.79	4.09	5.10	8.28	4.82
	(6.95)	(7.57)	(7.92)	(9.35)	(7.12)
VoV1	$0.24^{\cdot}$				
	(0.13)				
VoV2		0.02			
		(0.08)			
VoV3			0.01		
			(1.75)		
VoV4				-0.03	
				(0.08)	
VoV5					0.03
					(0.34)
R^2	0.05	0.00	0.00	0.00	0.00
Adj. R^2	0.03	-0.01	-0.01	-0.01	-0.01
Num. obs.	69	69	69	69	69

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

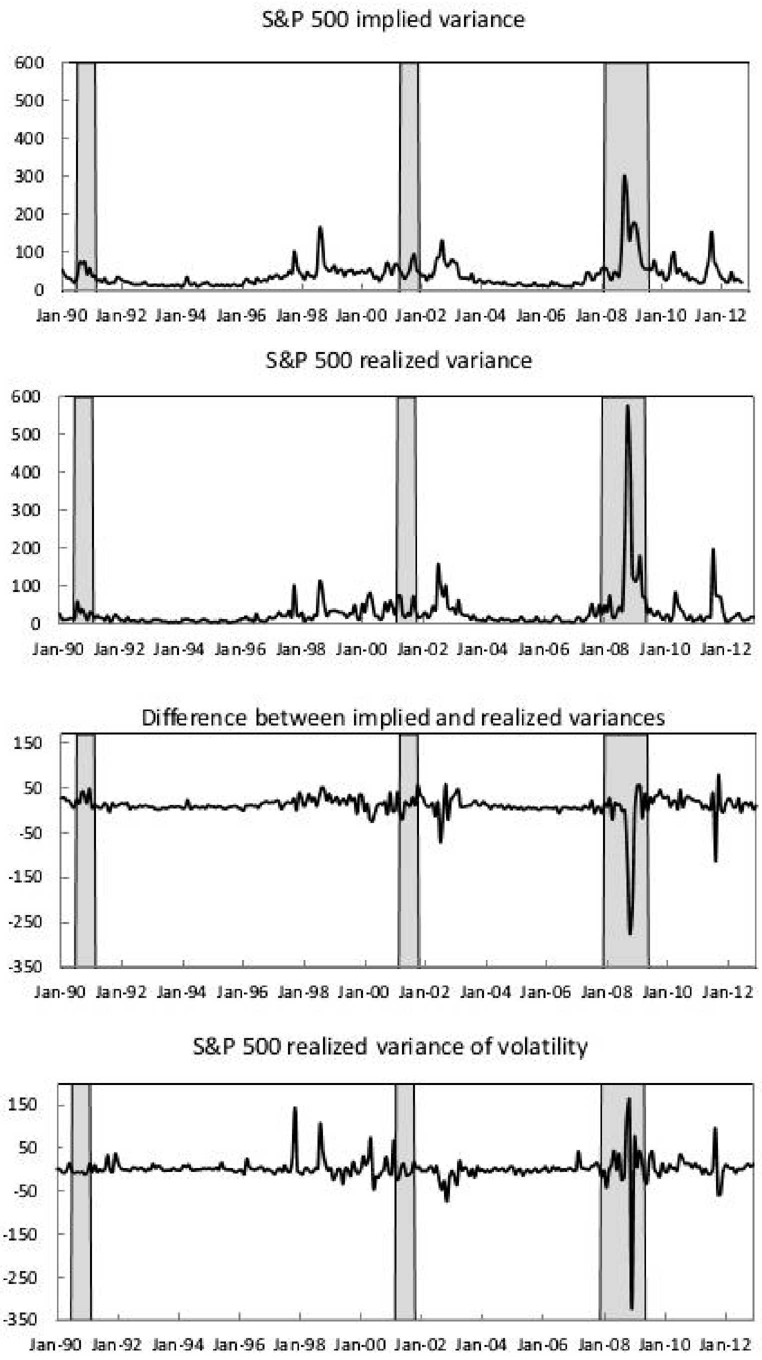

Figure 1: Realized variance,implied variance and volatility of volatility

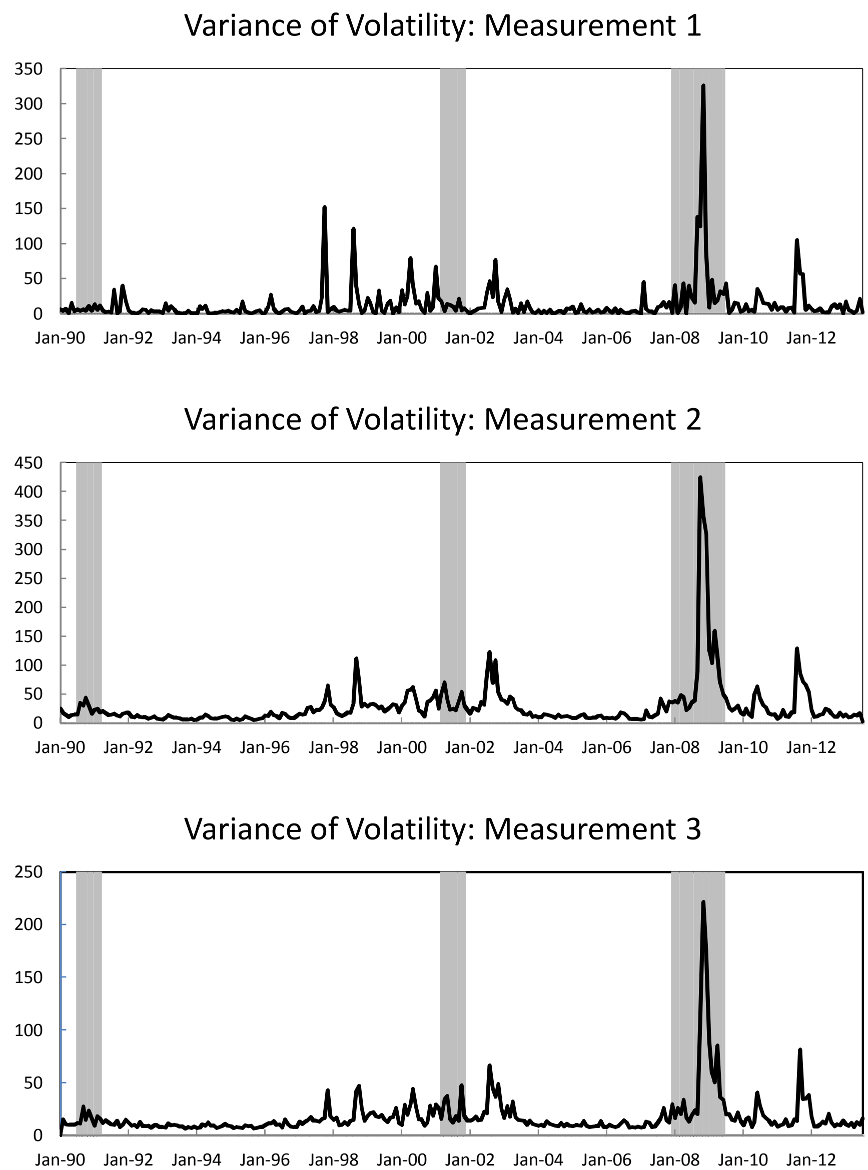

Figure 2: Measurement 1 to 3

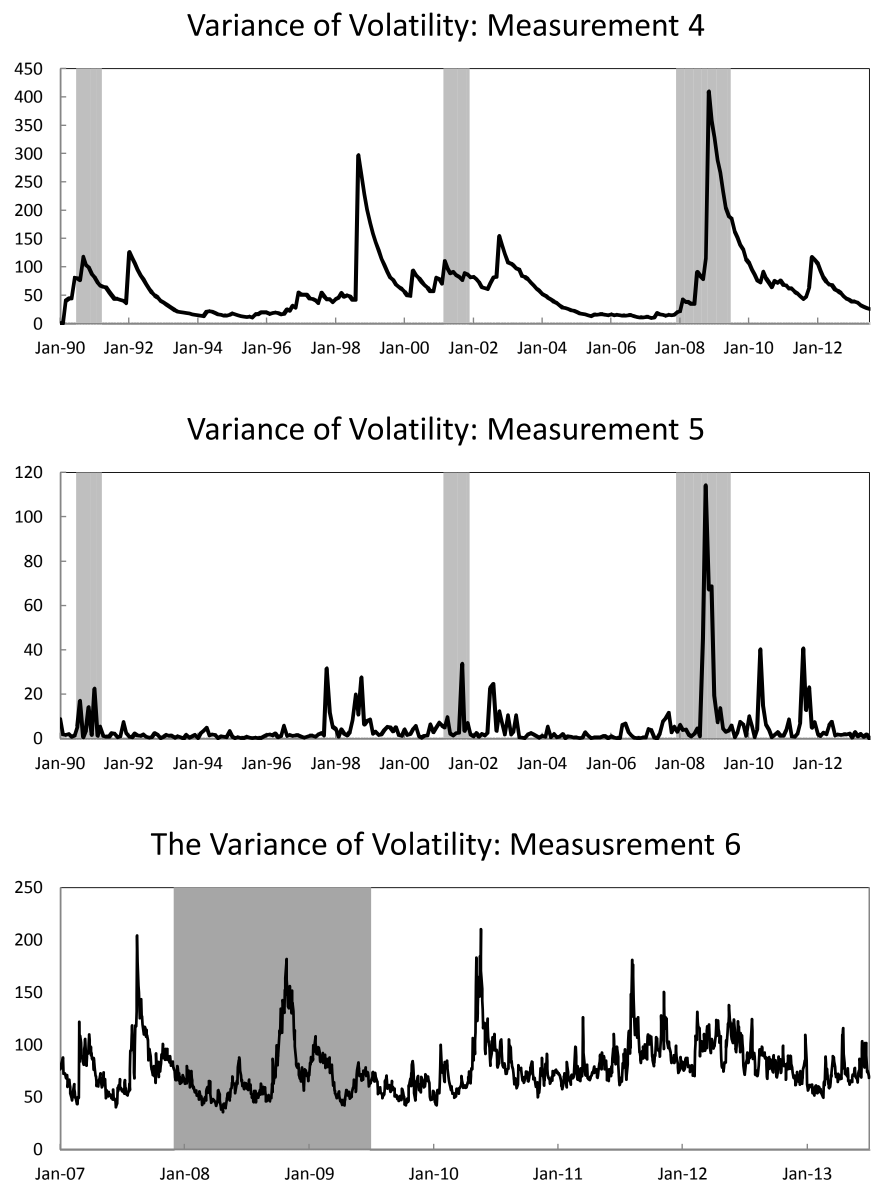

Figure 3: Measurement 4 to 6

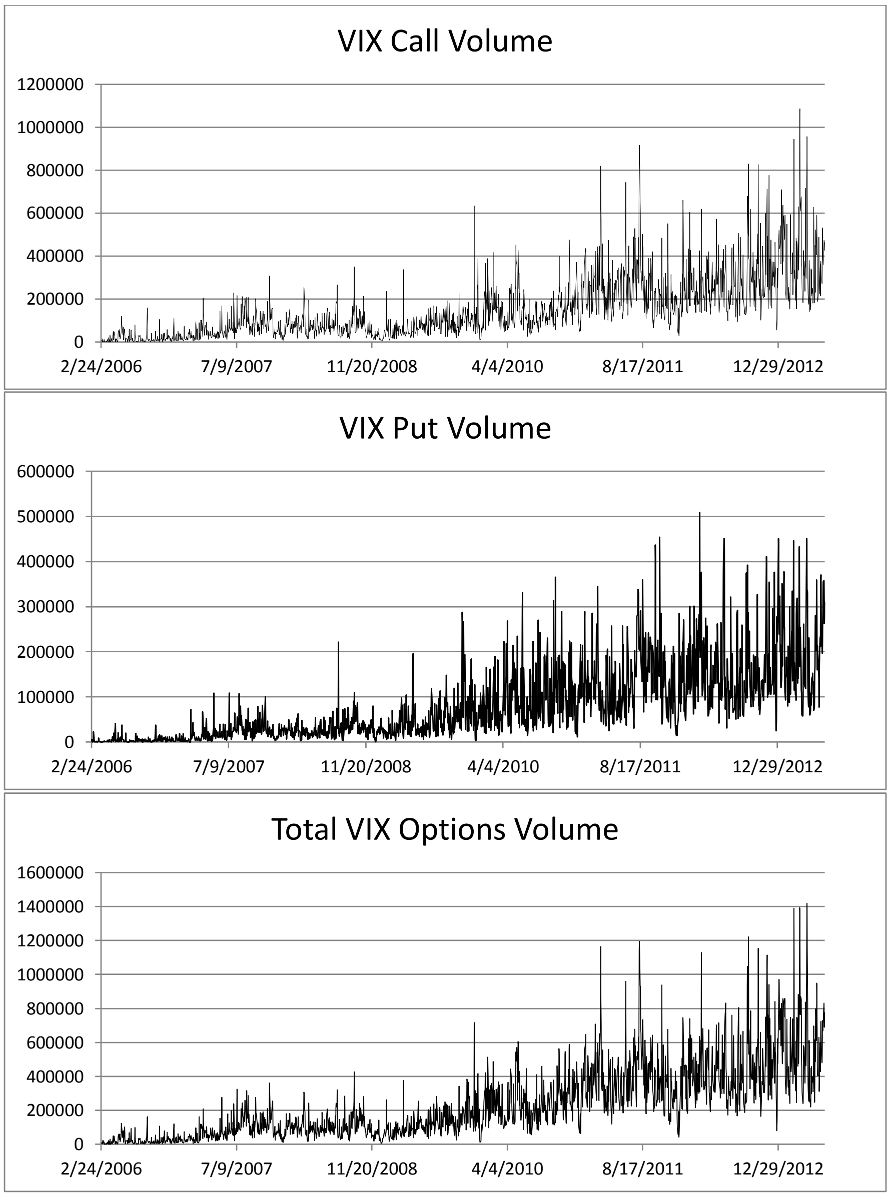

Figure 4: Trading volume of VIX options

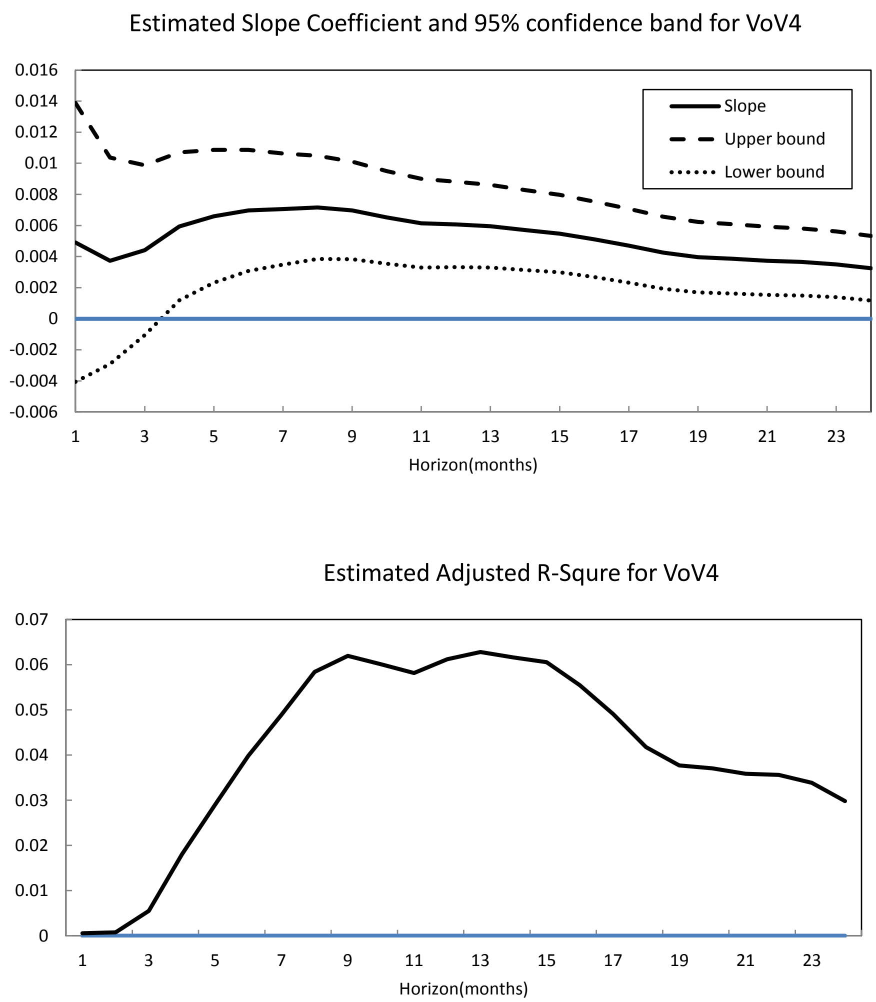

Figure 5: Predicting future return by VoV4

CHAPTER 2: LONG MEMORY IN VOLATILITY AND RETURN PREDICTABILITY

This chapter proposes a new way to measure the variance risk premium by applying a fractional cointegration relationship between implied variance and realized variance. To find the fractional cointegration coefficient between implied variance and realized variance, we develop a search method based on minimization of the score test statistic proposed by Robinson(1994). We use daily, weekly and monthly data of five stock market indexes (S&P500, S&P100, DJIA, NASDAQ100 and Russell2000) and their volatility indexes from the CBOE. We find our new measure improves the return prediction power of the variance risk premium both in-sample statically and out-of-sample dynamically, and the result is robust for the monthly data among all five indexes.

2.1 Introduction

Variance Risk Premium has been a concept that attracted lots of research attention recently because of its potential to exlain the difference between the market expected volatility and the realized volatility, and also its return prediction power, see Carr and Wu (2006), Carr and Wu (2009), Barndorff-Nielsen and Veraart (2011) and Drechsler and Yaron (2011). The standard way to measure the variance risk premium is the difference between implied variance and realized variance. The implied variance is usually measured daily from the option data, whereas the realized variance is usually measured from high frequency data. Because of the frequency difference of these two measurements, there is not a concensus about what the coefficient should be in front of the realized variance for us to get a better measurement of variance risk premium.

This chapter provides a method to find the coefficient between implied variance and realized variance from the well developed fractional cointegration literature. Because it's stylized fact that return is usually stationary where as volatility, different measurement of it, are always fractionally integrated, see Poon and Granger (2003) and Bandi and Perron (2006). Statistically speaking, the reason that return could be explained by variance risk premium migh partly due to that the stationarity of variance risk premium.

Given the possible fractional cointegration between implied variance and realized variance, we are able to find a fractional cointegration coefficient between them. Currently in the literature, there are two ways to find the degree of integration of a time series, see Robinson and Yajima (2002). One way is to measure it directly by minimizing a likelihood function, see Shimotsu (2007) and Shimotsu(2012), the other way is to assume that it equals to d_0 and test it by a score test statistic, see Gil-Alana (2000). We combined these two ways to find the fractional cointegration coefficient. We find the degree of two time series by the first way, and find the coefficient by minimizing the test statistic from the second way.

We started from a big pool of data, it's the daily, weekly and monthly data for the 5 stock indexes and their volatility indexes from Chicago Board of Exchange, the variables include return and volatility. Given the previous literature in volatility and volume, see Fleming and Kirby (2010), we also add stock indexes trading volume into our data, we start by estimating the degree of integration of these time series. Our finding is in line with the stylized facts that the return is always stationary whereas volatility and volume are usually fractionally integrated. However, because of the significant different degree of integration between volatility and volume, the only fractional cointegration relationship we find is between implied volatility and realized volaltility (and also different measures of them) for monthly data, which gives us a chance to improve the measurement of variance risk premium by using this fractional cointegration relationshi between implied variance and realized variance.

By minimizing the test statistic $\hat{r}^2$ proposed by Robinson(1994), we did find the fractional cointegration coefficient in front of realized variance could improve the return prediction power of variance risk premium.

The remainder of this chapter is organized as follows: Section 2 introduces the fractional integration and fractional cointegration model we use and proposed our way to estimate the frational cointegration coefficient. Section 3 describes the dataset and details of our construction of different measurement of realized volitility and implied volatility, and discusses the summary statitics of our data. Section 4 is the empirical results and investigated return forecast power by our improved measurement of variance risk premium. Section 5 offers some concluding remarks.

2.2 Model and Econometric Methodology

2.2.1 Fractionally Integration

We model the potential persistence in return, volatility and volume through a long memory model. The memory parameter d in the fractionally integrated processes is estimated in a semiparametric way. We report results for the "two step feasible exact local whittle estimator with detrending" [Shimotsu(2010)] [10]in our paper. We have also considered the gaussian semiparametric local whittle estimator [Kunsch(1987) and Robinson (1995)] as well as the exact local whittle estimator [Shimotsu and Phillips (2005)]. Our main findings are not affected by the choice of estimator. The corresponding results are available upon request. We briefly review the two step feasible exact local whittle estimator with detrending by Schimotsu as it serves as the basis for our estimation of the memory parameter d in this chapter.

2.2.1.1 The Exact Local Whittle Estimator

Consider a univariate series y_t which has the representation

$$\Delta^d y_t = (1-L)^d y_t = u_t \mathbf{1}\{t \geq 1\} \tag{5}$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function and u_t is assumed to be stationary with zero mean and spectral density $f_u(\lambda)$ satisfying $f_u(\lambda) \sim G$ for $\lambda \sim 0$. More specificitly, by expanding $(1-L)^d$, (5) can be rewritten as

$$\sum_{k=0}^{t} \frac{\Gamma(k-d)}{k!\Gamma(-d)} y_{t-k} = u_t$$

Define the Discrete Fourier Transform (DFT) and the periodogram of a time series at evaluated at the fundamental frequencies as

$$w_a(\lambda_s) = \frac{1}{\sqrt{2\pi n}} \sum_{t=1}^{n} a_t e^{it\lambda_s}, \qquad \lambda_s = \frac{2\pi s}{n}, s = 1, ..., n,$$
$$I_a(\lambda_s) = |w_a(\lambda_s)|^2$$

[10]We do appreciate that Dr. Shimotsu share the code with us.

Shimotsu and Phillips define the exact local whittle estimator as

$$\hat{d} = \arg\min R(d),$$

and

$$R(d) = \log \hat{G}(d) - \frac{2d}{m} \sum_{s=1}^{m} \log \lambda_j,$$

where

$$\hat{G}(d) = \frac{1}{m} \sum_{s=1}^{m} I_{\Delta^d y}(\lambda_j).$$

This exact local whittle estimator has been shown to be asymptotically normal:

$$\sqrt{m}(\hat{d} - d) \rightarrow \mathbf{N}(0, \frac{1}{4})$$

m here is a bandwidth parameter that determines the locality of the estimates; the choice of it usually involves a standard bias-variance tradeoff. To make sure our result is robust with the choice of this bandwidth parameter, we showed our results as m equals to $[n^{.5}]$, $[n^{.55}]$ and $[n^{.6}]$, where $[x]$ is the integer part of x.

2.2.1.2 The Two Step Feasible Exact Local Whittle Estimator with Detrending

For the two step feasible exact local whittle estimator with detrending, there are several elements that we need to review one by one:

1. Exact local whittle estimator with unknown mean

   If y_t is generated by a process with unknown mean:

   $$y_t = \mu_0 + y_t^0, \qquad y_t^0 = (1 - L)^d u_t \mathbf{1}\{t \geq 1\}$$

   then if we define

   $$\tilde{\mu}(d) = w(d)\bar{y} + (1 - w(d))y_1$$

where

$$w(d) = \begin{cases} 1 & \text{if } d \leq \frac{1}{2} \\ 0 & \text{if } d \geq \frac{3}{4} \\ \text{twice differentiable function} & \text{otherwise} \end{cases}$$

the estimator d now minimize the following objective function

$$R_F(d) = \log \hat{G}_F(d) - \frac{2d}{m}\sum_{s=1}^{m} \log \lambda_j, \qquad \text{where } \hat{G}_F(d) = \frac{1}{m}\sum_{s=1}^{m} I_{\Delta^d(y-\tilde{\mu}(d))}(\lambda_j) \tag{6}$$

2. Two step exact local whittle estimator

   Since it's hard to prove its global consistency in 6, Shimotsu proposed the two step exact local whittle estimator: if we denote $\hat{d}_T$ as the first-stage estimator, then the two-step exact local whittle estimator $\hat{d}_{2ELW}$ should be:

$$\hat{d}_{2ELW} = \hat{d}_T - R'_F(\hat{d}_T)/R''_F(\hat{d}_T)$$

3. Detrending

   Besides the unknown mean, if the data also have a polynomial time trend:

$$y_t = \mu_0 + \beta_{10}t + \beta_{20}t^2 + \cdots + \beta_{k0}t^k + y_t^0, \qquad y_t^0 = (1-L)^d u_t \mathbf{1}\{t \geq 1\}$$

   to detrend, regress y_t on $(1, t, t^2, ..., t^k)$ first and then apply the two-step estimation to the residual $y_t - \hat{y}_t$, the estimator d would be the two step exact loacal whittle estimator with detrending.

4. Feasible

   Again, because of the difficulty in proving the the global consistency of the two-step exact local whittle estimator, the feasible set to search for d to minimize the objective function would not be $[0, 1]$, but for arbitray small $\nu > 0$, the feasible set would be $[\nu, 1-\nu]$. so the two step exact local whittle estimator with detrending that search within the feasible set $[\nu, 1-\nu]$ would be the two step feasible exact local whittle estimator with detrending.

2.2.2 Fractional Cointegration

The definition of fractionally cointegration, as appeared in Cheung and Lai (1993) for the first time, comes from a generalization to fractional cointegration of the definition of standard cointegration as given in Engle and Granger (1987):

The component processes of an $(n \times 1)$ vector X_t are said to be fractionally cointegrated of order d, b with $b > 0$ (denoted $X_t \sim CI(d, b)$) if

1. for each i, $1 \leq i \leq n$, $X_{i,t} \sim I(d)$, and
2. there exists a vector $\xi \in R^n$ such that $Y_t = \xi' X_t \sim I(d-b)$.

This is so far the most used definition of fractional cointegration, we will discuss how to test it and how to find ξ next.

2.2.2.1 Robinson's (1994) score $\hat{r}^2$

Robinson (1994) proposed a way to test the degree of integration of a linear combination $Y_t = \xi' X_t$, by using a score statistic $\hat{r}^2$, which follows a χ^2_p distribution where p is the number of restrictions tested: given an $(n \times 1)$ vector X_t, the components of which are each $I(d)$ series, assume that the spectral density of the linear combination is given by $f(\omega; \tau, \sigma^2)$, for $0 < \omega < 2\pi$; with $\tau \in R^m$, where m is the number of frequencies we chose and $\sigma \in R$ is unknown, but f is linear in σ^2. So f could be written as:

$$f(\omega;\tau,\sigma^2) = \frac{\sigma^2}{2\pi} g(\omega;\tau),$$

Robinson(1994) choose Bloomfield (1973)'s exponential model as the base function $g(\omega,\tau)$, because "it leads to an especially neat version of frequency domain test statistic" [11]

$$g(\omega;\tau) = \exp[2\sum_{r=1}^{m} \tau_r \cos(r\omega)]$$

where this τ_r is just a nuisance parameter that could help to optimize the statistic $\hat{r}^2$, and we will talk about that later. So to test if the linear combination $\xi' X_t$ is integrated with degree d as $(1-L)^d \xi' X_t = u_t$, where u_t is a stationary process, Robinson's test statistic is:

$$\hat{r}^2 = T[(\frac{\hat{a}}{\hat{\sigma}^2})' \hat{A}^{-1} (\frac{\hat{a}}{\hat{\sigma}^2})]$$

which has a typical form of a score test statistic: where T is the sample size, or the length of the time series; $\hat{a}$ is the random variable which will be asymptotically normal;

$$\hat{a} = -\frac{2\pi}{T} \sum_{j=1}^{T-1} \psi(\omega_j) \frac{I(u_t;\omega_j)}{g(\omega_j;\hat{\tau})}$$

$\hat{\sigma}^2$ is a scaler to standardize $\hat{a}$,

$$\hat{\sigma}^2 = \min_{\tau} \sigma^2(\tau) = \min_{\tau} \frac{2\pi}{T} \sum_{j=1}^{T-1} \frac{I(u_t;\omega_j)}{g(\omega_j;\tau)}, \qquad \hat{\tau} = \arg\min_{\tau} \sigma^2(\tau)$$

whereas $\hat{A}$ is the typical estimator of variance-coveriance matrix,

$$\hat{A} = \frac{2}{T}[\sum_{j=1}^{T-1} \psi(\omega_j)\psi(\omega_j)' - \sum_{j=1}^{T-1} \psi(\omega_j)\hat{\epsilon}(\omega_j)'(\sum_{j=1}^{T-1} \hat{\epsilon}(\omega_j)\hat{\epsilon}(\omega_j)')^{-1} \sum_{j=1}^{T-1} \hat{\epsilon}(\omega_j)\psi(\omega_j)']$$

[11] majorly it was often possible to approximate the logarithm of an estimated spectral density by a truncated Foureier series.

The only difference here is $\hat{a}$, $\hat{\sigma}^2$ and $\hat{a}$ is in the frequency domain instead of time domain, that's why we have:

$$\psi(\omega_j) = \log[2\sin(\omega_j/2)], \qquad \hat{\epsilon}(\omega_j) = D_\tau[\log(g(\omega_j);\hat{\tau})], \qquad \omega_j = \frac{2\pi j}{T}$$

and $I(\cdot;\omega)$ is the periodogram of its argument evluated at ω,

$$I(u_t;\omega_j) = \frac{1}{2\pi T}|\sum_{t=1}^{T} u_t e^{it\omega_j}|^2$$

2.2.2.2 A Search Procedure for Identifying the Cointegrating Vector

Although Robinson(1994) proposed the score statistic $\hat{r}^2$ to test if the linear combination $\xi' X_t$ is integrated with degree d as $(1-L)^d \xi' X_t = u_t$, where u_t is a stationary process, but he didn't specify what ξ' should be, and theory actually suggests a range of possible values for the cointegrating vector. We propose a search procedure to find ξ by minimizing $\hat{r}^2$.

$$\hat{\xi} = \arg\min_{\xi\in R^{n-1}} \hat{r}^2 = \arg\min_{\xi\in R^{n-1}} T[(\frac{\hat{a}}{\hat{\sigma}^2})'\hat{A}^{-1}(\frac{\hat{a}}{\hat{\sigma}^2})]$$

We search ξ in R^{n-1} because we can always scale down one of them to 1 and then search for the other coefficients. To search ξ for a given d, since A is a scalar that will not be affected by the choice of ξ, to minimize $\hat{r}^2$, we are minimizing $\frac{\hat{a}}{\hat{\sigma}^2}$:

$$\frac{\hat{a}}{\hat{\sigma}^2} = \frac{-\frac{2\pi}{T}\sum_{j=1}^{T-1}\psi(\omega_j)\frac{I(u_t;\omega_j)}{g(\omega_j;\hat{\tau})}}{\frac{2\pi}{T}\sum_{j=1}^{T-1}\frac{I(u_t;\omega_j)}{g(\omega_j;\tau)}} = \frac{-\sum_{j=1}^{T-1}\psi(\omega_j)\frac{I(u_t;\omega_j)}{g(\omega_j;\hat{\tau})}}{\sum_{j=1}^{T-1}\frac{I(u_t;\omega_j)}{g(\omega_j;\tau)}}$$

Since $I(u_t;\omega_j)$ is like a variance, $\frac{\hat{a}}{\hat{\sigma}^2}$ is like the ratio of a weighted variance and an average variance. So the crucial part here is the weight $\psi(\omega_j)$ here. We plot it for $\omega \in (0, 2\pi)$ in figure .

It's obvious showed in the figure that the weight is positive between $\frac{\pi}{3}$ and $\frac{5\pi}{3}$, but when ω get close to 0 or 2π, the weight goes to negative infinite, therefore, a local

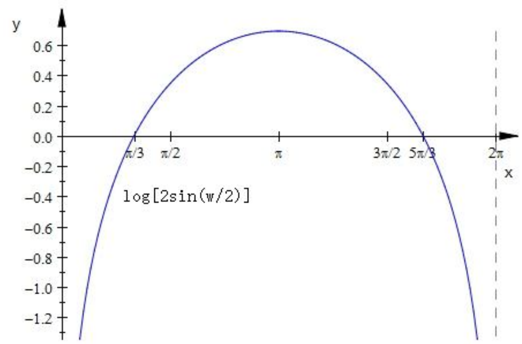

Figure 6: The weight function $\psi(\omega_j) = \log[2\sin(\frac{\omega_j}{2})]$ when $\omega_j \in (0, 2\pi)$

minimum of $\hat{r}^2$ will be at a point for which the "weighted variance" has the smallest magnatitude possible relative to the "average variance". Or, put it in another way, values of ξ that produce local minima of $\hat{r}^2$ correspond to linear combination of $\xi' X_t$ in which the frequencies near 0^+ and $2\pi^-$ contribute less variance than in other linear combination in a neighborhood near ξ.

2.3 Data and Measurements

We covered 5 major indexes in US market: SP500, SP100, DJIA, NASDAQ100 and Russell2000. The reason we chose these 5 indexes is because, first, they represent different companies in US market, from big to small, from technology to traditional firms; second, CBOE has implied volatility indexes based on these 5 indexes respectively. The start date of these 5 implied volatility index is documented in the following table. But for us to be able to compare these 5 indexes, we focus on the time period from Jan 2004 to Jun 2013.

Therefore, with 5 indexes and 3 different frequencies (Daily, Weekly and Monthly), we have 15 datasets to work with. For all the 5 daily datasets, we have data on

Table 21: Sample size for different data set

Market Index	Volatility Index	Start	End	Daily	Weekly	Monthly
SP500	VIX	1/2/1990	6/28/2013	5921	1225	282
SP100	VXO	1/2/1986	6/28/2013	6925	1433	330
DJIA	VXD	11/3/1997	6/28/2013	3938	816	188
NASDAQ100	VXN	2/1/2001	6/28/2013	3116	646	149
RUSSELL2000	RVX	1/2/2004	6/28/2013	2387	495	114

Return, Implied Volatility Index, Volume and Range; whereas for the 5 weekly and 5 monthly datasets, we have data on Return, Implied Volatility Index, Volume, Range, Realized Volatility and Variance Risk Premium, we will talk about the measurement of these variables one by one.

We measured index return by the formula:

$$r_t \equiv ln(\frac{P_t}{P_{t-1}}) * 100$$

and since it's from two consecutive prices, it's not annualized return. The implied volatility index comes from CBOE (Chicago Board of Exchange) website, VIX, VXD, VXN and RVX are all calculated by this "model-free" method:

$$E_t^Q(\sigma^2_{r,t+1}) \equiv IV \equiv 2\int_0^\infty \frac{C_t(t+1, \frac{K}{B(t,t+1)}) - C_t(t,K)}{K^2} dK$$

where $C_t(T, K)$ denote the price of a European call option at time t with the strike price at K and Maturity at T, $\frac{1}{B(t,t+1)}$ is the discount rate from t to $t+1$, and for out data, the maturity is always 30 days. But VXO are still calculated by the classic at-the-money Black-Scholes formula. We used the daily squared returns to estimate Realized Variance:

$$RV_t = \sum_{i=1}^{m} r^2_{t+\frac{i}{m}}$$

Since we are using daily return, we don't have data for realized variance for daily data, for weekly data, $m = 5$, and for monthly data m is around 22. And due to the lack of daily data for realized variance, we used another estimator to measure relized volatility, the high-low range-based volatility estimator, as proposed by Gallant, Hsu and Tauchen (1999), Alizadeh, Brandt and Diebold(2002) and Chernov(2007):

$$Range_t = (\max_{1\leq i\leq m} r_{t+\frac{i}{m}} - \min_{1\leq i\leq m} r_{t+\frac{i}{m}})^2$$

The variable volume is the trading volume for SP500, SP100, DJIA and NASDAQ100 indexes. But for Russell2000, we used the trading volume of Russell2000 ETF. We construct the weekly volume for each index by aggregating the all the trading volumes everyday during that week, and we did the same thing to get monthly volume.

Following Bollerslev, Tauchen and Zhou(2009), we defined the variance risk premium as the difference between the risk-neutral expectation of the future return variation over the $[t, t+1]$ time interval and the realized variance over the $[t-1, t]$ time interval,

$$VRP_t = IV_t - RV_t$$

where we define the implied variance following Bollerslev et al. (2013): for example, for VIX, since it's the annualized observation, we define the weekly implied variance as

$$WeeklyIV_t = VIX_t^2 = \frac{7}{365}(VIX_t^{CBOE})^2$$

and monthly implied variance as

$$MonthlyIV_t = VIX_t^2 = \frac{30}{365}(VIX_t^{CBOE})^2$$

And since we don't have daily data for realized variance, we don't have variance risk premium for daily data either. We summarize the variable we used for each data set in the following table:

Table 22: Variables for each dataset

	Implied Volatility Index	Range	Return	Volume	RV	VRP
Daily	√	√	√	√		
Weekly	√	√	√	√	√	√
Monthly	√	√	√	√	√	√

During the empirical tests, we also used other variables such as $\sqrt{RV}$, $\sqrt{Range}$, $\ln IV$, $\ln RV$ and $\ln Volume$, but since they are all calculated from the original data, we didn't list them in the above table.

Basic summary statistics for these variables in all these 15 datasets are given through Table 24 to Table 38. The mean excess return on SP100 over the sample period from Jan 2004 to Jun 2013 is around 2.83% annualy, whereare the same for NASDAQ100 is about 7.2% annually, the return relationship between these 5 indexes is:

$$r_{SP100} < r_{SP500} < r_{DJIA} < r_{Russell2000} < r_{NASDAQ100}$$

which make sense since SP100, SP500 and DJIA are indexes for big and mature companies and Russell2000 and NASDAQ100 are for smaller and growth companies. Their volatility relationship actually doesn't follow exactly as the return rank:

$$\sigma_{DJIA} < \sigma_{SP100} < \sigma_{SP500} < \sigma_{NASDAQ100} < \sigma_{Russell2000}$$

And this relationship is consistent among different measures of volatility: option implied volatility index, realized volatility and range. For the volume, we used the same volume for SP500 and SP100, and we use ishares volume for Russell2000. So lthe volume relationship is:

$$Volume_{Russell2000} < Volume_{DJIA} < Volume_{NASDAQ100} < Volume_{SP500}$$

which means SP100 and SP500 have the highest volume and the ETF for Russell2000 has the lowest volume.

The implied volatility index and volume, for all 5 indexes and through 3 different frequencies, are highly persistent with first-order autocorrelation ranging from 0.94 to 0.99. In contrast, the serial correlation in the realized variance and range are between 0.59 to 0.78, it might because of the square form of these two variables. Return is following the efficient market hypothesis, especially for the weekly return, with the first-order autocorrelation between -0.01 and -0.07, and daily return between -0.08 to 0.12.

The sample autocorrelation between return and volume are always negative, and become stronger with the frequency from daily to monthly. But all three volatility measures are positively correlated with the volume. The return and 3 different measurements of volatility are always negatively correlated as the risk-return tradeoff. We also calculated the correlation of return and volatility among these 5 indexes in Table 39 to Table 44. From daily data to monthly data, the correlation of volatility and return among these 5 indexes actually dropped, which might indicate that the diversification among these 5 indexes will be more efficient when adjusting positions in the portfolio less frequently. However, the correlation of volume among these 5 indexes (from Table 45 to Table 47)are less correlated, which might be due to investors diversification among these 5 indexes.

2.4 Empirical Results

In this section, we do the fractional integration test for the variables we are interested in, then we test if they are fractionally cointegrated, then we use the method we proposed to get the coefficient for fractionally integrated variables, and at the end, we use the fractionally cointegration relationship identified by our coefficents to predict future return.

2.4.1 Memory Parameter d for each variable

First, we test if these three sets of variables we are interested in (return, volatility and volume) are fractionally integrated. We used "the 2 step feasible exact local whittle estimator with detrending" proposed by shimotsu (2010) to test if the memory parameter d equals to zero or not, for daily, weekly and monthly data and for all 5 indexes, respectively.

Since $\hat{d}$ follows a normal distribution:

$$\sqrt{m}(\hat{d} - d) \to \mathbf{N}(0, \frac{1}{4})$$

and we chose m as $[n^{.5}]$, $[n^{.55}]$ and $[n^{.6}]$, where the sample size for daily data is 2387, weekly 495 and monthly 114. so for each indexes, the standard deviation for $\hat{d}$ are the same, there is a set of $(3 \times 3 = 9)$ standard deviations for each variables for each indexes. We documented the estimatmion of the memory parameter d from table 48 to table 52, for all the varaibles we are interested in, even including possible transformations for these variables, such as taking square root of realized variance and range and taking logarithm of implied variance, realized variance and volume.

From the results of our estimation, we find that the excess returns, no matter it's daily, weekly or monthly, and no matter which index we use, they are almost always $I(0)$, and so does the variance risk premium. Therefore, the prediction power of variance risk premium on excess return by using monthly data as suggested by

Bollerslev, Tauchen and Zhou (2009) might come from the similar memory parameter d; it might also be why variance risk premium perform better than the different forms of volatility (implied variance, realized variance and range, as well as taking square root and logarithms of them): since volatility, as showed in our results are always fractionally integrated with a d around 0.3 to 0.7, whereas returns can not reject the null hypothesis that $d = 0$.

We also find that the daily implied volatility index (including the implied variance and the logarithm of it), for all 5 of them, are always almost $I(1)$, but not true with weekly or monthly data, which implied the highly persistency of the daily implied volatility index. We also find that the memory parameter of volume (including the logarithm of it) is always significantly from 0, sometimes even close to 1, but it varies a lot from daily to monthly, from index to index.

2.4.2 Fractional Cointegration

Here, we are interested in the relationship among return, volatility and volume, we want to find if there is any fractionally cointegrating relationship among these 3 variables for all the 5 indexes. But since we've already know that return is almost always $I(0)$, so we will only test if there is any cointegrating relationship among volatility and volume, but since we have 3 different measures of volatility (realized volatility, implied volatility and range) and several transformation of volatility, we will do a comprehensive test to see if there is any frationally cointegrating relationship between different form of volatility, and also between volatility and volume.

2.4.2.1 Same Degree of Integration

Since the first requirement for two time series to be fractionally cointegrated would be they have the same degree of integration. We used a t-statistic to test it:
If two time series x_t and y_t are fractionally integrated with degree d_1 and d_2, which means:

$$(1-L)^{d_1}x_t = u_t, \qquad (1-L)^{d_2}y_t = v_t$$

where both u_t and v_t are stationary processes. Since our two-step feasible exact whittle estimater with detrending $\hat{d}_1$ and $\hat{d}_2$ have the property as

$$\sqrt{m}(\hat{d}_1 - d_1) \sim N(0, \frac{1}{4}), \qquad \sqrt{m}(\hat{d}_2 - d_2) \sim N(0, \frac{1}{4})$$

to test the null hypothesis if $d_1 = d_2$, the t-statistic would be

$$\frac{\bar{d}_1 - \bar{d}_2}{\sqrt{\frac{1}{m}}} \sim t_{2m-2}$$

So we used this t-statistic to test if two time series have the same degree of integration. We tested for different forms of implied volatility with different forms of realized volatility and range; we also tested for different forms of volaitlity with different forms of volume for all 5 indexes with daily, weekly and monthly data.

We put the results in Table 53 to Table 57. To be more conservative, we use 1.96 as the 5% critical value for t-statistics, which means, if the t statistic is bigger than 1.96, then the null hypothesis that $d_1 = d_2$ is rejected, and we will not look further for fractionally cointegrating relationships between these two variables.

From the results we can see that sometimes different forms of volatility and volume might have the same degree of integration here and there, but it's not consistent through all the 5 indexes. Actually, we can see some evidence that volatility and volume might be fractionally cointegrated for DJIA-VXD data and NASDAQ100-VXN data, but not for SP500 and Russell2000. So it might really depends on what kind of volume data we use (not like other indexes, we use ETF data for Russell2000, and that might be the reason why there is no evidence of fractionally cointegration between volatility and volume in Russell2000). So we are not pursuing volatility-volume relationship in this chapter any more.

Actually, the only thing we found consistent with all 5 indexes is the relationship between implied volatility and realized volatility (or range) in monthly data. This fractionally cointegrating relationship between implied volatility and realized volatility in monthly data is so robust that it actually worked also in other forms: such as $ln(IV)$ and $ln(RV)$, VIX and $\sqrt{RV}$, IV and Range, VIX and $\sqrt{Range}$,we summarized them in table 58. So the next step would be to find the fractionally cointegrating vector by the searched procedure we proposed in section 2.2.2.

2.4.2.2 Coefficient for the Fractionally Cointegration

Here, we consider the fractional cointegration relationship between 5 sets of two variables (IV and RV, $ln(IV)$ and $ln(RV)$, VIX and $\sqrt{RV}$, IV and Range, VIX and $\sqrt{Range}$) for the monthly data for all 5 indexes. so $\xi' = (\xi_1, \xi_2)$ and $X = (X_1, X_2)'$, first, we record $d(X_1)$ and $d(X_2)$, when $\xi = (1, 0)$ and $\xi = (0, 1)$, then by using the searching method we proposed in section 2.2.2, we find the $(1, \xi^*)$ and record the degree of integration $d(\xi'_{\min \hat{r}^2} X_t)$. To compare the results, with other method, we also tested the degree of integration when $\xi_{unity} = (1, -1)$ and $\xi_{OLS} = (1, \hat{\beta}_1)$, where

$\hat{\beta}_1$ is the OLS estimation from the regression

$$X_1 = \beta_0 + \beta_1 X_2 + \epsilon$$

We still used the same "2 step feasible exact local whittle estimator with detrending" to estimate the degree of integration for $\xi' X$ with different ξ. We repeated this process for different number of frequency, for $m = n^{0.5}$, $m = n^{0.55}$ and $m = n^{0.6}$. And we documented the result of this whole process from table 59 to table 61.

From the results we can tell that first of all, our estimation of the coefficient of fractional cointegration through the minimization of the $\hat{r}^2$ does not really get affected by the choice of number of frequency. Second, for all five sets of variables and 5 different indexes, both our searching for minimum $\hat{r}^2$ way, and unity difference, reduce the degree of integration for both variables in most cases, which corresponds to that theory that there is a range of possible values for the cointegrating vector. But the ξ_{OLS} does not work that way, it always get a degree of integration for $d(\xi'_{OLS} X)$ is always between $d(X_1)$ and $d(X_2)$. Comparing the results for $\xi_{\min \hat{r}^2}$ and ξ_{unity}, for most of cases, $\xi_{\min \hat{r}^2}$ will get a lower degree of integration for $d(\xi' X)$. However, we still want to test if this could help to predict future return.

2.4.3 Return Forecast

Here, since ξ'_{OLS} can not really deintegrate $\xi'_{OLS} X$ to a lower degree of integration, we only compare the forecast power of $\xi'_{\min \hat{r}^2} X$ and $\xi'_{unity} X$. And we only test it for the variables $X = (IV, RV)'$ for all the 5 indexes because the economic meaning behind it.

$$\xi'_{unity} X = (1, -1)(IV, RV)' = IV - RV = VRP_{standard}$$

is actually the so-called Variance Risk Premium, which attracted a lot of research attention recently. Our estimation

$$\xi'_{\min \hat{r}^2} X = (1, \xi^*)(IV, RV)' = IV - \xi^* RV = VRP_{improved}$$

is an attempt to find the correct coefficient between implied variance and realized variance for variance risk premium, because previous research find that there is a almost always postive difference between implied variance and realized variance, but because of the scale of measurement of realized variance, it's actually hard to determine what is the correct coefficient for realized variance when estimating the variance risk premium. So our return forecast power comparison is actually between the old variance risk premium and this new variance risk premium improved by our fractional cointegration coefficient.We compared the return forecast power by two ways: static in sample way and dynamic out of sample way.

2.4.3.1 Static In-Smaple Prediction

The static in sample forecast is based on the estimation of $\xi_{\min \hat{r}^2}$ from the whole sample, and then construct the predictor time series as $IV - \xi'_{\min \hat{r}^2} RV$; then we followed Bollerslev, Tauchen and Zhou (2009)'s way to compare our prediction result by the new improved variance risk premium with the result by the standard variance risk premium from their paper.

$$\frac{1}{h} \sum_{j=1}^{h} r_{t+j} = b_0(h) + b_1(h)(VRP_{t-1}) + u_{t+h,t} \tag{7}$$

so the explanatory power, as measured by the coefficient of determination,

$$R^2(h) = \frac{Cov(\frac{1}{h} \sum_{j=1}^{h} r_{t+j}, VRP_{t-1})}{Var(\frac{1}{h} \sum_{j=1}^{h} r_{t+j}) Var(VRP_{t-1})}$$

To compare the return forecasting power of the 2 different Variance Risk Premium (the new improved Variance Risk Premium proposed by us and the standard Variance Risk Premium), we plot the results of R^2 with different horizons, from $h = 1$ to 24, for all the 5 indexes in Figure 7. From this figure we can tell that the R^2 of the newly improved variance risk premium almost always dominate R^2 for the standard variance risk premium from horizon $h = 1$ to 24, especially when $h \geq 3$. And this result is robust for all the 5 indexes we tested.

Since at this point, the results are based on the estimation of the fractional cointegration coefficient from the whole sample, and we believe that since the implied variance and realized variance have this long-term relationship, ξ^* should be quite stable. So the performance should be good if we just use this ξ^* we estimated here for future forecast. However, we still tested the out of sample performance of the new variance risk premium.

2.4.3.2 Dynamic Out-of-Sample Prediction

To see the out of sample prediction power, we did the 1 step ahead forecast dynamically: we start from $0.15T$, where T is the length of the time series, we used the first $0.15T$ data to estimate the fractionally cointegration coefficient ξ^* by minimizing the score statistic $\hat{r}^2$, then get the $VRP_{improved} = IV - \xi^* RV$, then run OLS regression between r_{t+1} and $VRP_{improved}(t)$, get $\hat{b}_0(t)$ and $\hat{b}_1(t)$ from equation 7, then for the next step, when we observe $VRP_{improved}(t+1)$, we use it to predict the future return, so the prediction is

$$E_{t+1}(r_{t+2}) = \hat{b}_0(t) + \hat{b}_1(t) VRP_{improved}(t+1)$$

Then compare the forecast $E_{t+1}(r_{t+2})$ with the true value of r_{t+2}, and do this stepwise from $t = 0.15T$ to T, so the forecast power is

$$R^2 = 1 - \frac{\sum_{t=0.15T}^{T}[r_{t+2} - E_{t+1}(r_{t+2})]^2}{\sum_{t=0.15T}^{T} r_{t+2}^2}$$

For the forecast power for the standard variance risk premium, we did the similar thing without estimate ξ^*, because we just keep it as 1. Then we documented R^2 from both ways for all 5 indexes in the following table:

Table 23: Comparison of return forecasting power for all 5 indexes

R^2	VIX	VXO	VXD	VXN	RVX
$\min \hat{r}^2$	0.7237	0.494	0.6483	0.3149	0.6957
VRP	0.6863	0.4097	0.5627	0.2976	0.3336
Ratio	0.948321	0.829352	0.867962	0.945062	0.479517

From the result we can see that the dynamic out-of-sample prediction by our "improved" variance risk premium always out perform the standard variance risk premium.

2.5 Concluding Remarks

We did a comprehensive study of the fractionally cointegrated relationship among return, volume and volatility, by using 5 indexes (SP500, SP100, DJIA, NASDAQ100 and Russell2000) for daily, weekly and monthly data, we also tried different measures of volatility (implied volatility, realized volatility and range). From the results we can tell that the time series of index return is always stationary, whereas the degree of various form of volatility and volume is always between 0.3 and 0.7. However, there is not significant evidence that volatility and volume are fractionally cointegrated.

The only robust fractional cointegration relationship we found among these data are between different measurements of implied volatility and realized volatility for monthly data. We developed our own way to estimate the frational cointegration coefficient by minimizing the score test statistic $\hat{r}^2$ proposed by Robinson (1995). By this way, we also improved the standard measurement of variance risk premium and then compared the return forecast power on these two measurement of variance risk premium. And our results showed that the prediction by our measurement of variance risk premium almost always outperform the standard measurement of variance risk premium, both in-sample statically and out-of-sample dynamically.

For the future study, we believe that our way to estimate the fractional cointegration coefficient could be used in pair trading when determining the relative holding position in this pair.

Table 24: Summary statistics for daily SP500-VIX data

	Ex. Return	VIX	Range	Volume
Mean	0.015397	20.48553	3.188024	3.54E+09
Std.Dev	1.313964	10.05466	8.412467	1.58E+09
Skewness	-0.31635	2.258896	8.162741	0.722556
Kurtosis	13.62777	9.570921	86.38276	3.733853
AR(1)	-0.113051	0.99658	0.6874	0.98293
Return	1			
VIX	-0.13311	1		
Range	-0.07866	0.661246	1	
Volume	-0.03149	0.701753	0.465645	1

Table 25: Summary statistics for weekly SP500-VIX data

	Ex. Return	VIX	Range	Volume	RV	VRP
Mean	0.074936	20.38422	17.55966	1.71E+10	8.330006	1.628528
Std.Dev	2.592809	10.19607	45.2569	7.53E+09	19.4987	11.15739
Skewness	-0.94518	2.33447	9.621347	0.628409	6.194703	-6.73895
Kurtosis	11.99928	10.18902	127.9165	3.320049	50.99497	67.4479
AR(1)	-0.061019	0.98905	0.67246	0.97698	0.74514	0.04325
Return	1					
VIX	-0.26631	1				
Range	-0.31289	0.707237	1			
Volume	-0.10107	0.713661	0.469533	1		
RV	-0.17488	0.767291	0.78374	0.500688	1	
VRP	-0.02714	-0.20969	-0.4283	-0.1542	-0.75969	1

Table 26: Summary statistics for monthly SP500-VIX data

	Ex. Return	VIX	Range	Volume	RV	VRP
Mean	0.32266	20.48912	72.745	7.43E+10	36.1706	5.421871
Std.Dev	4.342834	9.327422	148.23	3.06E+10	72.53436	37.67888
Skewness	-1.05285	1.857956	4.967921	0.451539	4.947516	-4.57124
Kurtosis	5.752631	6.882143	30.75245	2.871369	31.8925	31.96977
AR(1)	0.23191	0.97749	0.67459	0.98475	0.78276	0.42985
Return	1					
VIX	-0.47087	1				
Range	-0.44033	0.798796	1			
Volume	-0.20536	0.803232	0.564535	1		
RV	-0.51087	0.813386	0.936159	0.587108	1	
VRP	0.354421	-0.381	-0.72087	-0.25069	-0.83512	1

Table 27: Summary statistics for daily SP100-VXO data

	Ex. Return	VXO	Range	Volume
Mean	0.01194	20.27243	3.007663	3.54E+09
Std.Dev	1.267834	10.64045	8.183512	1.58E+09
Skewness	-0.26201	2.376938	8.631755	0.725158
Kurtosis	13.98328	10.36119	97.44118	3.737536
AR(1)	-0.123306	0.99571	0.69327	0.98329
Return	1			
VXO	-0.14144	1		
Range	-0.06771	0.66249	1	
Volume	-0.03242	0.691929	0.462067	1

Table 28: Summary statistics for weekly SP100-VXO data

	Ex. Return	VXO	Range	Volume	RV	VRP
Mean	0.054588	20.36543	16.24501	1.71E+10	7.757294	2.467179
Std.Dev	2.480214	10.89137	43.52697	7.55E+09	18.27395	10.04207
Skewness	-0.95499	2.445295	10.69109	0.636998	6.123977	-5.48126
Kurtosis	12.67049	11.05924	156.5459	3.353797	48.97961	55.37069
AR(1)	-0.071335	0.98704	0.62025	0.97676	0.74857	-0.148289
Return	1					
VXO	-0.28457	1				
Range	-0.32183	0.710473	1			
Volume	-0.09992	0.702335	0.46103	1		
RV	-0.16111	0.77115	0.762539	0.501036	1	
VRP	-0.16423	-0.00269	-0.17313	-0.03325	-0.58822	1

Table 29: Summary statistics for monthly SP100-VXO data

	Ex. Return	VXO	Range	Volume	RV	VRP
Mean	0.235769	20.30298	68.2362	7.43E+10	33.68317	8.226233
Std.Dev	4.151149	9.92729	137.325	3.06E+10	67.97647	32.29908
Skewness	-0.91382	1.951168	5.048676	0.459463	5.015302	-3.67462
Kurtosis	4.767258	7.268549	31.96599	2.884506	32.60511	29.75585
AR(1)	0.22227	0.97428	0.67103	0.98437	0.78382	0.27413
Return	1					
VXO	-0.48286	1				
Range	-0.42553	0.802812	1			
Volume	-0.1845	0.780293	0.565754	1		
RV	-0.48264	0.816314	0.936266	0.582031	1	
VRP	0.208855	-0.20505	-0.58161	-0.13952	-0.70959	1

Table 30: Summary statistics for daily DJIA-VXD data

	Ex. Return	VXD	Range	Volume
Mean	0.015648	18.68246	2.973224	2.21E+08
Std.Dev	1.202541	9.008344	8.203976	84926194
Skewness	-0.06801	2.355877	9.388949	1.244964
Kurtosis	13.67838	10.16727	117.315	5.92239
AR(1)	-0.110053	0.99654	0.69001	0.96179
Return	1			
VXD	-0.12972	1		
Range	-0.03429	0.62555	1	
Volume	-0.0446	0.352357	0.385386	1

Table 31: Summary statistics for weekly DJIA-VXD data

	Ex. Return	VXD	Range	Volume	RV	VRP
Mean	0.072576	18.82889	15.48024	1.07E+09	6.973678	1.421818
Std.Dev	2.428721	9.13272	42.08829	3.7E+08	16.29313	10.76802
Skewness	-1.01046	2.293158	11.39926	0.876426	6.594742	-8.88362
Kurtosis	13.38716	9.564089	174.7235	4.688911	58.66029	128.0126
AR(1)	-0.072195	0.98881	0.59985	0.96872	0.75743	0.45208
Return	1					
VXD	-0.04959	1				
Range	-0.30962	0.55446	1			
Volume	-0.11585	0.379579	0.357173	1		
RV	-0.13339	0.707603	0.767971	0.406747	1	
VRP	0.157325	-0.11785	-0.58552	-0.22529	-0.76294	1

Table 32: Summary statistics for monthly DJIA-VXD data

	Ex. Return	VXD	Range	Volume	RV	VRP
Mean	0.311428	18.63965	62.90537	4.63E+09	30.28201	4.11469
Std.Dev	4.013598	8.466726	125.7395	1.38E+09	61.57219	44.76955
Skewness	-0.90628	2.064496	5.492772	0.808422	5.519517	-5.60347
Kurtosis	5.09417	8.086338	38.25379	4.851008	39.34235	46.16755
AR(1)	0.1763	0.97521	0.62885	0.97962	0.74656	0.43566
Return	1					
VXD	-0.13758	1				
Range	-0.39995	0.65402	1			
Volume	-0.23961	0.365292	0.452119	1		
RV	-0.45182	0.674941	0.945818	0.452826	1	
VRP	0.500886	-0.07828	-0.72878	-0.27316	-0.77093	1

Table 33: Summary statistics for daily NASDAQ100-VXN data

	Ex. Return	VXN	Range	Volume
Mean	0.028638	23.16799	3.617515	2E+09
Std.Dev	1.432362	9.093661	7.798821	1.46E+09
Skewness	-0.12269	2.441344	8.560389	42.89744
Kurtosis	10.42264	11.03018	100.9107	2006.667
AR(1)	-0.089144	0.9977	0.65777	0.66988
Return	1			
VXN	-0.12432	1		
Range	-0.08689	0.647934	1	
Volume	-0.03934	0.093649	0.112595	1

Table 34: Summary statistics for weekly NASDAQ100-VXN data

	Ex. Return	VXN	Range	Volume	RV	VRP
Mean	0.138816	23.41281	21.0073	9.63E+09	9.901493	2.247437
Std.Dev	2.865919	9.246246	37.97679	3.6E+09	19.98787	13.69058
Skewness	-0.50014	2.333065	6.321876	13.19802	7.817407	-10.0511
Kurtosis	5.724656	10.01252	53.79559	246.7741	84.82735	138.9452
AR(1)	-0.010531	0.99095	0.7755	0.89481	0.61497	0.13951
Return	1					
VXN	-0.04076	1				
Range	-0.22673	0.658163	1			
Volume	-0.11512	0.159203	0.213505	1		
RV	-0.13457	0.688246	0.800931	0.17989	1	
VRP	0.162262	-0.11919	-0.52715	-0.12986	-0.78635	1

Table 35: Summary statistics for monthly NASDAQ100-VXN data

	Ex. Return	VXN	Range	Volume	RV	VRP
Mean	0.600148	23.13368	94.34052	4.18E+10	42.9941	6.793093
Std.Dev	5.39977	8.43809	148.9011	9.07E+09	68.58144	36.27729
Skewness	-0.69082	1.994339	4.402492	4.130837	4.916042	-4.06816
Kurtosis	3.950406	7.739138	24.78702	30.9811	32.09293	29.11609
AR(1)	0.17361	0.9802	0.7307	0.96473	0.78738	0.31664
Return	1					
VXN	-0.40899	1				
Range	-0.44192	0.7879	1			
Volume	-0.2113	0.321488	0.318342	1		
RV	-0.48475	0.814243	0.912437	0.322961	1	
VRP	0.381139	-0.33447	-0.67428	-0.24793	-0.80695	1

Table 36: Summary statistics for daily Russell2000-RVX data

	Ex. Return	RVX	Range	Volume
Mean	0.024353	26.83433	4.71745	54475269
Std.Dev	1.703458	10.5987	9.847043	35593837
Skewness	-0.33802	2.073579	7.577981	1.851465
Kurtosis	8.068723	8.268848	82.6967	10.38643
AR(1)	-0.095064	0.99803	0.66159	0.94172
Return	1			
RVX	-0.11502	1		
Range	-0.06067	0.65322	1	
Volume	-0.06801	0.477446	0.478304	1

Table 37: Summary statistics for weekly Russell2000-RVX data

	Ex. Return	RVX	Range	Volume	RV	VRP
Mean	0.112227	27.09234	28.71136	2.63E+08	13.98715	2.336138
Std.Dev	3.390251	10.83443	56.98056	1.6E+08	25.64745	15.98082
Skewness	-0.45244	2.05861	7.312416	1.483982	5.137575	-5.52669
Kurtosis	6.218215	8.033777	75.84479	7.349192	35.35512	44.40461
AR(1)	-0.044926	0.99138	0.71384	0.94765	0.73236	0.32489
Return	1					
RVX	-0.03122	1				
Range	-0.24253	0.654351	1			
Volume	-0.17646	0.477619	0.468598	1		
RV	-0.14814	0.75567	0.781207	0.49098	1	
VRP	0.218485	-0.21855	-0.54065	-0.36175	-0.78879	1

Table 38: Summary statistics for monthly Russell2000-RVX data

	Ex. Return	RVX	Range	Volume	RV	VRP
Mean	0.493486	26.98333	125.6372	1.14E+09	60.73803	8.257451
Std.Dev	5.840154	10.59859	239.3675	6.29E+08	93.85868	64.97962
Skewness	-0.82046	2.062968	5.254285	0.902631	3.741556	-5.00612
Kurtosis	4.817933	8.086289	33.87714	4.029912	18.61797	34.75251
AR(1)	0.12158	0.97755	0.67053	0.95437	0.79434	0.35879
Return	1					
RVX	-0.04937	1				
Range	-0.40155	0.582048	1			
Volume	-0.33722	0.409721	0.42574	1		
RV	-0.38576	0.711932	0.904089	0.527236	1	
VRP	0.532615	-0.01303	-0.70957	-0.40404	-0.69193	1

Table 39: Correlation between daily implied volatility

	VIX	VXO	VXD	VXN	RVX
VIX	1				
VXO	0.994503	1			
VXD	0.99727	0.994015	1		
VXN	0.972489	0.977028	0.97462	1	
RVX	0.98311	0.973754	0.980186	0.96611	1

Table 40: Correlation between weekly implied volatility

	VIX	VXO	VXD	VXN	RVX
VIX	1				
VXO	0.994145	1			
VXD	0.941993	0.93805	1		
VXN	0.922487	0.925001	0.974576	1	
RVX	0.923129	0.911529	0.979849	0.966882	1

Table 41: Correlation between monthly implied volatility

	VIX	VXO	VXD	VXN	RVX
VIX	1				
VXO	0.994446	1			
VXD	0.878027	0.864372	1		
VXN	0.969374	0.975851	0.838792	1	
RVX	0.811369	0.791327	0.965763	0.773064	1

Table 42: Correlation between daily return

	SP500	SP100	DJIA	NASDAQ100	Russell2000
SP500	1				
SP100	0.995238	1			
DJIA	0.983111	0.986371	1		
NASDAQ100	0.922431	0.913956	0.893549	1	
Russell2000	0.927738	0.906342	0.895118	0.885807	1

Table 43: Correlation between weekly return

	SP500	SP100	DJIA	NASDAQ100	Russell2000
SP500	1				
SP100	0.993484	1			
DJIA	0.978033	0.983357	1		
NASDAQ100	0.907311	0.893494	0.872302	1	
Russell2000	0.933028	0.904838	0.887318	0.893846	1

Table 44: Correlation between monthly return

	SP500	SP100	DJIA	NASDAQ100	Russell2000
SP500	1				
SP100	0.989752	1			
DJIA	0.972904	0.981285	1		
NASDAQ100	0.895705	0.875377	0.835526	1	
Russell2000	0.930485	0.894003	0.874625	0.871615	1

Table 45: Corr between daily volume

	SP500	DJIA	N100	R2000
SP500	1			
DJIA	0.33215	1		
N100	0.172919	0.160759	1	
R2000	0.655507	0.374464	0.189114	1

Table 46: Corr between weekly volume

	SP500	DJIA	N100	R2000
SP500	1			
DJIA	0.349622	1		
N100	0.35108	0.304612	1	
R2000	0.700175	0.403072	0.360424	1

Table 47: Corr between monthly volume

	SP500	DJIA	N100	R2000
SP500	1			
DJIA	0.25554	1		
N100	0.48632	0.344026	1	
R2000	0.715911	0.339159	0.512349	1

Table 48: Fractional integration test for SP500-VIX data

	Daily			Weekly			Monthly		
	$m = n^{0.5}$	$m = n^{0.55}$	$m = n^{0.6}$	$m = n^{0.5}$	$m = n^{0.55}$	$m = n^{0.6}$	$m = n^{0.5}$	$m = n^{0.55}$	$m = n^{0.6}$
IV	0.871	0.957345	0.968464	0.623752	0.537907	0.6438	0.490848	0.548288	0.619079
Std	0.072169	0.058926	0.048564	0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
RV				0.396238	0.39361	0.550635	0.339813	0.38584	0.420753
Std				0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
Range	0.666563	0.562819	0.547064	0.34153	0.315104	0.434639	0.338574	0.374537	0.457022
Std	0.072169	0.058926	0.048564	0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
Return	0.095123	0.030598	0.023371	0.36543	-0.03407	0.022181	0.012913	0.084224	0.224852
Std	0.072169	0.058926	0.048564	0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
$\sqrt{RV}$				0.590341	0.534949	0.610847	0.494192	0.534222	0.540497
Std				0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
$\sqrt{Range}$	0.700944	0.620232	0.65102	0.553999	0.484754	0.580081	0.450287	0.495165	0.583686
Std	0.072169	0.058926	0.048564	0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
VIX	0.876165	0.908418	0.996039	0.778061	0.649862	0.728937	0.594331	0.632383	0.65162
Std	0.072169	0.058926	0.048564	0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
VRP				-0.03945	0.08488	0.312142	0.008723	0.050416	0.087217
Std				0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
Volume	0.633842	0.605349	0.577143	0.694153	0.681763	0.635072	1.04474	0.983367	0.804471
Std	0.072169	0.058926	0.048564	0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
lnIV	0.805947	0.821079	0.931592	0.846416	0.725534	0.761509	0.697789	0.7176	0.686525
Std	0.072169	0.058926	0.048564	0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
lnRV				0.660336	0.598018	0.544875	0.577114	0.622222	0.556472
Std				0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
lnVolume	0.63838	0.589587	0.551779	0.770458	0.729394	0.62279	1.118447	1.078007	0.91596
Std	0.072169	0.058926	0.048564	0.1066	0.091287	0.078087	0.158114	0.138675	0.121268

Table 49: Fractional integration test for SP100-VXO data

	Daily			Weekly			Monthly		
	$m=n^{0.5}$	$m=n^{0.55}$	$m=n^{0.6}$	$m=n^{0.5}$	$m=n^{0.55}$	$m=n^{0.6}$	$m=n^{0.5}$	$m=n^{0.55}$	$m=n^{0.6}$
IV	0.915155	0.886422	0.933258	0.601523	0.527043	0.663921	0.477861	0.555983	0.632503
Std	0.072169	0.058926	0.048564	0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
RV				0.39483	0.389359	0.545799	0.332521	0.384345	0.422779
Std				0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
Range	0.661579	0.56414	0.544926	0.345108	0.3134	0.427463	0.343635	0.378524	0.472625
Std	0.072169	0.058926	0.048564	0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
Return	0.073195	0.025953	0.019484	0.356347	-0.04303	0.012651	0.015955	0.091505	0.238359
Std	0.072169	0.058926	0.048564	0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
$\sqrt{RV}$				0.599605	0.526616	0.601298	0.493975	0.549027	0.55262
Std				0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
$\sqrt{Range}$	0.698532	0.61475	0.653715	0.571603	0.486364	0.574651	0.466483	0.507685	0.604786
Std	0.072169	0.058926	0.048564	0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
VXO	0.888981	0.87641	0.98628	0.773961	0.649299	0.744566	0.577037	0.647508	0.667407
Std	0.072169	0.058926	0.048564	0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
VRP				-0.09038	0.046495	0.258318	-0.22531	-0.16114	-0.07632
Std				0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
Volume	0.630907	0.603541	0.580396	0.691264	0.676924	0.629691	1.050605	0.980026	0.803111
Std	0.072169	0.058926	0.048564	0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
lnIV	0.803715	0.807137	0.931623	0.84507	0.729231	0.766668	0.66774	0.735054	0.696763
Std	0.072169	0.058926	0.048564	0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
lnRV				0.668289	0.57531	0.526616	0.592469	0.655958	0.566914
Std				0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
lnVolume	0.636352	0.588319	0.553581	0.768545	0.722486	0.616247	1.124435	1.073439	0.913356
Std	0.072169	0.058926	0.048564	0.1066	0.091287	0.078087	0.158114	0.138675	0.121268

Table 50: Fractional integration test for DJIA-VXD data

	Daily			Weekly			Monthly		
	$m = n^{0.5}$	$m = n^{0.55}$	$m = n^{0.6}$	$m = n^{0.5}$	$m = n^{0.55}$	$m = n^{0.6}$	$m = n^{0.5}$	$m = n^{0.55}$	$m = n^{0.6}$
IV	0.870776	1.005474	0.982677	0.731147	0.625839	0.747737	0.493587	0.56169	0.689927
Std	0.072169	0.058926	0.048564	0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
RV				0.382657	0.357461	0.50102	0.340715	0.38493	0.410718
Std				0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
Range	0.616789	0.561484	0.508345	0.336528	0.289732	0.389279	0.372725	0.404266	0.475653
Std	0.072169	0.058926	0.048564	0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
Return	0.052128	-0.00018	0.000863	0.376505	-0.05652	-0.01153	0.048453	0.211768	0.301509
Std	0.072169	0.058926	0.048564	0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
$\sqrt{RV}$				0.59687	0.504969	0.584259	0.498488	0.554111	0.547711
Std				0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
$\sqrt{Range}$	0.698074	0.630152	0.638052	0.577219	0.462963	0.551568	0.496882	0.537265	0.598947
Std	0.072169	0.058926	0.048564	0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
VXD	0.881338	0.925716	0.998829	0.873897	0.730572	0.8173	0.608142	0.664961	0.745303
Std	0.072169	0.058926	0.048564	0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
VRP				-0.17533	-0.09319	0.103959	-0.21432	-0.16432	-0.10811
Std				0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
Volume	0.575736	0.52101	0.51978	0.595535	0.495706	0.527047	0.505919	0.668365	0.575962
Std	0.072169	0.058926	0.048564	0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
lnIV	0.803551	0.814382	0.916711	0.912351	0.789145	0.803794	0.712898	0.763808	0.793778
Std	0.072169	0.058926	0.048564	0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
lnRV				0.648076	0.557423	0.521756	0.555927	0.635951	0.562548
Std				0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
lnVolume	0.560944	0.51601	0.488469	0.653657	0.512848	0.49625	0.612252	0.75329	0.674087
Std	0.072169	0.058926	0.048564	0.1066	0.091287	0.078087	0.158114	0.138675	0.121268

Table 51: Fractional integration test for NASDAQ100-VXN data

	Daily			Weekly			Monthly		
	$m = n^{0.5}$	$m = n^{0.55}$	$m = n^{0.6}$	$m = n^{0.5}$	$m = n^{0.55}$	$m = n^{0.6}$	$m = n^{0.5}$	$m = n^{0.55}$	$m = n^{0.6}$
IV	0.820823	0.86455	0.88977	0.709361	0.598788	0.712552	0.486015	0.550537	0.616184
Std	0.072169	0.058926	0.048564	0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
RV				0.38626	0.380142	0.496803	0.359044	0.383899	0.427252
Std				0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
Range	0.622257	0.544413	0.557456	0.335962	0.312289	0.397092	0.375879	0.320363	0.352358
Std	0.072169	0.058926	0.048564	0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
Return	0.053199	0.02275	0.037907	0.001127	-0.0633	0.032199	-0.21066	-0.2483	-0.10708
Std	0.072169	0.058926	0.048564	0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
$\sqrt{RV}$				0.58371	0.509609	0.55475	0.497155	0.529145	0.544442
Std				0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
$\sqrt{Range}$	0.658899	0.579106	0.61419	0.499996	0.466163	0.509975	0.449915	0.444652	0.472163
Std	0.072169	0.058926	0.048564	0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
VXN	0.827204	0.830392	0.919597	0.831701	0.691793	0.768975	0.56858	0.635445	0.646855
Std	0.072169	0.058926	0.048564	0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
VRP				-0.13157	-0.04384	0.088627	0.001459	0.034898	0.082913
Std				0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
Volume	0.239801	0.18873	0.173667	0.374454	0.291894	0.246717	0.255037	0.341037	0.366956
Std	0.072169	0.058926	0.048564	0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
lnIV	0.78681	0.76793	0.870184	0.85643	0.74946	0.777937	0.644332	0.721938	0.666885
Std	0.072169	0.058926	0.048564	0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
lnRV				0.665909	0.552742	0.515645	0.545805	0.627934	0.548425
Std				0.1066	0.091287	0.078087	0.158114	0.138675	0.121268
lnVolume	0.370105	0.401783	0.364126	0.439192	0.373265	0.312602	0.347203	0.425561	0.476911
Std	0.072169	0.058926	0.048564	0.1066	0.091287	0.078087	0.158114	0.138675	0.121268

Table 52: Fractional integration test for Russell2000-RVX data

	Daily			Weekly			Monthly		
	$m = n^{0.5}$	$m = n^{0.55}$	$m = n^{0.6}$	$m = n^{0.5}$	$m = n^{0.55}$	$m = n^{0.6}$	$m = n^{0.5}$	$m = n^{0.55}$	$m = n^{0.6}$
IV	0.955985	0.91285	0.93607	0.716443	0.67796	0.83853	0.406843	0.466216	0.56989
Std	0.072169	0.058926	0.048566	0.1066	0.091287	0.078087	0.158114	0.138675	0.121265
RV				0.487526	0.502183	0.65483	0.35369	0.4227	0.47657
Std				0.1066	0.091287	0.078087	0.158114	0.138675	0.121265
Range	0.755233	0.498183	0.50775	0.373291	0.35755	0.51894	0.272418	0.309764	0.39378
Std	0.072169	0.058926	0.048566	0.1066	0.091287	0.078087	0.158114	0.138675	0.121265
Return	0.018738	-0.03508	-0.01063	0.135131	-0.10767	-0.04639	-0.09264	-0.13023	-0.03232
Std	0.072169	0.058926	0.048566	0.1066	0.091287	0.078087	0.158114	0.138675	0.121265
$\sqrt{RV}$				0.621291	0.626544	0.64998	0.450276	0.498197	0.52304
Std				0.1066	0.091287	0.078087	0.158114	0.138675	0.121265
$\sqrt{Range}$	0.705559	0.567625	0.60647	0.548801	0.517754	0.60682	0.345469	0.403444	0.53977
Std	0.072169	0.058926	0.048566	0.1066	0.091287	0.078087	0.158114	0.138675	0.121265
RVX	0.927438	0.893486	0.9853	0.797572	0.756789	0.8672	0.501712	0.541654	0.6085
Std	0.072169	0.058926	0.048566	0.1066	0.091287	0.078087	0.158114	0.138675	0.121265
VRP				-0.00139	0.105584	0.2352	-0.15174	-0.10829	-0.03979
Std				0.1066	0.091287	0.078087	0.158114	0.138675	0.121265
Volume	0.555809	0.54611	0.5714	0.666897	0.66308	0.63607	0.639148	0.742566	0.63319
Std	0.072169	0.058926	0.048566	0.1066	0.091287	0.078087	0.158114	0.138675	0.121265
lnIV	0.847108	0.825778	0.93796	0.823469	0.797693	0.84079	0.581689	0.608079	0.64626
Std	0.072169	0.058926	0.048566	0.1066	0.091287	0.078087	0.158114	0.138675	0.121265
lnRV				0.71927	0.731178	0.56636	0.505328	0.547455	0.5286
Std				0.1066	0.091287	0.078087	0.158114	0.138675	0.121265
lnVolume	0.632694	0.626107	0.59717	0.844125	0.772443	0.73352	0.848259	0.923797	0.82682
Std	0.072169	0.058926	0.048566	0.1066	0.091287	0.078087	0.158114	0.138675	0.121265

Table 53: T statistics for integration in SP500-VIX data

		$m = n^{0.5}$			$m = n^{0.55}$			$m = n^{0.6}$		
		Daily	Weekly	Monthly	Daily	Weekly	Monthly	Daily	Weekly	Monthly
$IV \sim RV$	d(IV)=d(RV)		2.134277	0.955227		1.580696	1.171429		1.193095	1.635437
	d(lnIV)=d(lnRV)		1.745592	0.763212		1.396866	0.687785		2.774271	1.072446
	d(VIX)=d($\sqrt{RV}$)		1.760965	0.633333		1.258802	0.707853		1.512287	0.91635
	d(IV)=d(Range)	2.832765	2.647477	0.963063	6.695333	2.440689	1.252932	8.67717	2.678569	1.336353
	d(VIX)=d($\sqrt{Range}$)	2.427928	2.101883	0.911016	4.890677	1.808665	0.989491	7.104369	1.906282	0.560199
$V \sim Volume$	d(IV)=d(Volume)	3.286157	0.660411	3.503122	5.973569	1.575862	3.137403	8.057799	0.111774	1.528781
	d(VIX)=d(Volume)	3.357724	0.787128	2.848638	5.143244	0.349464	2.530983	8.625599	1.202055	1.260436
	d(lnIV)=d(Volume)	2.321876	0.712555	2.660475	3.928556	0.04229	2.598932	7.82082	1.776468	1.89197
	d(RV)=d(Volume)		2.794688	4.458349		3.156558	4.308832		1.081321	3.164219
	d($\sqrt{RV}$)=d(Volume)		0.973837	3.481971		1.608267	3.238836		0.310232	2.176786
	d(lnRV)=d(Volume)		0.317229	2.957524		0.917383	2.604258		1.155083	2.045048
	d(Range)=d(Volume)	0.453392	3.307889	4.466185	0.721763	4.016551	4.390335	0.619371	2.566795	2.865135
	d($\sqrt{Range}$)=d(Volume)	0.929796	1.314755	3.759654	0.252567	2.158129	3.520474	1.521229	0.704227	1.820635
$V \sim lnVolume$	d(IV)=d(lnVolume)	3.22328	1.376218	3.969285	6.241061	2.097634	3.819862	8.580064	0.269054	2.448149
	d(VIX)=d(lnVolume)	3.294848	0.071321	3.314801	5.410735	0.871237	3.213442	9.147864	1.359335	2.179804
	d(lnIV)=d(lnVolume)	2.321876	0.712555	2.660475	3.928556	0.04229	2.598932	7.82082	1.776468	1.89197
	d(RV)=d(lnVolume)		3.510495	4.924513		3.678331	4.991292		0.924042	4.083587
	d($\sqrt{RV}$)=d(lnVolume)		1.689644	3.948134		2.130039	3.921295		0.152953	3.096154
	d(lnRV)=d(lnVolume)		1.033037	3.423687		1.439156	3.286718		0.997804	2.964416
	d(Range)=d(lnVolume)	0.390516	4.023696	4.932349	0.454272	4.538323	5.072795	0.097107	2.409515	3.784503
	d($\sqrt{Range}$)=d(lnVolume)	0.86692	2.030562	4.225817	0.520058	2.679902	4.202933	2.043494	0.546947	2.740003

Table 54: T statistics for integration in SP100-VXO data

		$m = n^{0.5}$			$m = n^{0.55}$			$m = n^{0.6}$		
		Daily	Weekly	Monthly	Daily	Weekly	Monthly	Daily	Weekly	Monthly
$IV \sim RV$	d(IV)=d(RV)		1.938949	0.919208		1.508256	1.237702		1.512698	1.729432
	d(lnIV)=d(lnRV)		1.658351	0.476053		1.686117	0.570375		3.074157	1.070762
	d(VIX)=d($\sqrt{RV}$)		1.635605	0.525329		1.343926	0.710158		1.834717	0.946558
	d(IV)=d(Range)	3.513656	2.405387	0.84892	5.469308	2.340341	1.279675	7.996235	3.02813	1.31839
	d(VIX)=d($\sqrt{Range}$)	2.638936	1.89829	0.699202	4.440521	1.78486	1.008276	6.847936	2.17597	0.51638
$V \sim Volume$	d(IV)=d(Volume)	3.938661	0.841842	3.622354	4.800635	1.64186	3.057816	7.265863	0.438348	1.406868
	d(VIX)=d(Volume)	3.575977	0.775772	2.995113	4.63074	0.302619	2.397822	8.357662	1.471107	1.119049
	d(lnIV)=d(Volume)	2.319049	0.717864	2.888395	3.713455	0.073887	2.440126	7.784346	1.926318	1.786074
	d(RV)=d(Volume)		2.78079	4.541562		3.150116	4.295518		1.07435	3.1363
	d($\sqrt{RV}$)=d(Volume)		0.859833	3.520442		1.646545	3.107981		0.36361	2.065607
	d(lnRV)=d(Volume)		0.215519	2.897509		1.113127	2.336892		1.320006	1.947734
	d(Range)=d(Volume)	0.425005	3.247229	4.471275	0.668672	3.982202	4.337492	0.730372	2.589782	2.725259
	d($\sqrt{Range}$)=d(Volume)	0.937041	1.122518	3.694315	0.19022	2.087479	3.406098	1.509726	0.704863	1.635429
$V \sim lnVolume$	d(IV)=d(lnVolume)	3.863214	1.56681	4.089293	5.058962	2.140964	3.731425	7.81801	0.610514	2.315969
	d(VIX)=d(lnVolume)	3.50053	0.050804	3.462051	4.889068	0.801722	3.071431	8.90981	1.643273	2.02815
	d(lnIV)=d(lnVolume)	2.319049	0.717864	2.888395	3.713455	0.073887	2.440126	7.784346	1.926318	1.786074
	d(RV)=d(lnVolume)		3.505758	5.0085		3.649219	4.969127		0.902184	4.045401
	d($\sqrt{RV}$)=d(lnVolume)		1.584802	3.987381		2.145648	3.781589		0.191443	2.974708
	d(lnRV)=d(lnVolume)		0.940488	3.364448		1.61223	3.010501		1.14784	2.856835
	d(Range)=d(lnVolume)	0.349559	3.972197	4.938213	0.410345	4.481305	5.011101	0.178225	2.417616	3.63436
	d($\sqrt{Range}$)=d(lnVolume)	0.861594	1.847487	4.161254	0.448547	2.586583	4.079707	2.061873	0.532696	2.54453

Table 55: T statistics for integration in DJIA-VXD data

		$m = n^{0.5}$			$m = n^{0.55}$			$m = n^{0.6}$		
		Daily	Weekly	Monthly	Daily	Weekly	Monthly	Daily	Weekly	Monthly
$IV \sim RV$	d(IV)=d(RV)		3.269125	0.966845		2.939931	1.274638		3.159518	2.302414
	d(lnIV)=d(lnRV)		2.47912	0.99277		2.53838	0.921988		3.61185	1.906769
	d(VIX)=d($\sqrt{RV}$)		2.598745	0.69351		2.471359	0.799349		2.984374	1.629391
	d(IV)=d(Range)	3.519357	3.701849	0.764397	7.534766	3.681865	1.135206	9.767109	4.590499	1.766948
	d(VIX)=d($\sqrt{Range}$)	2.539383	2.783087	0.703669	5.015891	2.931511	0.920832	7.428857	3.403033	1.206882
$V \sim Volume$	d(IV)=d(Volume)	4.088204	1.272149	0.077994	8.221623	1.425529	0.769239	9.531648	2.82621	0.939777
	d(VIX)=d(Volume)	4.234554	2.611266	0.646513	6.868081	2.572826	0.024543	9.864232	3.717047	1.396424
	d(lnIV)=d(Volume)	3.361664	2.426767	0.63654	5.06355	3.026677	0.075845	8.818044	3.938484	0.986995
	d(RV)=d(Volume)		1.996976	1.044839		1.514403	2.043877		0.333308	1.362637
	d($\sqrt{RV}$)=d(Volume)		0.012521	0.046997		0.101467	0.823892		0.732673	0.232967
	d(lnRV)=d(Volume)		0.492873	0.316282		0.676076	0.233736		0.067763	0.110616
	d(Range)=d(Volume)	0.568847	2.4297	0.84239	0.686857	2.256336	1.904445	0.23546	1.764289	0.827171
	d($\sqrt{Range}$)=d(Volume)	1.695171	0.171822	0.057157	1.85219	0.358686	0.945374	2.435375	0.314014	0.189541
$V \sim lnVolume$	d(IV)=d(lnVolume)	4.293167	0.726923	0.750506	8.306491	1.237749	1.381646	10.17638	3.220607	0.13062
	d(VIX)=d(lnVolume)	4.439516	2.066039	0.025999	6.952949	2.385046	0.63695	10.50896	4.111444	0.587266
	d(lnIV)=d(lnVolume)	3.361664	2.426767	0.63654	5.06355	3.026677	0.075845	8.818044	3.938484	0.986995
	d(RV)=d(lnVolume)		2.542202	1.717351		1.702182	2.656284		0.061089	2.171795
	d($\sqrt{RV}$)=d(lnVolume)		0.532706	0.719509		0.086313	1.436299		1.12707	1.042125
	d(lnRV)=d(lnVolume)		0.052353	0.35623		0.488297	0.846143		0.326634	0.919774
	d(Range)=d(lnVolume)	0.773809	2.974926	1.514902	0.771725	2.444116	2.516852	0.409267	1.369892	1.636328
	d($\sqrt{Range}$)=d(lnVolume)	1.900133	0.717048	0.729669	1.937058	0.546465	1.557781	3.080102	0.708411	0.619616

Table 56: T statistics for integration in NASDAQ100-VXN data

		$m = n^{0.5}$			$m = n^{0.55}$			$m = n^{0.6}$		
		Daily	Weekly	Monthly	Daily	Weekly	Monthly	Daily	Weekly	Monthly
$IV \sim RV$	d(IV)=d(RV)		3.030959	0.803031		2.395152	1.201644		2.762938	1.557973
	d(lnIV)=d(lnRV)		1.787239	0.62314		2.154942	0.677874		3.358973	0.976842
	d(VIX)=d($\sqrt{RV}$)		2.326362	0.451732		1.995729	0.766545		2.743419	0.844515
	d(IV)=d(Range)	2.751405	3.502797	0.696561	5.432909	3.138441	1.659809	6.842771	4.039857	2.175568
	d(VIX)=d($\sqrt{Range}$)	2.332092	3.111672	0.750508	4.264465	2.471648	1.375829	6.288708	3.316819	1.440541
$V \sim Volume$	d(IV)=d(Volume)	8.05088	3.141707	1.46083	11.46905	3.361852	1.510731	14.74546	5.965595	2.05519
	d(VIX)=d(Volume)	8.139294	4.289355	1.983021	10.88936	4.380669	2.123011	15.35964	6.688161	2.308107
	d(lnIV)=d(Volume)	5.774031	3.914039	1.879207	6.213721	4.121013	2.137205	10.42038	5.959202	1.566566
	d(RV)=d(Volume)		0.110748	0.657799		0.9667	0.309087		3.202657	0.497217
	d($\sqrt{RV}$)=d(Volume)		1.962994	1.531289		2.38494	1.356465		3.944742	1.463592
	d(lnRV)=d(Volume)		2.734093	1.838977		2.857436	2.068847		3.443963	1.496439
	d(Range)=d(Volume)	5.299475	0.36109	0.764269	6.036139	0.223411	0.149077	7.902687	1.925738	0.120379
	d($\sqrt{Range}$)=d(Volume)	5.807202	1.177683	1.232513	6.624897	1.909022	0.747182	9.070929	3.371342	0.867566
$V \sim lnVolume$	d(IV)=d(lnVolume)	6.245326	2.534415	0.877921	7.853425	2.470487	0.901216	10.82368	5.121861	1.148475
	d(VIX)=d(lnVolume)	6.33374	3.682063	1.400111	7.273738	3.489304	1.513495	11.43786	5.844427	1.401392
	d(lnIV)=d(lnVolume)	5.774031	3.914039	1.879207	6.213721	4.121013	2.137205	10.42038	5.959202	1.566566
	d(RV)=d(lnVolume)		0.496544	0.074889		0.075335	0.300428		2.358923	0.409498
	d($\sqrt{RV}$)=d(lnVolume)		1.355702	0.94838		1.493575	0.74695		3.101008	0.556877
	d(lnRV)=d(lnVolume)		2.126801	1.256067		1.966071	1.459331		2.600229	0.589724
	d(Range)=d(lnVolume)	3.493921	0.968382	0.18136	2.420516	0.667954	0.758593	3.980906	1.082004	1.027094
	d($\sqrt{Range}$)=d(lnVolume)	4.001648	0.570391	0.649604	3.009274	1.017656	0.137666	5.149148	2.527608	0.039149

Table 57: T statistics for integration in Russell2000-RVX data

		$m = n^{0.5}$			$m = n^{0.55}$			$m = n^{0.6}$		
		Daily	Weekly	Monthly	Daily	Weekly	Monthly	Daily	Weekly	Monthly
$IV \sim RV$	d(IV)=d(RV)		2.147437	0.336172		1.925549	0.313802		2.352512	0.769552
	d(lnIV)=d(lnRV)		0.977473	0.48295		0.728632	0.43716		3.514425	0.970269
	d(VIX)=d($\sqrt{RV}$)		1.653661	0.325308		1.426768	0.313373		2.781779	0.704736
	d(IV)=d(Range)	2.781702	3.219054	0.85018	7.037124	3.50992	1.128194	8.819279	4.092757	1.45227
	d(VIX)=d($\sqrt{Range}$)	3.074452	2.333678	0.988164	5.53005	2.618502	0.996643	7.800261	3.334497	0.566774
$V \sim Volume$	d(IV)=d(Volume)	5.544999	0.464789	1.469223	6.223783	0.163007	1.992784	7.5087	2.592758	0.521996
	d(VIX)=d(Volume)	5.14944	1.225839	0.869223	5.895169	1.026535	1.448798	8.522366	2.959914	0.203603
	d(lnIV)=d(Volume)	2.971005	0.193777	1.685939	3.388532	0.276596	2.276676	7.017002	1.373729	1.488967
	d(RV)=d(Volume)		1.682649	1.805395		1.762542	2.306586		0.240246	1.291548
	d($\sqrt{RV}$)=d(Volume)		0.427822	1.19453		0.400233	1.762172		0.178135	0.908339
	d(lnRV)=d(Volume)		0.491301	0.846352		0.745977	1.40696		0.892725	0.862489
	d(Range)=d(Volume)	2.763297	2.754265	2.319403	0.813342	3.346913	3.120978	1.310579	1.499999	1.974266
	d($\sqrt{Range}$)=d(Volume)	2.074988	1.107839	1.857386	0.365119	1.591967	2.445442	0.722105	0.374583	0.770377
$V \sim lnVolume$	d(IV)=d(lnVolume)	4.479649	1.197763	2.79176	4.866194	1.035006	3.29966	6.978086	1.344787	2.118743
	d(VIX)=d(lnVolume)	4.08409	0.436713	2.191759	4.53758	0.171478	2.755675	7.991751	1.711943	1.80035
	d(lnIV)=d(lnVolume)	2.971005	0.193777	1.685939	3.388532	0.276596	2.276676	7.017002	1.373729	1.488967
	d(RV)=d(lnVolume)		3.345201	3.127932		2.960555	3.613462		1.007726	2.888295
	d($\sqrt{RV}$)=d(lnVolume)		2.090374	2.517067		1.598246	3.069048		1.069836	2.505086
	d(lnRV)=d(lnVolume)		1.171251	2.168889		0.452036	2.713837		2.140697	2.459236
	d(Range)=d(lnVolume)	1.697947	4.416817	3.64194	2.17093	4.544926	4.427854	1.841193	2.74797	3.571013
	d($\sqrt{Range}$)=d(lnVolume)	1.009638	2.770391	3.179923	0.99247	2.78998	3.752318	0.191491	1.622555	2.367124

Table 58: Estimates of d: monthly data for all 5 indexes

		VIX		VXO		VXD		VXN		RVX	
	std.dev	IV	RV	IV	RV	IV	RV	IV	RV	IV	RV
$m = n^{0.5}$	0.158114	0.490848	0.33981304	0.477861	0.3325215	0.493587	0.34071541	0.486015	0.359044354	0.406843	0.35369
$m = n^{0.55}$	0.138675	0.548288	0.3858397	0.555983	0.38434506	0.56169	0.38492982	0.550537	0.383899385	0.466216	0.4227
$m = n^{0.6}$	0.121265	0.619079	0.42075279	0.632503	0.42277905	0.689927	0.41071815	0.616184	0.427251986	0.56989	0.47657
		lnIV	lnRV	lnIV	lnRV	lnIV	lnRV	lnIV	lnRV	lnIV	lnRV
$m = n^{0.5}$	0.158114	0.697789	0.57711437	0.66774	0.59246904	0.712898	0.55592748	0.644332	0.545804942	0.581689	0.505328
$m = n^{0.55}$	0.138675	0.7176	0.62222158	0.735054	0.65595759	0.763808	0.6359512	0.721938	0.627934101	0.608079	0.547455
$m = n^{0.6}$	0.121265	0.686525	0.55647219	0.696763	0.56691379	0.793778	0.56254801	0.666885	0.54842547	0.64626	0.5286
		VIX	$\sqrt{RV}$	VIX	$\sqrt{RV}$	VIX	$\sqrt{RV}$	VIX	$\sqrt{RV}$	VIX	$\sqrt{RV}$
$m = n^{0.5}$	0.158114	0.594331	0.49419207	0.577037	0.49397467	0.608142	0.49848805	0.56858	0.497155317	0.501712	0.450276
$m = n^{0.55}$	0.138675	0.632383	0.53422153	0.647508	0.54902686	0.664961	0.55411126	0.635445	0.52914456	0.541654	0.498197
$m = n^{0.6}$	0.121265	0.65162	0.54049661	0.667407	0.55261961	0.745303	0.54771078	0.646855	0.544442261	0.6085	0.52304
		IV	Range	IV	Range	IV	Range	IV	Range	IV	Range
$m = n^{0.5}$	0.158114	0.490848	0.33857409	0.477861	0.34363487	0.493587	0.3727253	0.486015	0.375878822	0.406843	0.272418
$m = n^{0.55}$	0.138675	0.548288	0.37453725	0.555983	0.37852435	0.56169	0.40426557	0.550537	0.320363363	0.466216	0.309764
$m = n^{0.6}$	0.121265	0.619079	0.45702209	0.632503	0.47262514	0.689927	0.47565299	0.616184	0.352357525	0.56989	0.39378
		VIX	$\sqrt{Range}$	VIX	$\sqrt{Range}$	VIX	$\sqrt{Range}$	VIX	$\sqrt{Range}$	VIX	$\sqrt{Range}$
$m = n^{0.5}$	0.158114	0.594331	0.45028656	0.577037	0.46648296	0.608142	0.49688164	0.56858	0.449914707	0.501712	0.345469
$m = n^{0.55}$	0.138675	0.632383	0.49516533	0.647508	0.50768542	0.664961	0.53726469	0.635445	0.44465212	0.541654	0.403444
$m = n^{0.6}$	0.121265	0.65162	0.58368625	0.667407	0.60478641	0.745303	0.59894744	0.646855	0.472163485	0.6085	0.53977

Table 59: Fractionally cointegration coefficient when $m = n^{0.5}$

	VIX			VXO			VXD			VXN			RVX		
	ξ_1	ξ_2	$d(\xi'X)$	ξ_1	ξ_2	$d(\xi'X)$	ξ_1	ξ_2	$d(\xi'X)$	ξ_1	ξ_2	$d(\xi'X)$	ξ_1	ξ_2	$d(\xi'X)$
$X_1 = IV$	1	0	0.4908	1	0	0.4779	1	0	0.4936	1	0	0.4860	1	0	0.4068
$X_2 = RV$	0	1	0.3398	0	1	0.3325	0	1	0.3407	0	1	0.3590	0	1	0.3537
$\min r^2$	1	-0.7879	0.0011	1	-0.9091	0.1448	1	-0.7879	0.1554	1	-0.7879	0.0215	1	-0.7475	0.0708
Unity	1	-1	0.0087	1	-1	0.0290	1	-1	0.0426	1	-1	0.0015	1	-1	0.0889
OLS	1	-0.6833	0.4231	1	-0.7783	0.4164	1	-0.5761	0.4664	1	-0.7392	0.4313	1	-0.7036	0.4127
$X_1 = IV$	1	0	0.4908	1	0	0.4779	1	0	0.4936	1	0	0.4860	1	0	0.4068
$X_2 = Range$	0	1	0.3386	0	1	0.3436	0	1	0.3727	0	1	0.3759	0	1	0.2724
$\min r^2$	1	-0.3838	0.0755	1	-0.4646	0.0001	1	-0.4242	0.0961	1	-0.3838	0.0001	1	-0.3434	0.0051
Unity	1	-1	0.2454	1	-1	0.2273	1	-1	0.2299	1	-1	0.3002	1	-1	0.1280
OLS	1	-0.3329	0.4212	1	-0.3840	0.4183	1	-0.2728	0.4706	1	-0.3343	0.4392	1	-0.2461	0.3925
$X_1 = VIX$	1	0	0.5943	1	0	0.5770	1	0	0.6081	1	0	0.5686	1	0	0.5017
$X_2 = \sqrt{RV}$	0	1	0.4942	0	1	0.4940	0	1	0.4985	0	1	0.4972	0	1	0.4503
$\min r^2$	1	-0.8687	0.2896	1	-0.9495	0.1798	1	-0.8283	0.0490	1	-0.9091	0.2740	1	-0.7879	0.0740
Unity	1	-1	0.2572	1	-1	0.1450	1	-1	0.0134	1	-1	0.2619	1	-1	0.0371
OLS	1	-1.0285	0.5392	1	-1.0750	0.5354	1	-0.9989	0.5723	1	-1.0422	0.5318	1	-1.0111	0.5008
$X_1 = VIX$	1	0	0.5943	1	0	0.5770	1	0	0.6081	1	0	0.5686	1	0	0.5017
$X_2 = \sqrt{Range}$	0	1	0.4503	0	1	0.4665	0	1	0.4969	0	1	0.4499	0	1	0.3455
$\min r^2$	1	-0.6667	0.0157	1	-0.7475	0.0029	1	-0.6667	0.1673	1	-0.6667	0.0063	1	-0.5859	0.3402
Unity	1	-1	0.2321	1	-1	0.2043	1	-1	0.1724	1	-1	0.2544	1	-1	0.0236
OLS	1	-0.7207	0.5177	1	-0.7507	0.5214	1	-0.6884	0.5680	1	-0.6938	0.5088	1	-0.6833	0.4532
$X_1 = \ln IV$	1	0	0.6978	1	0	0.6677	1	0	0.7129	1	0	0.6443	1	0	0.5817
$X_2 = \ln RV$	0	1	0.5771	0	1	0.5925	0	1	0.5559	0	1	0.5458	0	1	0.5053
$\min r^2$	1	-0.9495	0.2903	1	-0.9899	0.2918	1	-0.8687	0.1214	1	-0.9899	0.2809	1	-0.9495	0.1489
Unity	1	-1	0.3180	1	-1	0.2987	1	-1	0.0804	1	-1	0.2837	1	-1	0.0160
OLS	1	-1.1222	0.6243	1	-1.1383	0.6267	1	-1.1074	0.6399	1	-1.0894	0.5876	1	-1.0798	0.5640

Table 60: Fractionally cointegration coefficient when $m = n^{0.55}$

	VIX			VXO			VXD			VXN			RVX		
	ξ_1	ξ_2	$d(\xi'X)$	ξ_1	ξ_2	$d(\xi'X)$	ξ_1	ξ_2	$d(\xi'X)$	ξ_1	ξ_2	$d(\xi'X)$	ξ_1	ξ_2	$d(\xi'X)$
$X_1 = IV$	1	0	0.5483	1	0	0.5560	1	0	0.5617	1	0	0.5505	1	0	0.4662
$X_2 = RV$	0	1	0.3858	0	1	0.3843	0	1	0.3849	0	1	0.3839	0	1	0.4227
$\min r^2$	1	-0.7475	0.0520	1	-0.9091	0.0001	1	-0.7879	0.0001	1	-0.7879	0.0198	1	-0.7475	0.0758
Unity	1	-1	0.0504	1	-1	0.1205	1	-1	0.1482	1	-1	0.0349	1	-1	0.0513
OLS	1	-0.6833	0.4776	1	-0.7783	0.4827	1	-0.5761	0.5437	1	-0.7392	0.4736	1	-0.7036	0.4992
$X_1 = IV$	1	0	0.5483	1	0	0.5560	1	0	0.5617	1	0	0.5505	1	0	0.4662
$X_2 = Range$	0	1	0.3745	0	1	0.3785	0	1	0.4043	0	1	0.3204	0	1	0.3098
$\min r^2$	1	-0.3838	0.0001	1	-0.4646	0.0001	1	-0.4242	0.0001	1	-0.3838	0.0001	1	-0.3434	0.0001
Unity	1	-1	0.2682	1	-1	0.2371	1	-1	0.2292	1	-1	0.1970	1	-1	0.1434
OLS	1	-0.3329	0.4706	1	-0.3840	0.4763	1	-0.2728	0.5418	1	-0.3343	0.4444	1	-0.2461	0.4666
$X_1 = VIX$	1	0	0.6324	1	0	0.6475	1	0	0.6650	1	0	0.6354	1	0	0.5417
$X_2 = \sqrt{RV}$	0	1	0.5342	0	1	0.5490	0	1	0.5541	0	1	0.5291	0	1	0.4982
$\min r^2$	1	-0.8687	0.2493	1	-0.9495	0.1635	1	-0.8283	0.0459	1	-0.9091	0.2852	1	-0.7879	0.2157
Unity	1	-1	0.2627	1	-1	0.1520	1	-1	0.0046	1	-1	0.2803	1	-1	0.1139
OLS	1	-1.0285	0.5824	1	-1.0750	0.6004	1	-0.9989	0.6452	1	-1.0422	0.5791	1	-1.0111	0.5651
$X_1 = VIX$	1	0	0.6324	1	0	0.6475	1	0	0.6650	1	0	0.6354	1	0	0.5417
$X_2 = \sqrt{Range}$	0	1	0.4952	0	1	0.5077	0	1	0.5373	0	1	0.4447	0	1	0.4034
$\min r^2$	1	-0.6667	0.0113	1	-0.7475	0.0001	1	-0.6667	0.0981	1	-0.6667	0.0001	1	-0.5859	0.0001
Unity	1	-1	0.2798	1	-1	0.2168	1	-1	0.1837	1	-1	0.1968	1	-1	0.0341
OLS	1	-0.7207	0.5614	1	-0.7507	0.5765	1	-0.6884	0.6294	1	-0.6938	0.5357	1	-0.6833	0.5225
$X_1 = \ln IV$	1	0	0.7176	1	0	0.7351	1	0	0.7638	1	0	0.7219	1	0	0.6081
$X_2 = \ln RV$	0	1	0.6222	0	1	0.6560	0	1	0.6360	0	1	0.6279	0	1	0.5475
$\min r^2$	1	-0.9495	0.3635	1	-0.9899	0.3435	1	-0.8687	0.0626	1	-0.9899	0.4238	1	-0.9495	0.1167
Unity	1	-1	0.3875	1	-1	0.3499	1	-1	0.1631	1	-1	0.4254	1	-1	0.0110
OLS	1	-1.1222	0.6647	1	-1.1383	0.6950	1	-1.1074	0.7226	1	-1.0894	0.6699	1	-1.0798	0.6191

Table 61: Fractionally cointegration coefficient when $m = n^{0.6}$

	VIX			VXO			VXD			VXN			RVX		
	ξ_1	ξ_2	$d(\xi'X)$	ξ_1	ξ_2	$d(\xi'X)$	ξ_1	ξ_2	$d(\xi'X)$	ξ_1	ξ_2	$d(\xi'X)$	ξ_1	ξ_2	$d(\xi'X)$
$X_1 = IV$	1	0	0.6191	1	0	0.6325	1	0	0.6899	1	0	0.6162	1	0	0.5699
$X_2 = RV$	0	1	0.4208	0	1	0.4228	0	1	0.4107	0	1	0.4273	0	1	0.4766
$\min r^2$	1	-0.7475	0.0048	1	-0.8687	0.0001	1	-0.7879	0.0736	1	-0.7879	0.0194	1	-0.7475	0.2488
Unity	1	-1	0.0872	1	-1	0.1887	1	-1	0.2155	1	-1	0.0829	1	-1	0.1191
OLS	1	-0.6833	0.5278	1	-0.7783	0.5345	1	-0.5761	0.6298	1	-0.7392	0.5269	1	-0.7036	0.5951
$X_1 = IV$	1	0	0.6191	1	0	0.6325	1	0	0.6899	1	0	0.6162	1	0	0.5699
$X_2 = Range$	0	1	0.4570	0	1	0.4726	0	1	0.4757	0	1	0.3524	0	1	0.3938
$\min r^2$	1	-0.3838	0.0001	1	-0.4646	0.0001	1	-0.4242	0.0020	1	-0.3838	0.0001	1	-0.3434	0.0001
Unity	1	-1	0.3511	1	-1	0.3385	1	-1	0.2886	1	-1	0.2124	1	-1	0.2237
OLS	1	-0.3329	0.5505	1	-0.3840	0.5634	1	-0.2728	0.6519	1	-0.3343	0.4982	1	-0.2461	0.5739
$X_1 = VIX$	1	0	0.6516	1	0	0.6674	1	0	0.7453	1	0	0.6469	1	0	0.6085
$X_2 = \sqrt{RV}$	0	1	0.5405	0	1	0.5526	0	1	0.5477	0	1	0.5444	0	1	0.5230
$\min r^2$	1	-0.8687	0.2727	1	-0.9495	0.2202	1	-0.8283	0.0122	1	-0.9091	0.3032	1	-0.7879	0.1371
Unity	1	-1	0.2903	1	-1	0.2136	1	-1	0.0524	1	-1	0.3099	1	-1	0.0303
OLS	1	-1.0285	0.5917	1	-1.0750	0.6074	1	-0.9989	0.6695	1	-1.0422	0.5922	1	-1.0111	0.6045
$X_1 = VIX$	1	0	0.6516	1	0	0.6674	1	0	0.7453	1	0	0.6469	1	0	0.6085
$X_2 = \sqrt{Range}$	0	1	0.5837	0	1	0.6048	0	1	0.5989	0	1	0.4722	0	1	0.5398
$\min r^2$	1	-0.6667	0.1347	1	-0.7475	0.0293	1	-0.6667	0.0394	1	-0.6667	0.0001	1	-0.5859	0.0323
Unity	1	-1	0.4183	1	-1	0.3699	1	-1	0.2977	1	-1	0.1937	1	-1	0.2105
OLS	1	-0.7207	0.6222	1	-0.7507	0.6428	1	-0.6884	0.6874	1	-0.6938	0.5688	1	-0.6833	0.6285
$X_1 = \ln IV$	1	0	0.6865	1	0	0.6968	1	0	0.7938	1	0	0.6669	1	0	0.6463
$X_2 = \ln RV$	0	1	0.5565	0	1	0.5669	0	1	0.5625	0	1	0.5484	0	1	0.5286
$\min r^2$	1	-0.9495	0.3115	1	-0.9899	0.2665	1	-0.8687	0.1151	1	-0.9899	0.3447	1	-0.9495	0.0494
Unity	1	-1	0.3305	1	-1	0.2715	1	-1	0.1814	1	-1	0.3461	1	-1	0.0963
OLS	1	-1.1222	0.6074	1	-1.1383	0.6210	1	-1.1074	0.6644	1	-1.0894	0.5976	1	-1.0798	0.6052

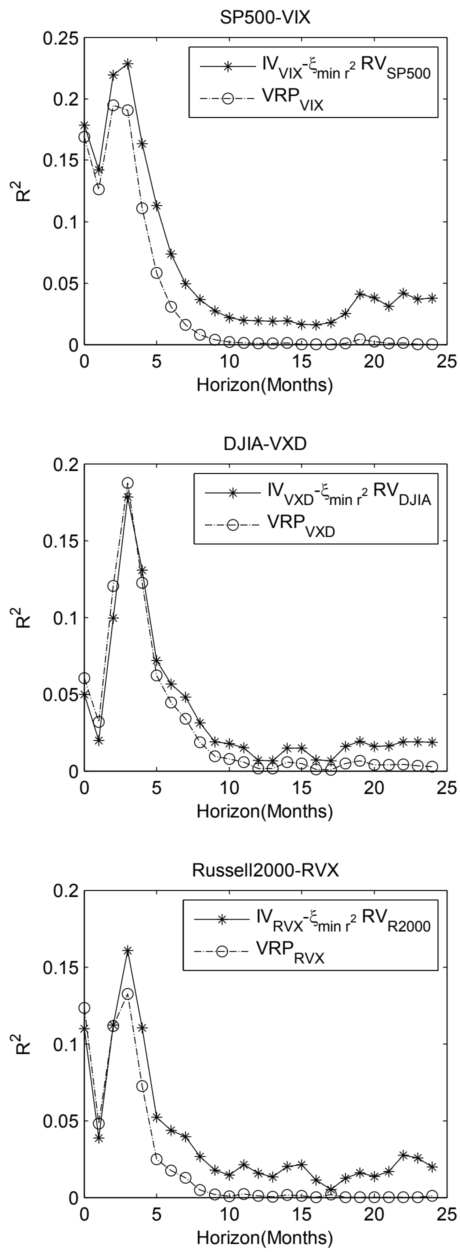

Figure 7: 24 months static in sample forecast by different dataset

CHAPTER 3: DOES OCCUPANCY STATUS MATTER IN SUBPRIME MORTGAGE?

By using submortgage data, we found that investors are being charged with a significant risk premium over owner occupants; besides that, they are also facing a more restricted loan; with the market getting hotter, this risk premium and restrictions are getting even worse. Being treated like that, our findings show that investors were actually not more risky than owner occupants in terms of both prepayment and default.

We suspect the reason for this puzzle is that when the market getting hotter, there are more speculative investors who commit occupancy fraud to get a more favorable loan. And these speculative investors were actually recorded as owner occupants on loan documents, which increased our estimation of the hazard of owner occupants group. And our information asymmetry test actually reaffirmed our suspect. Therefore, this paper, for the first time, give statistical evidence on occupancy fraud, and we also proposed a statistical scanning way to reduce to potential occupancy fraud.

3.1 Introduction

It is common knowledge among real estate investors that when they fill a pre-assessment table as in table 62; for the item “Home will be”, once “Investment” is checked, they have to pay more interest rate and meet higher requirement(e.g. more down payment) to get the loan.

For a rational investor who is trying to minimize his cost; apparently, to avoid a check on the “Investment” item is a good choice, and it’s also an easy choice for them since in some cases, they only need to claim that they “intend” to live in that house (though they know they won’t).

But investors know that “the easy choice” will be occupancy fraud, which is violating the federal law. Investors need to balance the cost of going to jail and the benefit of saving some interest. If banks only focus on the volume and put less effort to detect the fraud while mortgage brokers and real-estate agents eager to close a home sale so that they can tolerate that fraud; the temptation to commit fraud can be substantial and actually “encouraged” to some extent by both banks and agents.

And also, when home prices fall, non-occupancy investors tend to be more likely to walk away from their purchases than ordinary homebuyers.If occupancy fraud happens a lot, there will be much risk hidden in the loans which are supposed to be backed by those homebuyers who live in their house but turn out to be non-occupancy investors. The loans backing these non-occupancy purchases turn out to be riskier than the rating agencies and those investors who bought mortgage-backed securities once thought.

So when the home prices really fall, there will be lots of “unexpected” defaults whose risk is not priced into the loans or the pools, so it will cause problems for the whole housing and mortgage market, and then the whole financial market, so the financial crisis comes.

Academically, there are lots of papers from law school’s that discuss this occupancy issue from “a lawyer’s” point of view. Simon and Corkery (2008) analyzed the whole process including the incentive of non-occupancy investors to defraud on occupancy and how banks and real estate agencies encourage or at least tolerate them to do so, and how these occupancy frauds can be part of the cause of problems on the housing and mortgage market today, and at last they conclude that speculators (especially those who did occupancy fraud) may have accelerated housing downturn. This small article in the Wall Street Journal got lots of attention both from the industry and academia, and especially in law school.

The most recent one, Lefcoe (2009) recounts the extent to which speculating buyers contributed more than proportionately to housing price volatility and the rate of mortgage foreclosure; then the author disclosed the way that spec buyers deceived mortgage lenders by committing occupancy fraud, claiming falsely that they were buying as owner occupants so they could benefit from more favorable mortgage rates and terms; at last, the author explored the rational for a government imposed ban on home flipping.

Also, by examining the flipping activity in Las Vegas from 1994 through mid-2007, Depken et al (2009) found that flip homes tend to be older and smaller than non-flip homes; and flippers appear to purchase the flip home at a discount and they sell the flip home at a premium to otherwise similar properties, they also found that as the residential property market in Las Vegas begins to take off, flip homes become a more significant percentage of total sales.

This paper is structured as following: in section 2, we introduce the data we use for the empirical tests, including both the pool level data and loan level data; insection 3, we do empirical tests to see how much higher interest the investors are charged and how much more restrictions (in terms of LTV ratio) are put on them by banks; in section 4, we do empirical tests to see how dangerous these non-occupancy investors are and get an idea how dangerous those occupancy fraud performers are; in section 5, we test the information asymmetry and introduce some ways to prevent potential occupancy fraud, then we conclude in section 6.

3.2 Data

Here we used two sets of data, loan level and pool level.

3.2.1 Loan Level Data

For the loan level data, the author got it from FHFA (Federal Housing Finance Agency[12]). The data include static section and dynamic section. For the static section, it includes 20,000 individual subprime mortgage loans initiated from 10/9/2003 to 10/23/2007, all the loans are from two states:Arizona and Florida; these two states, according to Simon and Corkery (2008) are where “much of the occupancy fraud was concentrated”. Among our data, there are 15968 owner occupants, 1388 second home owner and 2639 investors, see figure 8.

[12]We appreciate Dr. Robert Dunsky’s help on getting the data.

All of these loans are adjusted rate mortgages (ARM). The number of loans issued each month in our data is 200 at the beginning of 2004, then it increased with the housing market getting hotter, from the second quarter of 2005 to the end of 2006, the number of loans issued each month trippled; however, after 2006, with this subprime mortgage crisis, the number of loans issued each month dropped dramatically to less than 100 at the end of 2007. We summarized origination date of our data in figure 9.

For all these loans, we have borrowers information at the loan origination, including borrower's FICO score at initiation, loan type(whether there is a balloon payment or not), loan purpose (purchase or refinance), document status (whether it's a "low doc" loan or not), which state it belongs to (Florida or Arizona), in which year this loan was originated (04, 05, 06 or 07), loan maturity (15, 20, 30 or 40 years) and the original amount of the loan. The description of all the variables we are using is summarized at table 63.

For the dynamic section, there are 555,512 monthly observations after the loan get initiated. It includes the monthly payment, current balance and the status of each loan (whether it's prepaid, delinquent or default). The data is right censored on 10/1/2007. By the right censored date, there are 80 percent of the loan still current (neither prepaid nor default yet), 3 percent prepaid and 17 percent default, see figure 10.

3.2.2 Pool Level Data

The pool level data [13] the author uses are from Freddie Mac PC(Participation Certificate). To be in this dataset, they must be 30 year fixed rate mortgages and have at least 150 mortgages included within the pool, and they are issued as Mortgage Backed Securities between Jan 2006 and Mar 2006. There are 434 PCs in total, including 142815 loans in them. The description of all the variables we are using is summarized at table 64.

3.3 Risk Premium

In this section, we are trying to quantify what is the “risk premium” being charged on investors over owner occupants. Here the “risk premium” contain two folds: one is the real risk premium in term of interest rate, which means how much higher the interest rate investors were being charged by subprime mortgage lender over owner occupants; the other is not the real risk premium, but the restrictions investors have to face. Since owner occupants, not only being charged with lower interest rate, but also qualify for much smaller down payment. So, as an investor, she/he also has to face a lower Loan To Value(LTV) ratio. So basically in this section, we quantified how much restrictions investors have to face, both in terms of interest rate and LTV ratio.

[13] I appreciate Jason Berkowitz who provided me with the pool level data which is hand collected by himself.

3.3.1 Risk Based Mortgage Pricing Model

Here, following Hendershott and Shilling (1989) , Ambrose, LaCour-little and Sanders (2004) and An, Do, Rosenblette and Yao (2012) , we use the following linear regression model to test the investor's risk premium while controling other risk factors:

$$r_i = X_i\beta + \epsilon_i, i = 1, \dots n$$

where here r_i is the subprime mortgage spread, X_i is a vector of pricing factors recorded on the loan documents, such as borrower's FICO score at initiation, loan type(whether there is a balloon payment or not), loan purpose (purchase or refinance), document status (whether it's a "low doc" loan or not), which state it belongs to (Florida or Arizona), in which year this loan was originated (04,05,06 or 07), loan maturity (15, 20, 30 or 40 years) and the original amount of the loan. ϵ_i here is the disturbance.

Here, we put our result in table 65, the basic model we just stated is model 1 in table 2; to test the dynamic trend of the risk premium investors being charged over owner occupants from 2004 to 2007, we added the cross product of the time dummy and the investor dummy, which is model 2 in table 65. From both of these model, we can see that investors were being charged by a statistically significant risk premium, which is around 0.5 percent per year; according to the result of model 2, we can see that this premium is even increasing over the years, from 0.5 percent in 2004 to almost 1 percent in 2007 (see figure 1 on this trend).The reason why this investor risk premium was increasing every year, we think it might because the housing market is getting hot from 2004 to 2007, so banks charge investors a higher premium every year to compensate their risk. And other parameters also get the predicted sign: for example, after controlling other variables, the higher the FICO score, the lower the interest rate being charged; and "low-doc" applicants in average got a higher interest rate than non "low-doc" applicants. But no matter which model we use, our pricing factors recorded on the loan documents could only explain less than 50 percent of the total variation in interest rate. We plot the risk premium in figure 11.

So, by running this risk based mortgage pricing model, we found that investors did get a statistically significant risk premium around 0.5 percent to 1 percent, and this risk premium is increasing from 2004 to 2007. After finding this statistically significant risk premium, the next question is how restricted in terms of LTV ratio investors have to face.

3.3.2 LTV Ratio

The LTV ratio is a very important indicator to see the applicants leverage. Usually, since favorable home mortgage terms are reserved for owner occupants, as typical first time home buyers in US, people only are required to pay a very low down payment, which means the LTV ratio of owner occupants could be very high, close to 100 percent. However, as investors they do not only have to make a lot of effort to demonstrate their abilities in managing properties for some history (at least two years), and more reserve of cash to pay property taxes and insurance, but they also have to pay a higher down payment, which will reduce their ability to leverage, so investors will have a lower LTV ratio compared to owner occupants.

Here, we want to quantify, despite the higher risk premium, what other restrictions investors have to face, so we use LTV ratio as an indicator of the restrictions investors have to face. We are running a similar model as the risk based mortgage pricing model:

$$LTV_i = X_i\beta + \epsilon_i, i = 1, \ldots n$$

where this LTV_i is the loan to value ratio, X_i is a vector of pricing factors stated in the previous model, and ϵ_i is the disturbance.

The result is in table 66, in which model 3 is just as stated above, while model 4 is trying to test the dynamic trend of this LTV ratio over the years. From the result in table 5, we can see that investors have to face a more restricted loan: with other factors beging controlled, investors in average are getting a loan with a LTV ratio 3.2 percent lower than the owner occupants in absolute value. Even worse, investors are getting loans with a LTV ratio from around 3 percent lower in 2004 to almost 6 percent lower in 2007 than owner occupants. (see figure 2 for this trend). Still, in these two models, the coeffcients for other control variables also make sense, such as: the bigger the original loan amount, the lower the LTV ratio (due to some restrictions on Jumbo mortgage); the higher the FICO score, the higher the LTV ratio the applicants could get; and "low-doc" applicants are in average getting a lower LTV ratio than non "low-doc" applicants.

Therefore, from the analysis of interest rate and LTV ratio investors have to face, we can get the result that as an investor, she/he has to be charged by a higher interest rate than occupant owner and has to face a more restricted loan (less leverages, lower LTV ratio than occupant owner. What makes things even worse is, with the housing market getting hotter from 2004 to 2007, the risk premium also increased and the restrictions tightened. To test the robustness of our result, we use the pool level data to see whether the result is similar to loan level data.

3.3.3 Pool level data

Here, by using the pool level data, we are still running the previous two models as

$$\begin{cases} r_i = X_i\beta + \epsilon_i, i = 1, \ldots n \\ LTV_i = X_i\beta + \epsilon_i, i = 1, \ldots n \end{cases}$$

the only difference here is that for pool level data, we don't have as much detail on applicant's information as in loan level data: for the prcing factors, we only have the average FICO score, average maturity, LTV ratio, percentage of purchasing (instead of refinancing), percentage of investors (instead of owner occupants), average loan amount at origination for each pool.

So we put all the information we get from the pool level data into the above model and our result is consistent with what we get from the loan level data (see table 67): apprently, even in pool level, we could still see evidence that investors were being charged with a statistically significant risk premium, and have to face a more restricted (much lower LTV ratio) loan. So this pool level data analysis could be a robustness check of our loan level data.

3.4 Investors' Hazards

From the previous section we found that investors have to face a statistically significant risk premium around 0.5 to 1 percent, and a statistically significant lower LTV ratio around 3 to 6 percent in absolute value. the next question we want to ask is "do investors deserve this risk premium?" or "are they really that risky?" so we used the competing risk model to test the hypothesis of whether investors are more risky (in terms of both default and prepayment) than owner occupants.

From previous literature on analyzing mortgage risk, Campbell and Dietrich (1983) , Cunningham and Capone (1990) , Archer, Ling and McGill (1996) and Calhoun and Deng (2002) used the multinomial logit model, but the problem with multinomial logit model is the assumption that alternative termination risks are independent, but apparently, they are not. Lots of people also used single risk cox partial likelihood model, such as Green and Shoven (1986) , Clapp, Goldberg, Harding and Lacour-Little (2001) , and Pavlov (2001) . But this single risk model usually only consider default, without recognizing the prepayment risk, which will also affect the default risk: for example, sometimes when people are close to default, they will try to refinance to reduce the payment first, if they were able to refinance, this loan itself become prepaid instead of default. So actually, prepayment risk and default risk are intercorrelated with each other, to recognize this property of mortgage, we are using the competing risk model.

3.4.1 Competing Risk Model

Competing risk model is a well developed model using in biological literature for many years, it's first used in analyzing mortgage data by Deng, Quigley and Order (2000) a decade ago. Following the classical textbook Lancaster (1990) , we define the prepayment hazard rate as

$$\lambda_P(t,x) = \lim_{dt\to 0} \frac{Pr_P(t \le T < t+dt|T \ge t, x)}{dt}$$

which is the conditional probability that an individual with covariates x prepays in the interval $[t, t+dt]$, given the individual was still "current" (not default not prepaid yet) just before time t. The default hazard rate as

$$\lambda_D(t,x) = \lim_{dt \to 0} \frac{Pr_D(t \le T < t + dt | T \ge t, x)}{dt}$$

which has the similar explanation as $\lambda_P(t,x)$ except it's default instead of prepayment this time. We assume the only risk for a mortgage is prepayment and default, so,

$$\lambda(t,x) = \lambda_D(t,x) + \lambda_P(t,x)$$

There for the survival function can be defined as

$$S(t,x) = exp\left\{-\int_0^t \lambda(u,x)du\right\} = exp\left\{-\int_0^t [\lambda_D(u,x) + \lambda_P(u,x)]du\right\}$$

which is the probability an individual with covariates x being "current" at time t. so the unconditional density that an individual default at time t would be

$$f_D(t,x) = \lim_{dt \to 0} \frac{Pr_D(t \le T < t + dt | x)}{dt} = \lambda_D(t,x)S(t,x)$$

the same thing for an individual prepay at time t would be

$$f_P(t,x) = \lim_{dt \to 0} \frac{Pr_P(t \le T < t + dt | x)}{dt} = \lambda_P(t,x)S(t,x)$$

Therefore the likelihood function for estimation is

$$L = \prod_{i=1}^{n} f(t_i, x_i) = \prod_{i=1}^{n} \prod_{j \in \{P,D\}} \lambda_j(t_i, x_i)^{d_{ij}} exp\left\{-\int_0^{t_i} \lambda_j(u, x_i)du\right\}$$

where d_{iD} is 1 if individual i default, 0 if individual i get censored, and also d_{iP} is 1 if individual i prepaid, 0 if individual i get censored.

Fine and Gray (1999) proposed a proportional hazard model for competing risk, they seperated the baseline hazard from the regression coefficient by assuming the hazard rate

$$\lambda_j(t_i, x_i) = \lambda_{j0}(t_i)exp\{x_i\beta\}$$

so the likelihood function become

$$L = \prod_{i=1}^{n} \prod_{j \in \{P,D\}} [\lambda_{j0}(t_i) exp\{x_i\beta\}]^{d_{ij}} exp\left\{- \int_0^{t_i} \lambda_{j0}(t_i) exp\{x_i\beta\}\, du\right\}$$

so we don't have to estimate the baseline hazard function anymore, so the whole estimation becomes easier, and we used this estimation method in R by following Fine and Gray (1999) .

3.4.2 Estimation Result

Here, we first estimated the predicted cumulative probability of default and prepayment and the result is in figure 13. From this figure we can see that after 5 years, the probability of prepayment is around 3-5 percent and the probability of default is almost 50 percent.

The next issue we want to address is whether there is any difference between investors and owner occupants in terms of their prepayment and default behavior. The result is in figure 14, from which we can not tell the difference between owner occupants and the investors; by using the competing risk model to test the null hypothesis that there is no significant difference between investors and owner occupants behavior on prepayment and default, we get the t statistic as following:

	t-statistic	Probability
Default	0.1643705	0.6851641
Prepayment	0.6007734	0.4382831

from which we can see that we can not actually reject the null hypothesis that there is no significant difference between investors and owner occupants behavior on prepayment and default.

Controlling other variables, we run the competing risk model again and the result is in table 68, from which we still could not reject the null hypothesis that there is no significant difference between investors and owner occupants in terms of prepayment and default. But the other variables still make sense, for example, the higher the FICO score, the lower the default hazard; low doc borrowers tend to be more likely to default than non low doc borrowers. So from this competing risk model, in terms of default and prepayment, owner occupants are at least as risky as investors.

so why do investors is being charged with a much higher risk premium and have to face a more restricted loan and then is not more risky than owner occupants? Is the market not so efficient to correct this error?

We think there are majorly two reasons. First, owner occupants get a lower interest rate and a more favorable loan because the government want to promote the idea that everyone has his own house, the government want to let people have their own house. so the most favorable loans are always reserved for the owner occupants instead of investors, even in terms of default and prepayment, owner occupants are as risky as investors.

The second reason is that in fact, investors are more risky than owner occupants, which is documented in some of previous literature, however, with the strong incentive to commit occupancy fraud, lots of investors were actually recorded as owner occupants to get a more favorable loan; and it's these investors who commit occupancy fraud but recorded as owner occupants in the data set increased the estimated hazard of default of the group of owner occupants, so we can not tell the difference between the investors and owner occupants in terms of default. To illustrate this, see the table as following:

	Record as Owner Occupants	Record as Investors
Real Owner Occupants	A(honest owner occupants)	B(no one)
Real Investors	C(Occupancy Fraud)	D(honest investors)

Apparently, no one would be in group B, because real owner occupants have no incentive to lie to become investor to get a less favorable loan. The problem comes from group C. As a group, investors are more risky than real owner occupants, but among this group, those who recorded as owner occupants are more risky than those investors who didn't commit occupancy fraud. With the number of group C increasing, the whole group as what we estimated as the owner occupants become as risky as group D. That's why by our estimation, we can not tell the difference between group A+C and group D in terms of riskiness.

To test whether our prediction about group C is correct or not, we will do some information asymmetry analysis in next section.

3.5 Information Asymmetry

3.5.1 the model

To test whether there is any information asymmetry during this process, there is a well developed two step procedure by Puelz and Snow (1994) and Kau et al (2012) . First we run that risk-based mortgage pricing model again,

$$r_i = X_i\beta + \epsilon_i, i = 1, \dots n$$

and get the residual $\hat{\epsilon_i}$ from this regression. Since the residual means the risk premium that can not be explained by our mortgage pricing model, or to be more clearly, if there are two individuals with the same characteristics, one accept an interest rate higher than the other, which can not be explained by the observable characteristics, we call that as the excess risk premium. It represents the extent of information asymmetry there.

Then the second step would be adding the excess risk premium into our competing risk model to see whether this excess risk premium, or the information asymmetry could explain the default hazard and the prepayment hazard which we observed in the data set.

$$L = \prod_{i=1}^{n} \prod_{j \in \{P,D\}} [\lambda_{j0}(t_i) exp\{x_i\beta + \gamma\hat{\epsilon}_i\}]^{d_{ij}} exp\left\{- \int_0^{t_i} \lambda_{j0}(t_i) exp\{x_i\beta + \gamma\hat{\epsilon}_i\}\, du\right\}$$

If this excess risk premium has some explanatory power on the competing risk model, then that's also some evidence for the occupancy fraud.

3.5.2 Estimation Result

Here, after adding the "excess risk premium" as an explanatory variable in the competing risk model to test whether this information asymmetry would have some prediction power over the future hazard of borrowers, we get some result in table 69, which showed that this "excess risk premium", or the infomration asymmetry do explain parts of the hazard, which affirms our suspect on the occupancy fraud. To test whether this information asymmetry has different effect on different groups, we devided our data into two groups, one is the group recorded as owner occupants, and the other is the group recorded as investor, then we did this information asymmetry test on both of these groups to see if there is any difference between these two groups in terms of the extent of information asymmetry, and the result is in table 70.

From this table, we find that among the group recorded as owner occupants, the information asymmetry problem is more serious than the group recorded as investors which reaffirmed our suspect in the last section that because in the group of owner occupants, it's not only true owner occupants in this group, but also those speculative investors who was trying to reduce their financing cost by commiting occupancy fraud. That's why this group include more information asymmetry than the group recorded as investors, which group is made up by honest investors.

3.5.3 Possible ways to prevent occupancy fraud

Here, we propose a "statistical scanning" way to help prevent the potential occupancy fraud. Since our ultimate goal would be to prevent the default, first, we will use historical data to "train" a competing risk model as

$$L = \prod_{i=1}^{n} \prod_{j \in \{P,D\}} [\lambda_{j0}(t_i) exp\left\{x_i\beta\right\}]^{d_{ij}} exp\left\{- \int_0^{t_i} \lambda_{j0}(t_i) exp\left\{x_i\beta\right\} du\right\}$$

then we will get the parameter estimation $\hat{\beta}$, after we get this estimation, we used this well-trained model to predict the cummulative probability of default

$$f_D(t,x) = \lim_{dt \to 0} \frac{Pr_D(t \leq T < t + dt|x)}{dt} = \lambda_D(t,x)S(t,x)$$

and if this probability is above some critical value, say q, then we should ask the loan officer to scrutinize this loan again. By this way, we don't have to scrutinize every loan, but we will significantly reduce the probability of future default.

3.6 Conclusion

By using submortgage data, we found that investors are being charged with a significant risk premium over owner occupants; besides that, they are also facing a more restricted loan; with the market getting hotter, this risk premium and restrictions are getting even worse. Being treated like that, our findings show that investors were actually not more risky than owner occupants in terms of both prepayment and default.

We suspect the reason for this puzzle is that when the market getting hotter, there are more speculative investors who commit occupancy fraud to get a more favorable loan. And these speculative investors were actually recorded as owner occupants on loan documents, which increased our estimation of the hazard of owner occupants group. And our information asymmetry test actually reaffirmed our suspect. Therefore, this paper, for the first time, give statistical evidence on occupancy fraud, and we also proposed a statistical scanning way to reduce to potential occupancy fraud.

For further study, we should check the robustness of our conclusion by using pool level data, or even with multinomial probit model. We can also run a logit regression to see the characteristic of tru e investors and using this model to find the potential occupancy fraud.

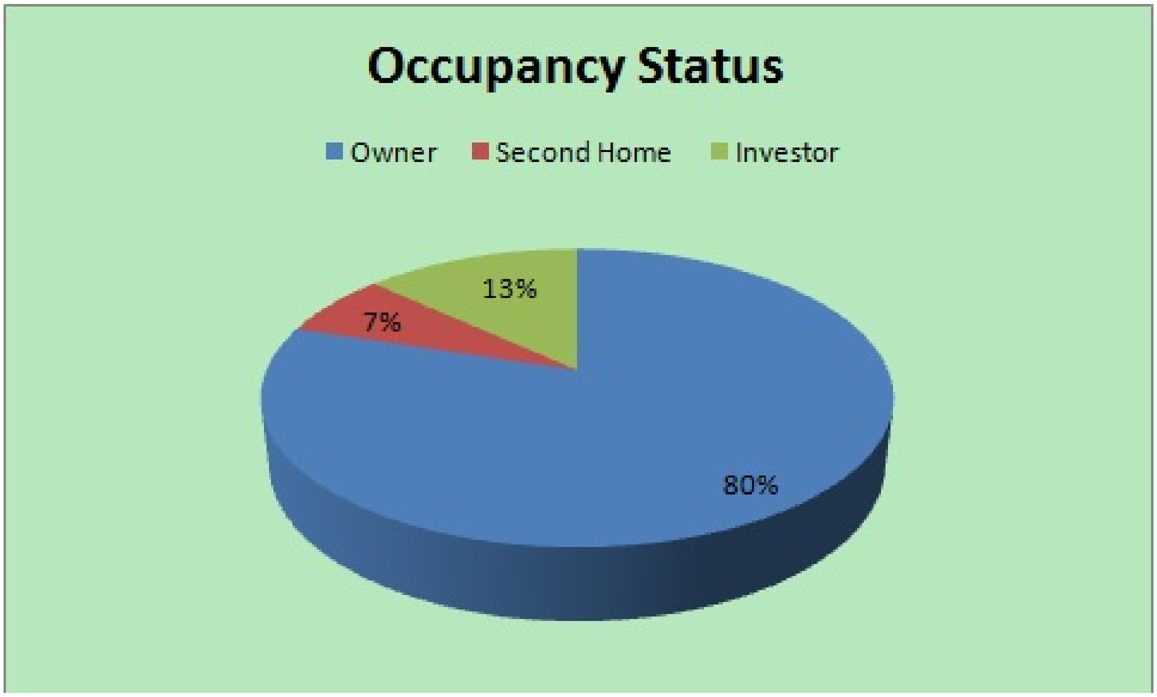

Figure 8: Occupancy status in our data

Table 62: Pre-assessment form by Universal American Mortgage Company

UAMC PRE-ASSESSMENT (FOR HOME PURCHASE ELIGIBILITY)

Builder: ____ Community: ____ Property State: ____
Estimated Sales Price: $ ____ Estimated Loan Amount: $ ____ Property Type: ____
Home will be: ☐ Primary ☐ Secondary ☐ Investment | Buyer Applying: ☐ Individually ☐ Jointly | ☐ FHA ☐ VA ☐ CONV | ☐ JUMBO ☐ ARM

GENERAL INFORMATION

	CUSTOMER 1	CUSTOMER 2
Married:	☐ Yes ☐ No — ☐ Separated	☐ Yes ☐ No — ☐ Separated
Name (First/Middle/Last/Suffix):		
Birthday (Month/Day/Year):		
Social Security Number:		
Current Address (Street):		
City/State/Zip:		
Phone:	( )	( )
Cell Phone:	( )	( )
Email:		
Time at Address:	____ Years ____ Months	____ Years ____ Months
	☐ Own ☐ Rent ☐ Rent Free	☐ Own ☐ Rent ☐ Rent Free
	Monthly Payment: $ ____	Monthly Payment: $ ____
U.S. Citizen:	☐ Yes ☐ No	☐ Yes ☐ No
Permanent Resident Alien:	☐ Yes ☐ No	☐ Yes ☐ No
Non-Resident Alien:	☐ Yes ☐ No	☐ Yes ☐ No

EMPLOYMENT

	CUSTOMER 1	CUSTOMER 2
	☐ Employed ☐ Self Employed	☐ Employed ☐ Self Employed
Current Employer:		
Length of Employment:	____ Years ____ Months	____ Years ____ Months
Monthly Income:	$	$
Monthly Debt Payment:	$	$

(Car Payment, Credit Cards, Alimony, Child Support, etc.) Have you or your co-borrower declared bankruptcy within the past 7 years? ☐ Yes ☐ No

$ DOWN

Down Payment Amount: $ ____ Sources: Sale of Present Home: $ ____
Savings: $ ____ Gift: $ ____ Sales of Stocks: $ ____
Other (Please indicate): ____
Do you need to sell your home in order to purchase this home? ☐ Yes ☐ No

AUTHORIZATION TO RELEASE INFORMATION AND TO ORDER CREDIT REPORT

By signing below, I/We hereby authorize Universal American Mortgage Company or Universal American Mortgage Company of California, herein referred to as "UAMC", to obtain my/our credit report for use in connection with my/our eligibility to be considered as a home purchaser.

X ____ X ____
Customer 1 Signature Date Customer 2 Signature Date

Joint credit application? Customer 1 and 2 each agree that they intend to apply for joint credit by initialing. **C1**: X ____ [Initial] **C2**: X ____ [Initial]

PERSONAL INFORMATION SHARING – SHARING WITH AFFILIATES

By law UAMC can and will share information about your pre-assessment status with their affiliated homebuilder without your permission in order to advise them of your purchase eligibility (i.e., information provided is satisfactory or more information is needed). Under the Fair Credit Reporting Act ("FCRA") UAMC is required to obtain your consent before they can share specific personal information that concerns your assets, income, and employment as well as credit reports and other credit related information ("FCRA-Covered Information") with their affiliated homebuilder.

I (we) agree to allow UAMC to share our FCRA-Covered Information with our affiliated homebuilder.

X ____ X ____
Customer 1 Signature Date Customer 2 Signature Date

UAMC CONTACT CONSENT

By signing below, I/We hereby authorize Universal American Mortgage Company to contact me/us regarding its mortgage loan products and services using any of the telephone number's (including cellular phones) and/or email addresses provided above unless otherwise specified here______________________________.

X ____ X ____
Customer 1 Signature Date Customer 2 Signature Date

EQUAL HOUSING LENDER

Universal American Mortgage Company, LLC; AZ: Mortgage Banker License #BK-0904844; CO: Supervised Lender License #987996; DE: Licensed Lender #8488; FL: Mortgage Lender License #ML0700915; IL: Residential Mortgage Licensee #MB0002095-2300 N. Barrington Road, Suite 750, Hoffman Estates, IL 60169; MD: Mortgage Lender License # 8101; MN: Mortgage Lender License # 8101; NV: Banker License #1310 and Mortgage Broker License #1055, 6750 Via Austi Parkway, Suite 130, Las Vegas, NV 89119 (702) 739-7933; NJ: Licensed by the NJ Department of Banking and Insurance #9933158, 800 West Main Street, Freehold, NJ 07728 (732) 625-2314. Licensee #L055199, 15550 Lighthouse Drive, Suite 200, Clearwater, Florida 33760 (727) 450-2733; NC: Mortgage Lender's License #L-108346; PA: Licensed by the PA Department of Banking pursuant to the Mortgage Act; SC: Mortgage Loan Broker # MB-0508401 and Supervised Lender License # S-5, 665, 688, 689, 690, S-6, 529, 572; TX: Regulated Loan License # 4514-34262; VA: Mortgage Lender/Broker License # MLB-817, licensed by the "Virginia State Corporation Commission"; Universal American Mortgage Company of California; CA: Licensed by the Department of Corporations under the CA Residential Mortgage Lending Act; NV: Banker License #3243 and Mortgage Broker License #3244, 10354 Professional Circle, Suite 120, Reno, NV 89521 (775)852-9980.
July 01, 2009

Table 63: Loan level data variable names and meaning

Variable Name	Meaning
Orig_Amt	Original Loan Amount
Inv_Pro	Dummy for Florida
Ini05	Dummy for loans initiated in 2005
Ini06	Dummy for loans initiated in 2006
Ini07	Dummy for loans initiated in 2007
FICO	FICO score
ldoc	Dummy for Low Doc and no doc status
LTV	Loan to Value Raio
Term 180	Loan term is 15 years or less
Term 240	Loan term is between 15 years and 20 years
Term480	Loan term is 40 years or more
Purchase	Dummy for Purchase (Not refinance)
Fixed	Dummy for Fixed rate loan product
Balloon	Dummy for balloon product
Age	Days from the initial date to the last payment
Vprep	Dummy for Voluntary Prepayment (not foreclosure)
ForeClosure	Dummy for foreclosure
Noncurrent	Dummy for Noncurrent states (prepaid, foreclosure, delinquincy)

Table 64: Pool level data variable name and meaning

Variable Name	Meaning
COUPON	dollar weighted average coupon rate of the pool
WAOLT	dollar weighted average of loan term
WAOCS	dollar weighted average of FICO score
OLTV	dollar weighted average of LTV ratio
WAOLS	dollar weighted average of loan size
Purchase	percentage of “purchasing” (not refinancing) borrowers in this pool
Investment	percentage of “investing” (not occupancy) borrowers in this pool

Table 65: The risk based mortgage pricing model

	Model 1		Model 2	
Dependent Variable	Interest Rate		Interest Rate	
Independent Variables	Coefficients	t-value	Coefficient	t-value
(Intercept)	11.400	74.095	11.410	74.095
Orig_Amt	-7.264E-07	-11.79	-7.250E-07	-11.769
Inv_pro	0.579	18.469	0.474	6.898
Florida	0.108	4.858	0.108	4.84
Ini05	0.455	5.951	0.451	5.851
Ini06	0.884	11.468	0.860	11.07
Ini07	0.870	10.307	0.838	9.738
FICO	-0.012	-68.023	-0.012	-68.025
Idoc	0.470	21.684	0.470	21.689
LTV	0.033	33.027	0.032	33.013
Term180	2.366	49.161	2.362	49.06
Term240	0.575	7.838	0.573	7.82
Term480	-0.701	-10.318	-0.704	-10.362
Purchase	0.255	11.063	0.258	11.179
Balloon	0.424	11.237	0.432	11.416
Inv_Pro:Ini05			0.058	0.702
Inv_Pro:Ini06			0.119	2.297
Inv_Pro:Ini07			0.253	1.952
Goodness of Fit				
R-square	0.4909		0.4911	
Adjusted R-square	0.4905		0.4906	

Table 66: The risk based mortgage pricing model

	Model 3		Model 4	
Dependent Variable	LTV Ratio		LTV Ratio	
Independent Variables	Coefficients	t-value	Coefficient	t-value
(Intercept)	42.190	45.086	42.210	45.042
Int_Rate	2.619	61.369	2.618	61.343
Orig_Amt	-8.148E-06	-18.315	-8.149E-06	-18.319
Inv_pro	-3.212	-13.852	-2.984	-5.856
Florida	-0.240	-1.456	-0.239	-1.449
Ini05	-2.320	-11.374	-2.378	-10.929
Ini06	-3.844	-18.119	-3.764	-16.803
Ini07	-3.339	-10.17	-3.091	-8.815
FICO	0.038	31.414	0.038	31.361
Idoc	-2.437	-15.124	-2.436	-15.119
Term180	2.366	49.161	2.362	49.06
Inv_Pro:Ini05			0.302	0.493
Inv_Pro:Ini06			-0.674	-1.049
Inv_Pro:Ini07			-1.866	-1.942
Goodness of Fit				
R-square	0.2112		0.2115	
Adjusted R-square	0.2108		0.211	

Table 67: The risk based mortgage pricing model

	Model 5		Model 6	
Dependent Variable	Interest Rate		LTV Ratio	
Independent Variables	Coefficients	t-value	Coefficient	t-value
(Intercept)	13.300	4.928	74.850	1.547
Term	-0.015	-2.004	0.054	0.407
FICO	-0.004	-5.346	-0.098	-7.919
LTV	0.020	7.543		
Purchase	-0.366	-3.54	19.060	12.052
Investment	2.198	10.96	-24.170	-6.367
Orig_Amt	-7.00E-07	-5.627	1.40E-05	6.493
coupon			6.021	7.543
Goodness of Fit				
R-square	0.4877		0.3972	
Adjusted R-square	0.4805		0.3887	

Table 68: Competing risk model to test the riskiness of investors

Competing Risk Model				
	Default Hazard		Prepayment Hazard	
Independent Variables	Coefficients	t-value	Coefficient	t-value
Orig_Amt	4.13E-07	4.136	7.09E-07	4.4243
Inv_pro	0.139	0.602	-0.006	-0.051
Int_Rate	0.183	14.870	0.234	7.251
Florida	0.03	0.647	-0.402	-4.660
Ini04	1.560	2.149	-0.247	-0.877
Ini05	2.970	4.097	-0.521	-1.898
Ini06	4.280	5.910	-0.677	-2.456
Ini07	5.340	7.318	-0.688	-2.198
FICO	-0.001	-4.447	-0.002	-2.944
Idoc	0.331	8.351	0.144	1.618
LTV	0.018	8.996	0.005	1.324
Term180	-4.490	-12.340	-0.725	-3.465
Term240	-3.520	-4.915	-0.125	-0.370
Term480	0.233	2.768	-0.259	-0.809
Purchase	0.127	3.089	0.143	1.517
Fixed	0.490	11.237	0.432	11.416
Balloon	0.105	1.882	0.085	0.632
Goodness of Fit				
Pseudo Log-likelihood	-27339		-5613	
Pseudo Likelihood ratio test	4205		169	

Table 69: Competing risk model to test information asymmetry

Competing Risk Model				
	Default Hazard		Prepayment Hazard	
Independent Variables	Coefficients	t-value	Coefficient	t-value
Orig_Amt	2.75E-07	2.75	5.36E-07	3.369
Inv_pro	0.256	0.8	0.132	1.058
Florida	0.047	1.01	-0.379	-4.383
Ini04	1.570	2.15	-0.246	-0.873
Ini05	3.080	4.26	-0.376	-1.381
Ini06	4.490	6.190	-0.429	-1.575
Ini07	5.550	7.6	-0.441	-1.4176
FICO	-0.004	-11.97	-0.005	-7.440
Idoc	0.419	10.65	0.256	2.943
LTV	0.023	12.17	0.013	3.34
Term180	-4.060	-11.250	1.173	-0.914
Term240	-3.410	-4.77	0.015	0.0437
Term480	0.093	1.12	-0.437	-1.3553
Purchase	0.174	4.24	0.202	2.131
Balloon	0.176	3.15	0.175	1.301
Excess Premium	0.187	15.3	0.235	7.152
Goodness of Fit				
Pseudo Log-likelihood	-27335		-5613	
Pseudo Likelihood ratio test	4212		169	

Table 70: Information asymmetry test on both groups

Competing Risk Model for Default Hazard				
	Owner Occupants		Investor	
Independent Variables	Coefficients	t-value	Coefficient	t-value
Orig_Amt	-2.35E-07	-4.12	1.22E-07	0.938
Florida	-0.252	-13.101	-0.066	-1.439
Ini04	0.052	0.608	0.246	-0.827
Ini05	-0.221	-2.633	0.300	1.024
Ini06	-0.287	-3.433	0.393	1.336
Ini07	0.09	1.058	0.744	2.491
FICO	-0.002	-14.418	-0.001	-3.512
Idoc	0.004	0.239	-0.040	-0.829
LTV	0.002	1.799	0.004	1.513
Term180	0.580	14.433	0.568	5.773
Term240	0.304	4.547	0.227	1.223
Term480	-0.042	-0.851	-0.120	-1.071
Purchase	-0.154	-7.961	0.126	2.368
Balloon	-0.121	-4.226	-0.106	-1.134
Excess Premium	0.113	8.079	0.030	4.989
Goodness of Fit				
Pseudo Log-likelihood	-122943		-13944	
Pseudo Likelihood ratio test	1292		142	

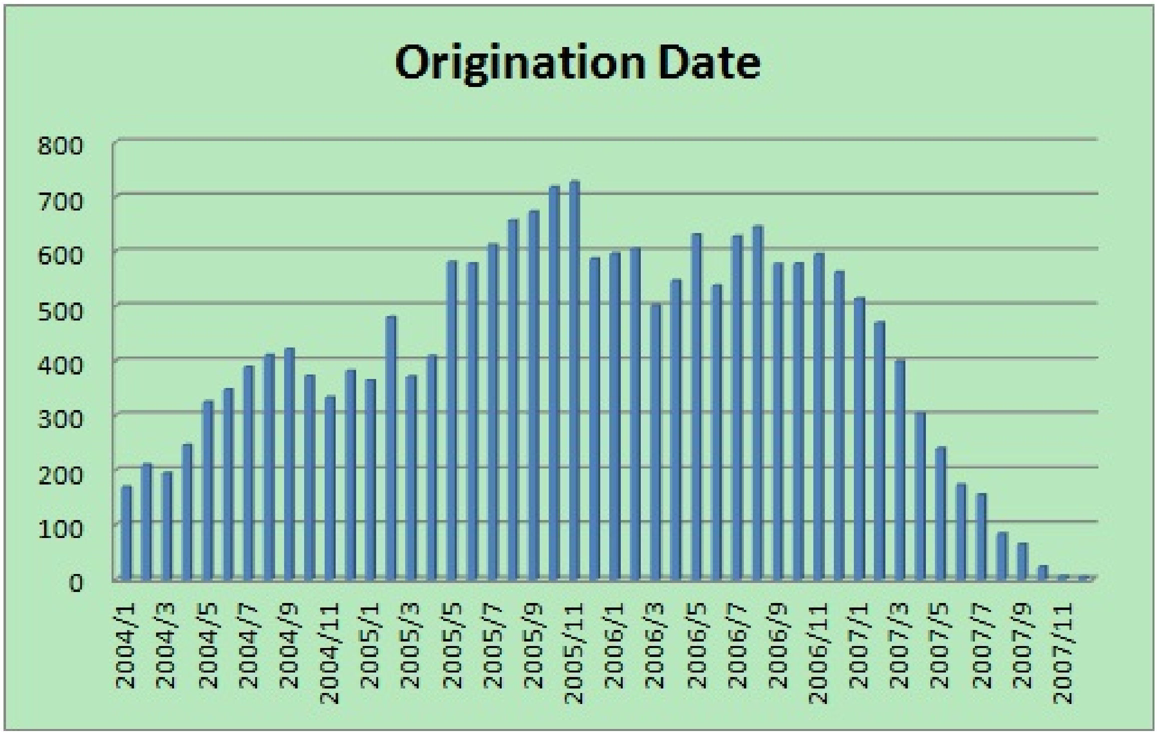

Figure 9: Origination date distribution in our data

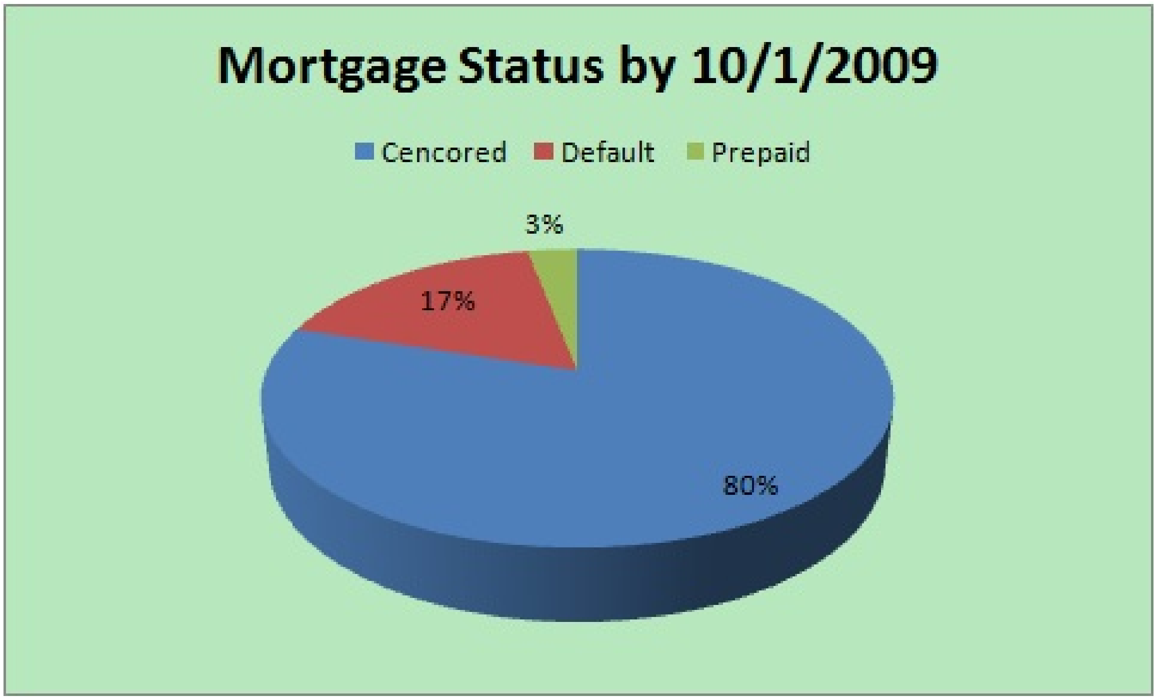

Figure 10: Mortgage status by the right censor date

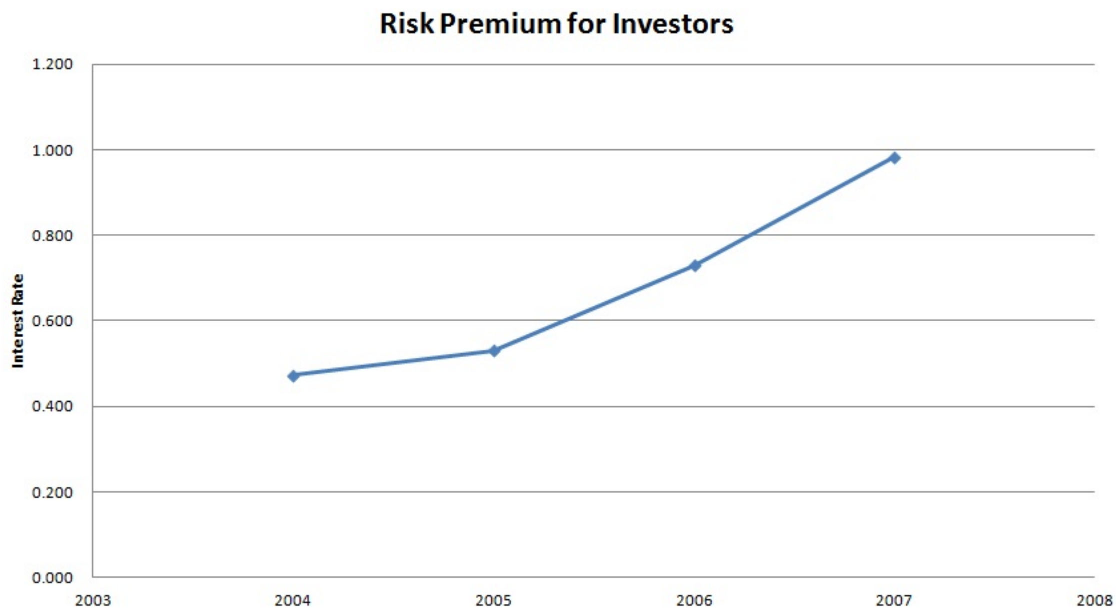

Figure 11: Risk premium for investors over the time

Figure 12: Restrictions for investors (LTV difference) over the time

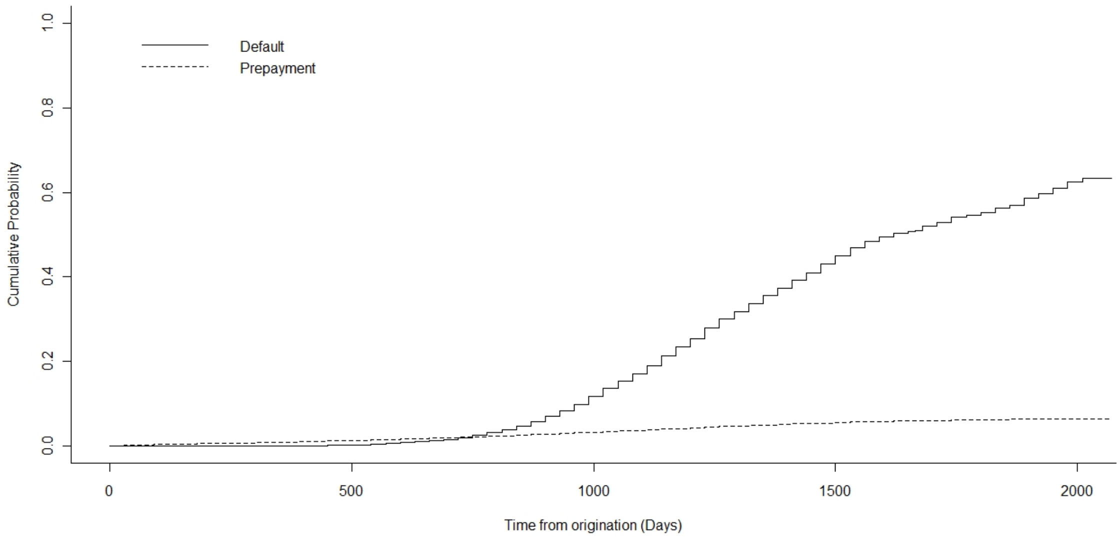

Figure 13: Cummulative probability for default and prepayment

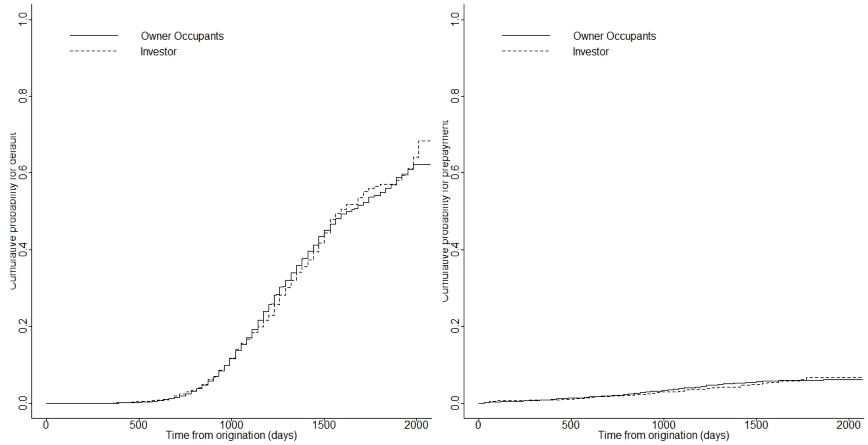

Figure 14: The difference between owner occupants and investors

REFERENCES

[Alizadeh et al., 2002] Alizadeh, S., Brandt, M. W., and Diebold, F. X. (2002). Range-based estimation of stochastic volatility models. *The Journal of Finance*, 57(3):pp. 1047–1091.

[Ambrose et al., 2004] Ambrose, B. W., LaCour-Little, M., and Sanders, A. B. (2004). The effect of conforming loan status on mortgage yield spreads: A loan level analysis. *Real Estate Economics*, 32(4):541 – 569.

[Andersen et al., 2001] Andersen, T. G., Bollerslev, T., Diebold, F. X., and Ebens, H. (2001). The distribution of realized stock return volatility. *Journal of Financial Economics*, 61(1):43–76.

[Ang and Bekaert, 2007] Ang, A. and Bekaert, G. (2007). Stock return predictability: Is it there? *Review of Financial Studies*, 20(3):651–707.

[Archer et al., 1996] Archer, W. R., Ling, D. C., and McGill, G. A. (1996). The effect of income and collateral constraints on residential mortgage terminations. *Regional Science and Urban Economics*, 26(34):235 – 261.

[Bandi and Perron, 2006] Bandi, F. M. and Perron, B. (2006). Long memory and the relation between implied and realized volatility. *Journal of Financial Econometrics*, 4(4):636–670.

[Barndorff-Nielsen and Shephard, 2002] Barndorff-Nielsen, O. E. and Shephard (2002). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society Series B*, 64(2):253–280.

[Barndorff-Nielsen and Veraart, 2012] Barndorff-Nielsen, O. E. and Veraart, A. E. (2012). Stochastic volatility of volatility and variance risk premia. *Journal of Financial Econometrics*, 11(1):1–46.

[Bloomfield, 1973] Bloomfield, P. (1973). An exponential model for the spectrum of a scalar time series. *Biometrika*, 60(2):pp. 217–226.

[Bollerslev et al., 2013] Bollerslev, T., Osterrieder, D., Sizova, N., and Tauchen, G. (2013). Risk and return: Long-run relations, fractional cointegration, and return predictability. *Journal of Financial Economics*, 108(2):409 – 424.

[Bollerslev et al., 2009] Bollerslev, T., Tauchen, G., and Zhou, H. (2009). Expected stock returns and variance risk premia. *Review of Financial Studies*, 22(11):4463–4492.

[Bollerslev and Todorov, 2011] Bollerslev, T. and Todorov, V. (2011). Tails, fears, and risk premia. *The Journal of Finance*, 66(6):2165–2211.

[Britten-Jones and Neuberger, 2000] Britten-Jones, M. and Neuberger, A. (2000). Option prices, implied price processes, and stochastic volatility. *The Journal of Finance*, 55(2):839–866.

[Calhoun and Deng, 2002] Calhoun, C. A. and Deng, Y. (2002). A dynamic analysis of fixed- and adjustable-rate mortgage terminations. *The Journal of Real Estate Finance and Economics*, 24:9–33.

[Campbell and Shiller, 1988] Campbell, J. Y. and Shiller, R. J. (1988). The dividend-price ratio and expectations of future dividends and discount factors. *Review of financial studies*, 1(3):195–228.

[Campbell and Dietrich, 1983] Campbell, T. S. and Dietrich, J. K. (1983). The determinants of default on insured conventional residential mortgage loans. *Journal of Finance*, 38(5):1569 – 1581.

[Carr and Wu, 2006] Carr, P. and Wu, L. (2006). A tale of two indices. *Journal of Derivatives*, 13(3):13 – 29.

[Carr and Wu, 2009] Carr, P. and Wu, L. (2009). Variance risk premiums. *Review of Financial Studies*, 22(3):1311–1341.

[Chernov, 2007] Chernov, M. (2007). On the role of risk premia in volatility forecasting. *Journal of Business & Economic Statistics*, 25(4).

[Cheung and Lai, 1993] Cheung, Y.-W. and Lai, K. S. (1993). A fractional cointegration analysis of purchasing power parity. *Journal of Business & Economic Statistics*, 11(1):pp. 103–112.

[Christensen and Prabhala, 1998] Christensen, B. J. and Prabhala, N. R. (1998). The relation between implied and realized volatility. *Journal of Financial Economics*, 50(2):125–150.

[Clapp et al., 2001] Clapp, J. M., Goldberg, G. M., Harding, J. P., and LaCour-Little, M. (2001). Movers and shuckers: Interdependent prepayment decisions. *Real Estate Economics*, 29(3):411–450.

[Corsi et al., 2008] Corsi, F., Mittnik, S., Pigorsch, C., and Pigorsch, U. (2008). The volatility of realized volatility. *Econometric Reviews*, 27(1-3):46–78.

[Cunningham and Capone, 1990] Cunningham, D. F. and Capone, Charles A., J. (1990). The relative termination experience of adjustable to fixed-rate mortgages. *The Journal of Finance*, 45(5):1687–1703.

[Demeterfi et al., 1999] Demeterfi, K., Derman, E., Kamal, M., and Zou, J. (1999). A guide to volatility and variance swaps. *The Journal of Derivatives*, 6(4):9–32.

[Deng et al., 2000] Deng, Y., Quigley, J. M., and Order, R. V. (2000). Mortgage terminations, heterogeneity and the exercise of mortgage options. *Econometrica*, 68(2):275–308.

[Depken et al., 2009] Depken, C. A., Hollans, H., and Swidler, S. M. (2009). An empirical analysis of residential property flipping. *Journal of Real Estate Finance and Economics, Vol. 39, No. 3, 2009*, pages 248–263.

[Drechsler and Yaron, 2011] Drechsler, I. and Yaron, A. (2011). What's vol got to do with it. *Review of Financial Studies*, 24(1):1–45.

[Engle and Granger, 1987] Engle, R. F. and Granger, C. W. J. (1987). Co-integration and error correction: Representation, estimation, and testing. *Econometrica*, 55(2):pp. 251–276.

[Epstein and Zin, 1989] Epstein, L. G. and Zin, S. E. (1989). Substitution, risk aversion, and the temporal behavior of consumption and asset returns: A theoretical framework. *Econometrica: Journal of the Econometric Society*, pages 937–969.

[Fine and Gray, 1999] Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94(446):pp. 496–509.

[Fleming and Kirby, 2011] Fleming, J. and Kirby, C. (2011). Long memory in volatility and trading volume. *Journal of Banking & Finance*, 35(7):1714 – 1726.

[Gallant et al., 1999] Gallant, A. R., Hsu, C.-T., and Tauchen, G. (1999). Using daily range data to calibrate volatility diffusions and extract the forward integrated variance. *The Review of Economics and Statistics*, 81(4):pp. 617–631.

[Gil-Alana, 2000] Gil-Alana, L. (2000). Evaluation of robinson's (1994) tests in finite samples. *Journal of Statistical Computation and Simulation*, 68(1):39–63.

[Green and Shoven, 1986] Green, J. and Shoven, J. B. (1986). The effects of interest rates on mortgage prepayments. *Journal of Money, Credit & Banking*, 18(1):41 – 59.

[Guo, 1996] Guo, D. (1996). The predictive power of implied stochastic variance from currency options. *Journal of Futures Markets*, 16(8):915–942.

[Hendershott and Shilling, 1989] Hendershott, P. H. and Shilling, J. D. (1989). The impact of the agencies on conventional fixed-rate mortgage yields. *The Journal of Real Estate Finance and Economics*, 2:101–115.

[Jiang and Tian, 2005] Jiang, G. J. and Tian, Y. S. (2005). The model-free implied volatility and its information content. *Review of Financial Studies*, 18(4):1305–1342.

[Jones, 2003] Jones, C. S. (2003). The dynamics of stochastic volatility: evidence from underlying and options markets. *Journal of econometrics*, 116(1):181–224.

[Julliard and Ghosh, 2012] Julliard, C. and Ghosh, A. (2012). Can rare events explain the equity premium puzzle? *Review of Financial Studies*, 25(10):3037–3076.

[Kau et al., 2012] Kau, J., Keenan, D., Lyubimov, C., and Slawson, V. (2012). Asymmetric information in the subprime mortgage market. *Journal of Real Estate Finance & Economics*, 44(1/2):67 – 89.

[Lamont, 1998] Lamont, O. (1998). Earnings and expected returns. *The Journal of Finance*, 53(5):1563–1587.

[Lancaster, 1990] Lancaster, T. (1990). *The Econometric Analysis of Transition Data*. Econometric Society Monographs. Cambridge University Press.

[Lefcoe, 2009] Lefcoe, G. (2009). How 'spec' condo and tract home buyers helped sink our housing and finance markets: Should the alienability of their interests be restrained by law? *Journal of Legis*, pages 1–17.

[Lettau and Ludvigson, 2001] Lettau, M. and Ludvigson, S. (2001). Consumption, aggregate wealth, and expected stock returns. *the Journal of Finance*, 56(3):815–849.

[Liu et al., 2005] Liu, J., Pan, J., and Wang, T. (2005). An equilibrium model of rare-event premia and its implication for option smirks. *Review of Financial Studies*, 18(1):131–164.

[Nielsen and Shimotsu, 2007] Nielsen, M. O. and Shimotsu, K. (2007). Determining the cointegrating rank in nonstationary fractional systems by the exact local whittle approach. *Journal of Econometrics*, 141(2):574–596.

[Pavlov, 2001] Pavlov, A. D. (2001). Competing risks of mortgage termination: Who refinances, who moves, and who defaults? *The Journal of Real Estate Finance and Economics*, 23:185–211.

[Phillips and Shimotsu, 2004] Phillips, P. C. B. and Shimotsu, K. (2004). Local whittle estimation in nonstationary and unit root cases. *The Annals of Statistics*, 32(2):pp. 656–692.

[Poon and Granger, 2003] Poon, S.-H. and Granger, C. W. (2003). Forecasting volatility in financial markets: A review. *Journal of Economic Literature*, 41(2):478–539.

[Puelz and Snow, 1994] Puelz, R. and Snow, A. (1994). Evidence on adverse selection: Equilibrium signaling and cross-subsidization in the insurance.. *Journal of Political Economy*, 102(2):236.

[Robinson, 1994] Robinson, P. M. (1994). Efficient tests of nonstationary hypotheses. *Journal of the American Statistical Association*, 89(428):pp. 1420–1437.

[Robinson and Yajima, 2002] Robinson, P. M. and Yajima, Y. (2002). Determination of cointegrating rank in fractional systems. *Journal of Econometrics*, 106(2):217–241.

[Schwert, 1989] Schwert, G. W. (1989). Why does stock market volatility change over time? *The journal of finance*, 44(5):1115–1153.

[Shimotsu, 2010] Shimotsu, K. (2010). Exact local whittle estimation of fractional integration with unknown mean and time trend. *Econometric Theory*, 26(02):501–540.

[Shimotsu, 2012] Shimotsu, K. (2012). Exact local whittle estimation of fractionally cointegrated systems. *Journal of Econometrics*, 169(2):266–278.

[Simon and Corkery, 2008] Simon, R. and Corkery, M. (2008). Speculators may have accelerated housing downturn; rising number of defaults also could complicate effort to help homeowners. *Wall Street Journal*, page 8.

[Weil, 1989] Weil, P. (1989). The equity premium puzzle and the risk-free rate puzzle. *Journal of Monetary Economics*, 24(3):401–421.

[Xudong. An and Yao, 2012] Xudong. An, Andy Q. Do, E. R. and Yao, V. W. (2012). Information asymmetries, excess premium and subprime mortgage default. *Working Paper*.