# *FORECASTING REALISED VOLATILITY AND VALUE-AT-RISK BACKTESTING*

*ACFI233 – Econometrics for Finance II*

*Date: 09/05/2025 | Word Count: 1956*

# Table of Contents

***Executive Summary* .......... 2**

Methodology .......... 2

Key Findings .......... 3

***Introduction* .......... 4**

***Empirical Analysis and Results* .......... 4**

Unbiased Test of VIX$^2$ .......... 4

Summary stats and Normality .......... 5

In-sample estimation of HAR and SHAR models .......... 6

Out-of-Sample Forecasting and Model averaging .......... 7

Value-at-Risk Forecasting and Back testing .......... 8

***Conclusion* .......... 9**

# Executive Summary

## Methodology

### Data

This report is based of analysis conducted in Python (VS code). The analysis uses daily returns for the SPDR S&P 500 ETF Trust, spanning from 16/11/2000 and ends 16/11/2023. The dataset includes the following all expressed as percentages.

- Realised variance (RV)
- Realised Quarticity (RQ)
- Realised Positive Semi variance (RSP)
- Realised Negative Semi variance (RSN)
- VIX$^2$ Index (VIX2)
- Close to Close Daily returns (Rt)

### Unbiased Test of VIX$^2$

- Employed a simple linear regression - $RV_t = \alpha + \beta VIX_t^2 + \epsilon_t$
- Individual t-test to hypothesise that $\alpha$ = 0 and $\beta = 1$

$$t_\alpha = \frac{\hat{\alpha} - 0}{SE(\widehat{\alpha})} \; AND \; t_\beta = \frac{\hat{\beta} - 1}{SE(\widehat{\beta})}$$

- Joint Wald test to hypothesise $\alpha$ = 0 and $\beta = 1$,

$$W = (R\hat{\beta} - r)^T \; [R\hat{V}R^T]^{-1}(R\hat{B} - r)$$

- The analysis was repeated using square root and logarithmic transformations of RV and VIX$^2$, to address non-normality and stabilise the variance.

### Summary Statistics and Normality

- Computation of summary statistics for RV and VIX$^2$, their square-root and log-transformed forms using Python functions.
- Jarque-Bera test at 5% significance level.

### In-sample estimation of HAR and SHAR models

- Using OLS with Newy-west standard errors (10 lags), Daily, Weekly and monthly RV were used as regressors, as well as RSP and RSN in Shar models.

$$RVt + 1 = \beta 0 + \beta d \; RV_t^{(d)} + \beta w \; RV_t^{(w)} + \beta m \; RV_t^{(m)} + \varepsilon t + 1$$

### Out-of-sample Forecasting and model averaging

- Rolling window of 1500 days, generate one-step ahead forecasts of RV using HAR-log and SHAR-sqrt models.

Forecast evaluation:

- Mincer-Zarnowitz to test unbiasedness
- Diebold-Mariano to test accuracy.

$$DM = \frac{\bar{d}}{\sqrt{\hat{\sigma}_d^2/T}} \; where \; d_t = L_1(t) - L_2(t)$$

### Value-at-Risk Forecasting and Back testing

- VaR forecasts at 1% and 5%, constructed from RV forecasts

Back testing includes:

- Proportion of failures test, Traffic light test

## Key Findings

### Unbiased Test of VIX$^2$

- The **none** and **sqrt transformation**: $\beta$close to 1 but $\alpha$ significantly different to 0. Suggesting VIX$^2$ is biased.
- **Log transformation**: $\beta$ different from 0 and $\alpha$ different from 1 - unbiased, but joint significance ($p<0.01$). Log best transformation for explanatory power.

### Summary stats and Normality

- All series are non-normally distributed ($p<0.05$)
- High skewness and Kurtosis - especially in RV and √RV

### In-sample estimation of HAR and SHAR models

- HAR-log is the best in-sample performance (highest R$^2$adj = 0.7422)
- SHAR-log also strong (0.6792)

### Out-of-sample Forecasting and model averaging

- MZ tests: HAR-log significant $\beta$>1, SHAR-sqrt $\beta$ not significantly different from 1
- Model average $\beta$=1.14 which is close, but t-stat = 1.76
- The model average offers a good compromise, moderate bias and strong R$^2$

### Value-at-Risk Forecasting and Back testing

- All models overpredict failures – VaR is too tight, PoF p-values = 0.000 - reject null of correct failure rate, Traffic light = RED across all forecasts and confidence levels.
- VaR models are not suitable for capital allocation or risk limit decisions

# Introduction

In financial risk management and asset pricing, accurate forecasting of volatility is very important, as volatility reflects market uncertainty and directly impacts investment decisions. Specifically, realised volatility offers an accurate reflection of actual market volatility compared to traditional methods.

Reliable volatility forecasting is important in today's complex and interconnected markets. Effective forecasting model can help investors reduce risk and optimise investment strategies resulting in more profitable/safer trading.

This report examines Heterogeneous Autoregressive (HAR) family of models, which is a widely recognised framework for forecasting realised volatility (RV). The analysis examines the accuracy and predictive power of various HAR models, and will assess practical use through, VaR back testing. This well help evaluate their ability to provide reliable measure of market risk, especially under times of uncertainty and economic turbulence.

# Empirical Analysis and Results

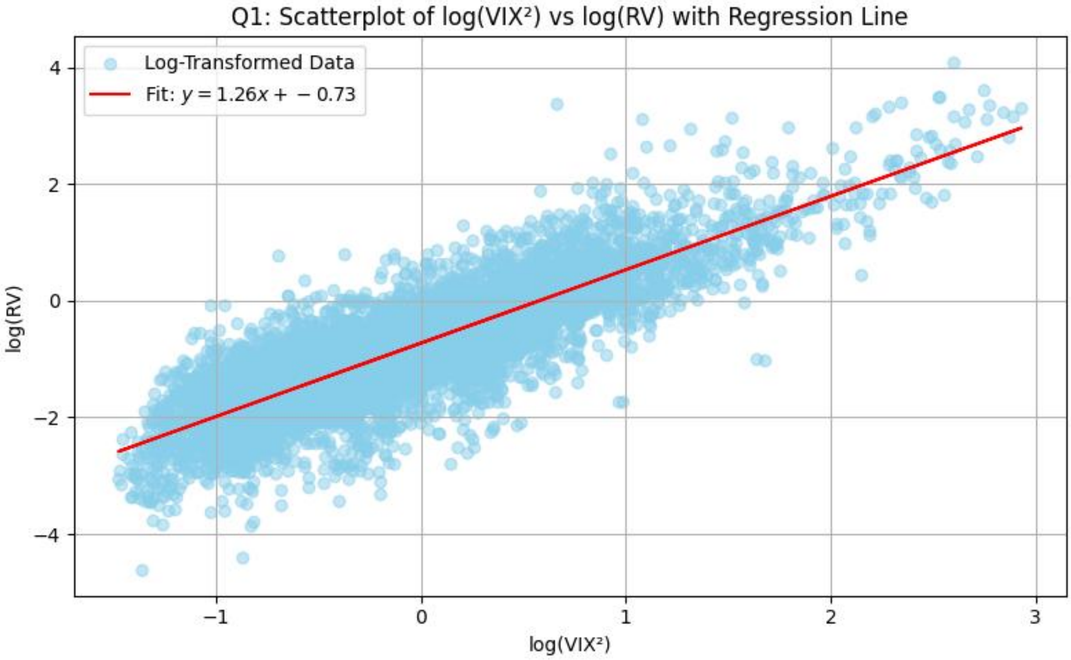
## Unbiased Test of VIX$^2$

To assess whether the VIX$^2$ index is an unbiased estimator of future realised variance, we employed a simple linear regression, with null hypothesis of unbiasedness requires $\alpha$ = 0 and $\beta = 1$.

| | $\hat{\alpha}$ | $t(\alpha = 0)$ | $\hat{\beta}$ | $t(\beta = 0)$ | *Wald Stat* | *p-value* |
|---|---|---|---|---|---|---|
| *None* | -0.5587 | -4.59 | 1.1464 | 1.30 | 341.21 | 0.0000 |
| *Square-Root* | -0.2893 | -7.15 | 1.0326 | 0.75 | 1648.65 | 0.0000 |
| *Logarithmic* | -0.7268 | -41.76 | 1.2584 | 11.37 | 1032.02 | 0.0000 |

The results (using Newey-west standard errors) indicate that for none transformed, alpha is significantly different from zero at both 1% and 5% levels, while beta is not significantly different from zero (p=0.1945). However, the joint Wald test rejects the null hypothesis, suggesting overall bias.

The square-foot model also rejects the joint null, although beta remains statistically close to 1. The log model produces significant deviations for both coefficients, with the strongest joint rejection. These findings suggest that VIX$^2$ is not an unbiased predictor of RV, the log transformed model captures the strongest statistical relationship. The following plot shows the relationship between log (VIX$^2$) and RV, supporting the regression analysis.



## Summary stats and Normality

Descriptive statistics were computed for RV, VIX$^2$ and their transformed versions, to help understand distributional properties and suitability for linear modelling.

| | *Mean* | *Std dev* | *Skewness* | *Kurtosis* | *JB p-value* |
|---|---|---|---|---|---|
| *RV* | 0.9313 | 2.1912 | 10.5393 | 171.5540 | 0.0 |
| *VIX$^2$* | 1.2998 | 1.4777 | 4.8514 | 37.3017 | 0.0 |
| *√RV* | 0.7902 | 0.5541 | 3.4714 | 24.0085 | 0.0 |
| *√VIX2* | 1.0454 | 0.4550 | 2.1875 | 10.7294 | 0.0 |
| *log (RV)* | -0.8008 | 1.0997 | 0.4234 | 3.4696 | 0.0 |
| *log (VIX2)* | -0.0588 | 0.7361 | 0.6855 | 3.4166 | 0.0 |

The output shows both RV and VIX$^2$ are positively skewed and show high kurtosis, which means there are extreme values and fait tails in the distributions. The standard deviation is relatively high compared to the meaning, hinting volatility. These patterns persist under transformation although reduced magnitude, as this is the reason we transform the variables to reduce skewness, stabilise variance and makes the distribution closer to normal.

To test the normality, we use the Jarque-Bera test at 5%, the test shows highly significance across all series (p-value =0.0), which confirms the rejection of the normality assumption. These results underscore the importance of using robust standard errors and non-linear transformations when modelling volatility.

## In-sample estimation of HAR and SHAR models

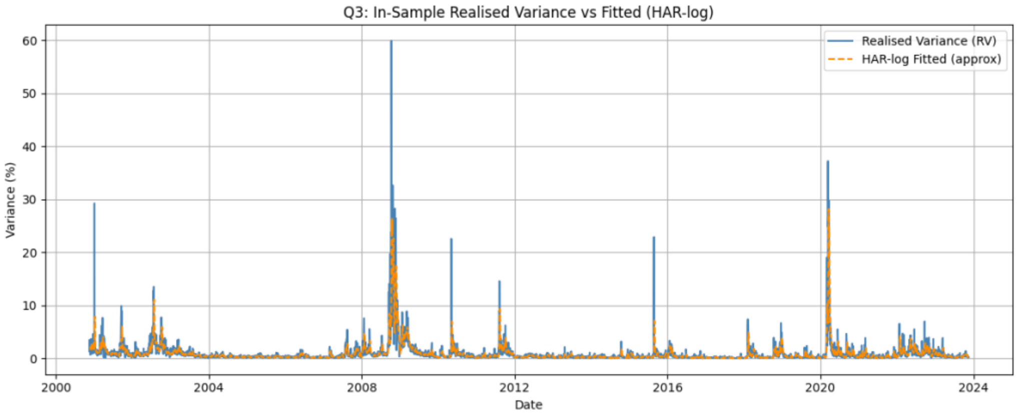
To capture the persistence and heterogeneity of volatility over time, we estimate the HAR model, and its extension SHAR model, using daily, weekly and monthly realised variance. These models investigate if past volatility has an influence on future volatility.

| | *Intercept* | *T-stat* | *Daily* | *T-stat* | *Weekly* | *T-stat* | *Monthly* | *T-stat* | *Adj.R$^2$* |
|---|---|---|---|---|---|---|---|---|---|
| | | | *Term* | | *Term* | | *Term* | | |
| *HAR* | 0.1069 | 3.22 | 0.3997 | 5.09 | 0.4237 | 4.47 | 0.0600 | 0.95 | 0.5528 |
| *SHAR* | 0.1124 | 2.70 | -0.0035 | -0.03 | 0.9057 | 5.01 | 0.3226 | 4.22 | 0.5297 |
| *√HAR* | 0.0495 | 3.87 | 0.5411 | 17.26 | 0.3246 | 6.90 | 0.0709 | 1.96 | 0.7299 |
| *√SHAR* | 0.0667 | 3.70 | 0.0913 | 1.83 | 0.4917 | 10.36 | 0.3715 | 6.70 | 0.6765 |
| *log (Har)* | -0.0438 | -4.32 | 0.5480 | 31.07 | 0.2799 | 10.99 | 0.1193 | 5.44 | 0.7422 |
| *log (SHAR)* | 0.1823 | 7.26 | 0.0881 | 4.26 | 0.2496 | 13.57 | 0.3813 | 9.72 | 0.6792 |

The results, summarised above, shows that most of the daily, all weekly and monthly coefficients are positive and statistically significant, aligning with wider financial theory volatility clustering and memory behaviour, which suggests past volatility does impact future volatility. The Daily SHAR term is insignificant, suggesting limited marginal contribution from daily semi variances.

When looking at the performance of the model, log (Har) has the highest Adj.R$^2$, suggesting a high explanatory power. √HAR also performs well (Adj.R$^2$ = 0.7299). Shar models tend to underperform compared to the HAR models, this is due to the increased complexity of the model, but SHAR is useful for capturing asymmetric volatilities.

Again, the performance improved under transformed specifications, suggesting the nonlinear transformations improve model fit. These patterns reinforce the relevance of using multi horizon volatility models. The plot shows the in-sample fitted values from log (HAR) compared to actual RV, validating the model's ability to capture volatility trends.



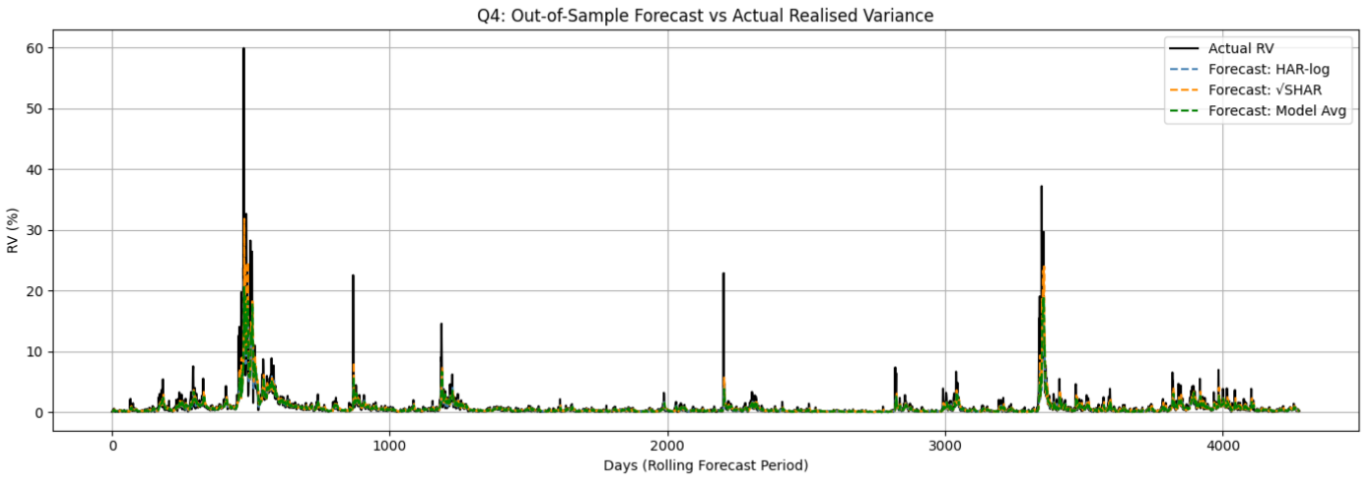
## Out-of-Sample Forecasting and Model averaging

Out-of-sample modelling gives us a more realistic assessment of how well our modules will perform in practice. To assess the predictive ability of the models, a rolling window approach is used to generate a one step ahead forecast of RV. Two models are selected based on in-sample fit: the HAR model with log transformation, and the SHAR model with the square-root transformation.

| | *MZ Beta* | *MZ p-value* | *MSE* | *QLIKE* | *$R^2$* |
|---|---|---|---|---|---|
| *Log (HAR)* | 1.2527 | 0.0016 | 2.5035 | 0.3248 | 0.5703 |
| *√SHAR* | 0.9690 | 0.6642 | 2.7647 | 0.3740 | 0.5255 |
| *Model Average* | 1.1403 | 0.0790 | 2.5127 | 0.3268 | 0.5687 |

From the Mincer-Zarnowitz test regression, we see that log (HAR) has a significant beta different from 1, suggesting it over-predicts RV but shows strong predictability. It also has the best explanatory power. The √SHAR model, performs worse across most metrics, it suggests it may be unbiased but lacking predictive power.

The model average shows a more balanced trade-off between bias and efficiency. It has the lowest QLIKE and near minimal MSE, indicating the forecasting combination improves predictive accuracy. It also has a MZ beta closer to one suggesting it is a less biased forecast overall. Although none are perfectly unbiased, the model average shows improved balanced between over and under forecasting and shows a more robust forecast for an effective strategy. However overall, there is very little difference between a lot of the metrics across the board. The figure below compares one step ahead out of sample forecasts from HAR, SHAR and model average against actual RV.



## Value-at-Risk Forecasting and Back testing

VaR is a widely used risk management tool that estimates the maximum expected loss over a given time. Using the out of sample RV, log (HAR) and √SHAR models we calculate VaR at 1% and 5% significance levels. This helps investors quantify potential losses to help make better informed decisions.

*1% significance level*

| | Obs Level | Failure | Expected | Fail rate | Excess Fail | Excess Ratio | Traffic Light |
|---|---|---|---|---|---|---|---|
| *Log (HAR)* | 0.9439 | 240 | 42.76 | 0.0561 | 197.24 | 5.61 | 🔴 |
| *√SHAR* | 0.9495 | 216 | 42.76 | 0.0505 | 173.24 | 5.05 | 🔴 |
| *Model Avg* | 0.0481 | 222 | 42.76 | 0.0519 | 179.24 | 5.19 | 🔴 |

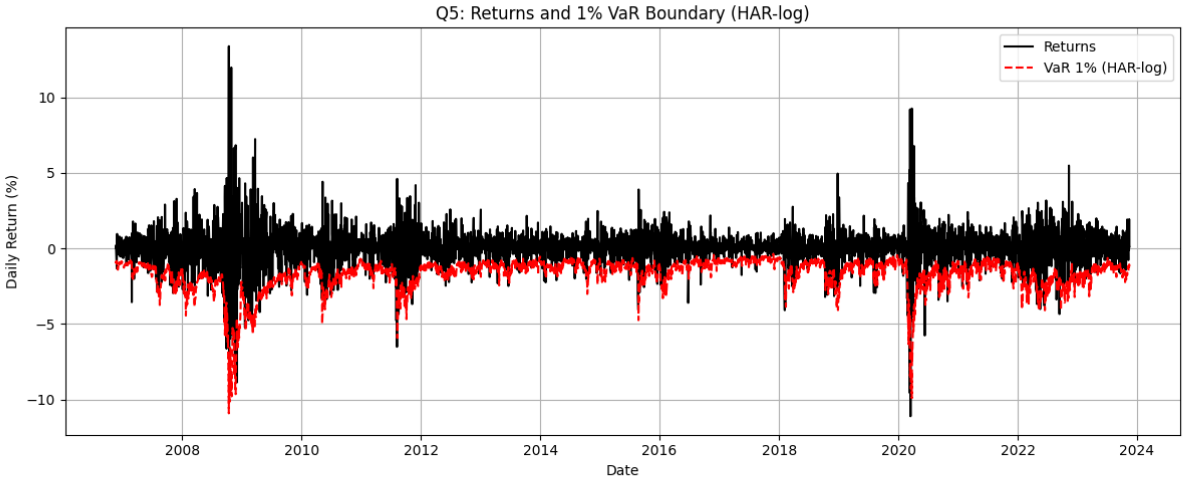
The 1% significance level means 99% of the time this estimation should be correct. The output shows all models under play risks, with all p-values = 0 showing strong statistical rejection of correct average. The Log (HAR) model had 240 failures (5.61%), which was higher than the SHAR and the model average failures, each exceeding the expected 42.76, meaning all models violated the acceptable threshold. The excess ratios all above 5, turned on the red light in the traffic light test, signalling poor VaR reliability. Additionally, in all cases PoF tests strongly against the null hypothesis, confirming observed violations are significantly greater than the expected.

*5% Significance Level*

| | Obs Level | Failure | Expected | Fail rate | Excess Fail | Excess Ratio | Traffic Light |
|---|---|---|---|---|---|---|---|
| *Log (HAR)* | 0.8870 | 483 | 213.80 | 0.1130 | 197.24 | 2.26 | 🔴 |
| *√SHAR* | 0.8987 | 433 | 213.80 | 0.1013 | 173.24 | 2.03 | 🔴 |
| *Model Avg* | 0.8954 | 447 | 213.8 | 0.1045 | 233.2 | 2.09 | 🔴 |

At 5% level of significance, the VaR forecasts again show under coverage of risk across all three models. The HAR model saw 483 failures (11.30%) more than double the expected, leading to an excess ratio of 2.26, indicating a serious underestimation of potential losses. Likewise did the SHAR model. The average model, which combines both, resulted in less failures than HAR, but more than SHAR, representing a slight improvement over the individual models, it still exceeds the expected failure count. All three models fail the traffic light test, signalling their unsuitability for risk management. Similarly, the PoF test at 5% shows significant results across all models but indicates underestimation of risk and models fail to meet back testing standards.

The Following figure overlays actual returns with 1% VaR forecasts from HAR model, providing assessment of risk violations.



# Conclusion

This report dived into the accuracy and practical application of the HAR models in forecasting realised volatility for the S&P 500. We used in-sample and out-of-sample models, we found the log (HAR) model demonstrated strong predictive power, while SHAR offered marginal gains in capturing asymmetric volatility effects. However, model averaging was the most appropriate and balanced approach, as it reduced forecasting errors and improved performance overall.

We found that the use of transformed variables was justified through statistical testing and normality diagnostics as it enhanced model efficiency. VaR back testing showed us that no models performed adequately and showed consistent red lights in the traffic light test. Highlighting a limitation in using HAR based forecasting.

Overall, HAR models offer valuable insist for volatility forecasting, but additional improvements are needed to ensure reliable risk assessment, particularly during times of market turbulence, which indeed is difficult.