Exploring the Effectiveness of Different Volatility Forecasting Models in Varying Market Conditions

FM442- Quantitative Methods for Finance and Risk Analysis

Candidate Number: 39270

January 20, 2025

Abstract

The goal of this research is to investigate the performance of different volatility models in varying market conditions. Specifically, the periods analysed were 2017-2018 representing low volatility, and 2020-2021 representing high volatility. The data chosen was the S&P 500 daily price (ticker: ^QSPC) from 2005-2024. Backtesting methods were then employed to assess the model performance in the specific time periods. The analysis compares GARCH, tGARCH, skewed tGARCH, and EWMA models. Results indicate that tGARCH and skewed tGARCH consistently outperform others across both periods, accurately capturing market risks and extreme events. Conversely, the EWMA model fails to meet robustness criteria during high-volatility periods, highlighting its limitations.

Introduction

Volatility is a cornerstone of financial markets, playing a key role in pricing, portfolio management, and risk control. Understanding and predicting market volatility remains a critical challenge for practitioners and academics alike. Volatility modelling, therefore, is one of the most fundamental tools in financial risk management. In a rapidly changing industry, where there are numerous models and parameters to choose from, having a deep understanding of when and which models perform best is essential for efficient and accurate risk analysis.

This research contributes to the growing body of literature on volatility modelling. One notable example is a thesis titled "*Modeling of Market Volatility with APARCH Model*" (Ding 2011), which focused on comparing the APARCH volatility model with other stochastic volatility models. Another study, "*Evaluating Volatility Forecasts in Various Equity Market Regimes*" (Felletter 2017), analysed GARCH and leveraged-GARCH models using the Volatility Index (VIX), with performance measured by metrics such as Mean Square Error and Theil U1.

This research aims to build upon and bridge these studies. It shares similarities with Ding (2011) in analysing a variety of volatility models and with Felletter (2017) in evaluating performance under varying market conditions. However, it diverges in its approach by employing backtesting methods, such as Violation Ratios and the Bernoulli Coverage Test, to assess model performance. By doing so, this study offers a novel perspective on evaluating volatility models and their practical applications in risk management.

The objective of this research is to evaluate the performance of multiple volatility models across varying market conditions, using backtesting methods to assess their reliability and effectiveness. This approach aims to provide both academics and practitioners with valuable insights into the comparative strengths and weaknesses of these models under dynamic market environments.

Data

The chosen data were daily prices for the S&P 500 Index from 2005-01-01 to 2024-12-31 and has been downloaded from WRDS. This data was chosen for its ability to represent broad market dynamics, its high liquidity, and its relevance in financial risk analysis. As shown in the next section this data is highly suitable for the purposes of this research.

The data was loaded into R and the log returns (r_t) were calculated by,

$$r_t = \log(P_t) - \log(P_{t-1})$$

where P_t is the price at time t.

Summary Data Analysis

Next some simple data analysis was conducted. We start by generating simple plots showing the price trend and returns of the S&P 500.

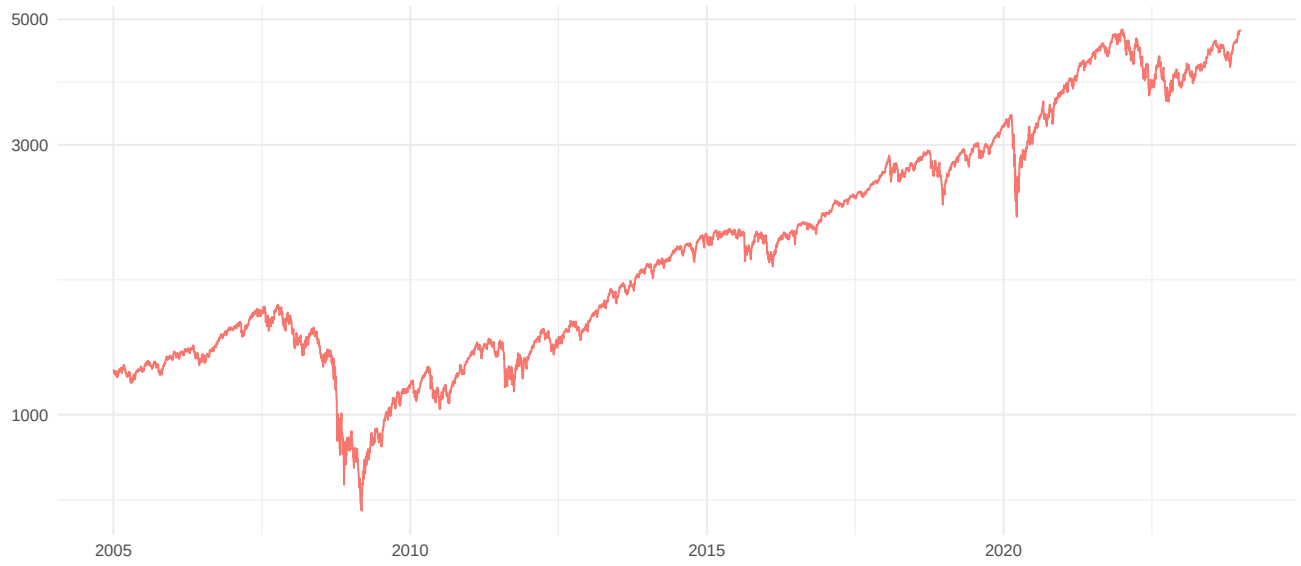

Figure 1: S&P-500 Index Price Trend (2005–2025)

This plot shows how the S&P 500 has changed over time, notably the financial crisis in 2008 and the COVID-19 financial crash is clearly visible. Following this we create a plot showing the returns.

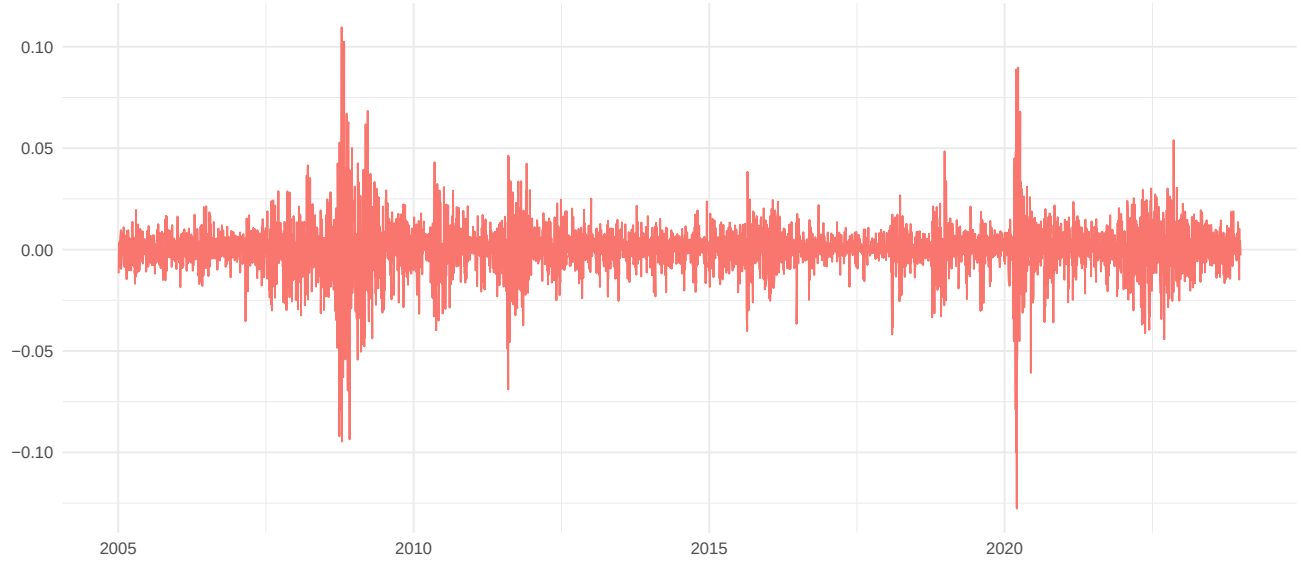

Figure 2: S&P-500 Index Log Returns

From these two plots we can see a period of relative stability from 2012-2019 and highly volatile conditions from COVID-19 and the 2008 Crisis. Ultimately for this research the periods chosen cover the years leading up to COVID-19, and the years encompassing the subsequent financial fallout.

Next we calculate certain important statistical results and conduct statistical tests to verify stationarity, volatility clustering, and autocorrelation of returns.

Table 1: Summary Statistics for Returns

	Result
Mean	0.000288
Unconditional Volatility	0.012283
Skewness	-0.522960
Kurtosis	15.858176
JB Test Statistic	33146.637516
ADF Test Statistic (Returns)	-16.763053
ADF P-Value (Returns)	0.010000
ADF Test Statistic (Squared Returns)	-9.593922
ADF P-Value (Squared Returns)	0.010000
LB Test Statistic (Squared Returns)	5027.831017
LB P-Value (Squared Returns)	0.000000

The mean return is close to zero which is expected for financial returns, similarly the volatility is consistent with expected volatility for daily returns. The negative skewness indicates the distribution of returns has a longer left tail, this means larger negative returns occur more frequently than large positive returns. This is common with financial returns, reflecting the greater downside risk. Kurtosis is typically compared to the normal distribution, which has kurtosis of three, and a value of 15.9 indicates fatter tails - i.e. more extreme returns than would occur if the returns were normally distributed.

Following this, different tests were conducted. First the Jarque-Bera normality test was done and a result of 33000 means the returns are not normally distributed. Secondly, ADF (Augmented Dickey-Fuller) tests were conducted to analyse stationarity, and the results indicate stationarity is satisfied. Finally, the Ljung-Box Test was conducted on the squared returns to detect volatility clustering and autocorrelation. The large test statistic and p-value close to zero strongly indicate volatility clustering is present.

The conclusion drawn from this analysis is that the chosen data is appropriate for the research in question, so next we move on to creating the volatility forecasting models.

Empirical Analysis

For the entirety of this analysis RStudio was used and numerous libraries such as `quantmod`, `tseries`, and most importantly, `rugarch` were used for the statistical analysis. Other libraries such as `ggplot` and `knitr` were used for formatting and presentation purposes.

The plan for analysis was as follows:

1) Create the forecasting models using the whole dataset.
2) Using the forecasting models to forecast data within the chosen time periods.
3) Using the forecasted volatility to create forecasts for Value-at-Risk (VaR).
4) Finally, conducting backtesting methods: Violation Ratios and Bernoulli Coverage Test to assess individual model performance.

Volatility Models

In this section the volatility models were generated. The chosen models to investigate were GARCH, tGARCH, Skewed tGARCH, and EWMA. Choosing the correct parameters was done by considering varying (p, q) values and calculating the Akaike Information Criterion (AIC) (Akaike, 1974), and choosing the parameters from the model with the most negative AIC. Similarly, for the tGARCH and Skewed tGARCH which contain an additional degrees of freedom parameter, different values were analysed and chosen based on the AIC.

The AIC results for all models are presented in the appendix.

Standard GARCH

The GARCH model (Generalised Autoregressive Conditional Heteroskedasticity), introduced by Bollerslev (1986) as an extension of the ARCH model developed by Engle (1982), has become a widely used tool for modelling conditional volatility.

It is defined by:

$$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i \varepsilon_{t-j}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2$$

where σ_t is the conditional volatility, (ω, α, β) are the main model parameters, and (p, q) are the lags in the volatility model. ε_t represents the residuals, in the case of GARCH these are assumed normally distributed with mean zero.

A function to create GARCH models for all combinations of $p = (1,\ 2,\ 3)$ and $q = (1,\ 2,\ 3)$ was made and at every iteration the AIC was stored. This is simply done by extracting the AIC from the `ugarchfit` which is saved as `infocriteria(fit)[1]`. The actual formula for AIC is defined as:

$$AIC = 2k - 2\log(\hat{\mathcal{L}})$$

where k is the number of estimated parameters and $\hat{\mathcal{L}}$ is the maximum value of the likelihood function.

After choosing the best model the log likelihood score was calculated, this is used to provide a simple analysis of model performance over the whole dataset.

tGARCH

tGARCH differs from GARCH by having $\varepsilon_t \sim t_\nu(0, \sigma_t^2)$, where t_ν is the Student's t distribution with ν degrees of freedom. This is instead of the normal distribution.

Analysis was done similar to the previous section, but models were also tested with varying degrees of freedom.

Skewed t-GARCH

The skewed tGARCH means one side of the distribution is fatter than the other, this is particularly useful for financial returns. As previously stated the returns have negative skew, so it is logical to model using a skewed distribution.

This means $\varepsilon_t \sim t(\xi, \nu)$ where ξ is the skewness parameters. $\xi > 1$ indicates positive skewness and $\xi < 1$ indicates negative skewness.

Fundamentally the formula remains unchanged,

$$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i \varepsilon_{t-j}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2$$

Exponentially Weighted Moving Average (EWMA)

The final model chosen was EWMA, a model popularised by JP Morgan in their RiskMetrics framework (JP Morgan, 1996), is a widely used tool for estimating conditional volatility due to its simplicity and efficiency.

As the data is daily, $\lambda = 0.94$, was chosen.

This volatility model is given by:

$$\hat{\sigma}_t^2 = (1 - \lambda)\ r_{t-1}^2\ + \lambda \hat{\sigma}_{t-1}^2,$$

where r_{t-1} is the return at time $t-1$ and $\hat{\sigma}_t^2$ is the EWMA variance forecast at time t. This model is considerably easier to use than the previous GARCH models, the simpler calculation also allows for expanding the testing window over longer time frames without greatly increasing computational requirements.

Backtesting

Since the objective of this project is to analyse how effective each model is, we test the models by doing backtesting. This involves using the models to forecast results, then conducting analysis with violation ratios and the Bernoulli Coverage Test by Kupiec (1995).

Rolling Window Analysis

We start by conducting a rolling window analysis over the data ranges (2016-01-01, 2017-12-31) and (2019-01-01, 2020-12-31) to represent the periods of low and high volatility, respectively. The conditional volatility of daily returns for these two periods were $\sigma_t^{(2017)} = 0.421\%$ and $\sigma_t^{(2020)} = 2.177\%$ compared to the unconditional volatility of $\sigma = 1.228\%$. This uses the models we previously made, then forecasting one day ahead from the start of the specified period and repeated over the whole testing window.

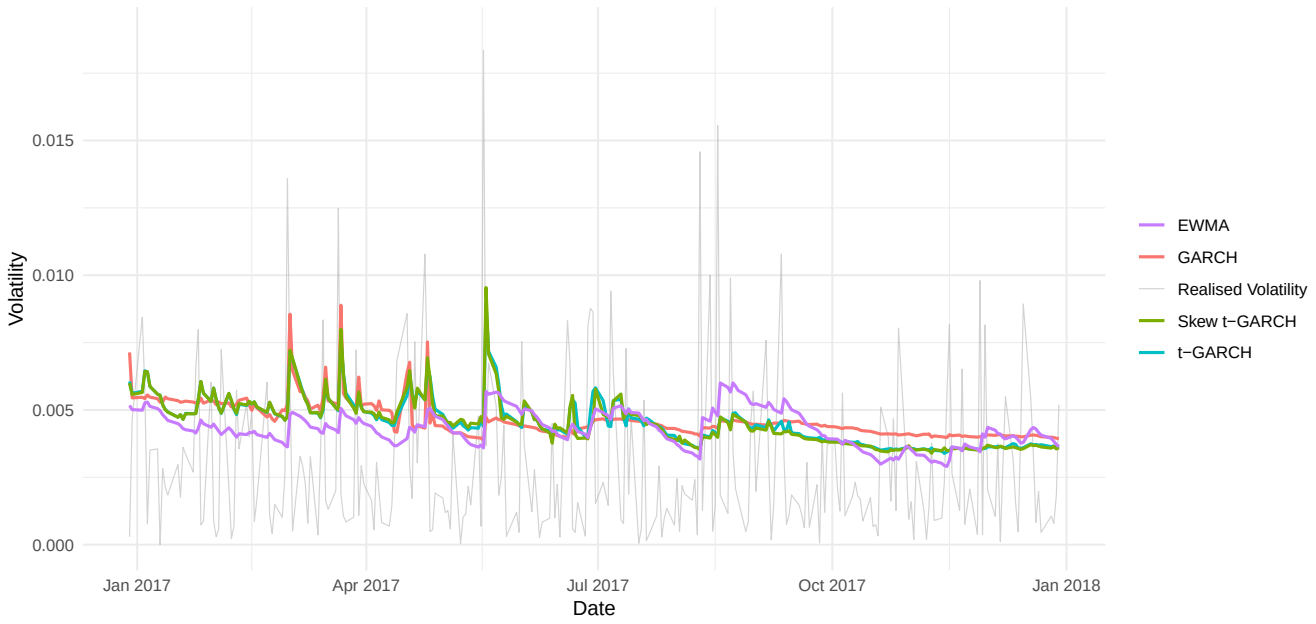

Figure 3: Rolling Window Volatility Forecasting (2017-2018)

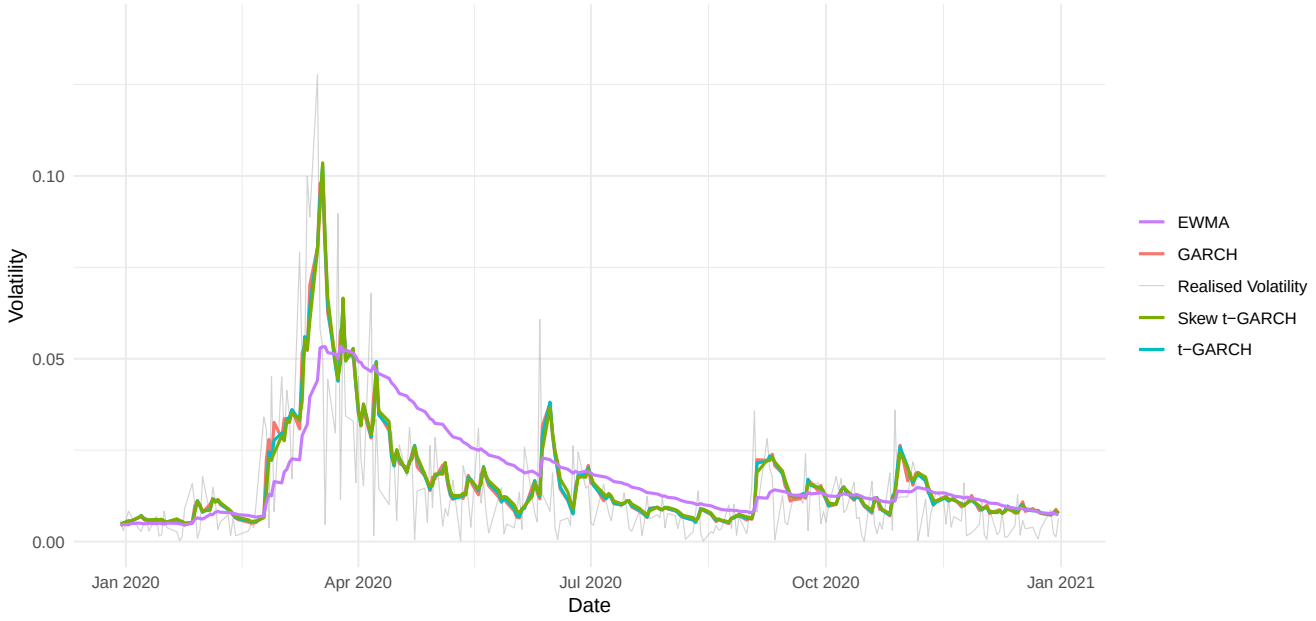

Figure 4: Rolling Window Volatility Forecasting (2020-2021)

From these results we can see that the different GARCH models appear similar, but EWMA deviates greatly.

Violation Ratios

In this section we use the volatility forecasts to also forecast a one-day $\text{VaR}_{0.99}$ and then compare this to the actual returns to obtain our observed number of violations. The expected number of violations are calculated by Testing Window Length * 0.01. Finally, we simply we find violation ratio as such,

$$\text{Violation Ratio} = \frac{\text{Observed Number of Violations}}{\text{Expected Number of Violations}}$$

After calculating the violation ratio for all four models and the two periods the results are presented in a plot with a threshold indicating where models would ideally lie.

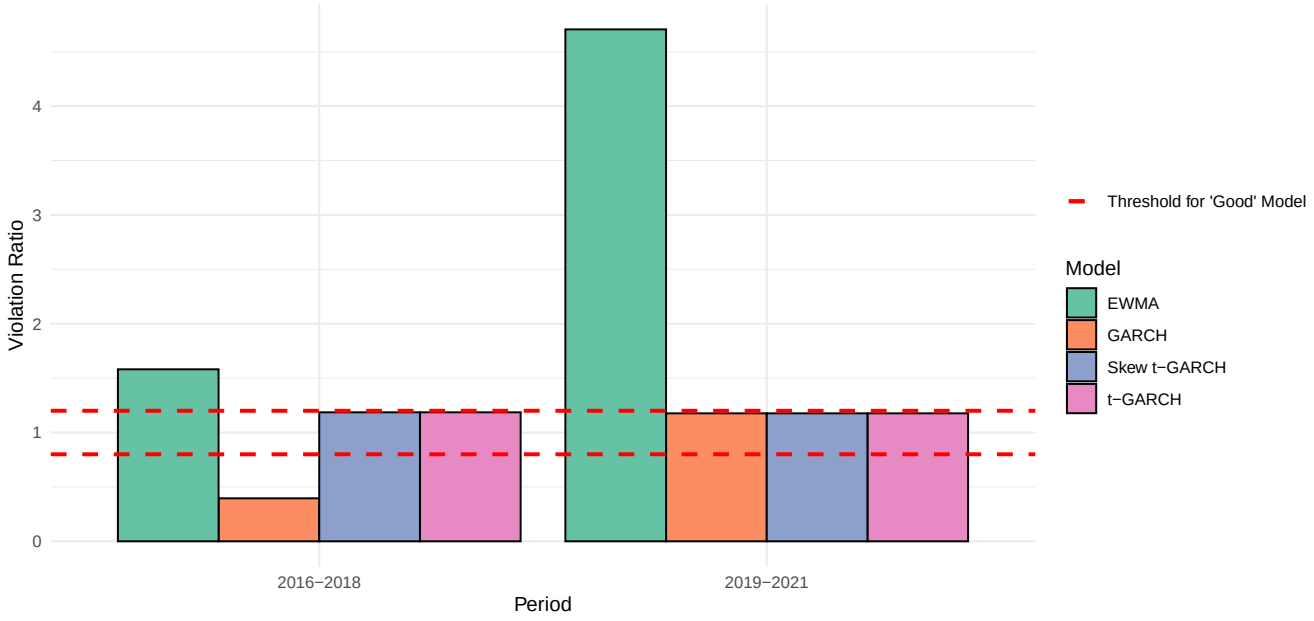

Figure 5: Violation Ratio Plot with Threshold

A good model would typically a violation ratio between 0.8 and 1.2. However, these results are not conclusive, and further testing was done to analyse the statistical significance.

Coverage Test

In this section we are particularly interested in analysing whether the observed violation matches with the expected violation, or if there is significant statistical difference. We do this by conducting a Bernoulli Coverage test with a significance level of 5%. The resulting test statistic is then compared against the critical value of $\chi^2_{(1)}(5\%) = 3.841$.

The formula is given as,

$$\text{Bernoulli Test Statistic} = -2\left(\left[\log(p)V + \log(1-p)(n-V)\right] - \left[n\log\left(\frac{V}{n}\right) + (n-V)\log\left(\frac{1-V}{n}\right)\right]\right)$$

where, $V = \sum_{i=1}^{n} V_i$ is the number of violations, n is the length of the data (the testing window), and p is the VaR % that was used - in the case of this research, $p = 0.01$.

We visualise these results similarly to the violation ratios by plotting the test statistic and marking the critical value.

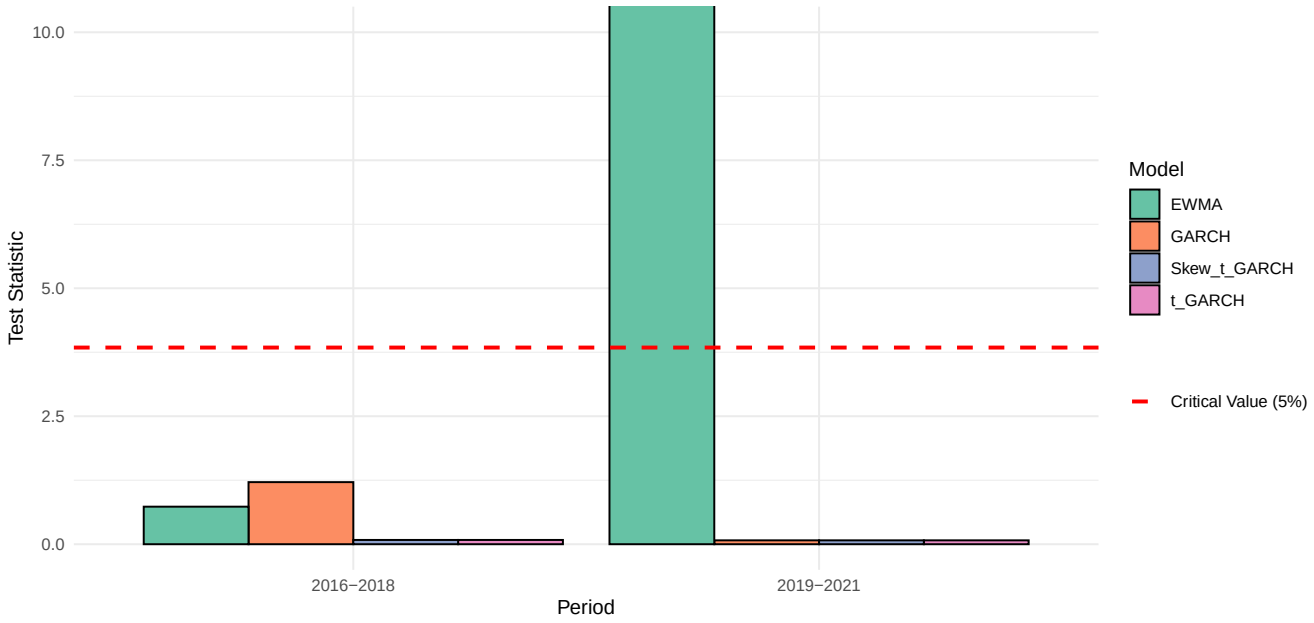

Figure 6: Bernoulli Coverage Test

Clearly visible is that in the period of low volatility all models are below the critical value, however during high volatility EWMA fails. This is consistent with the Violation Ratio test, except during low volatility when EWMA and GARCH did not meet the criteria of a good model.

Results and Analysis

The results for the log likelihood of the four models are presented below, this assesses how well the model fits the full data range. This is particularly important when choosing which model, but for the purposes of this research we are more interested in the performance over a specific data range.

Table 2: LogLikelihood For Different Models

Model	LogLikelihood
GARCH	15622.43
tGARCH	15755.44
Skew tGARCH	15770.96
EWMA	15491.73

From this we can see that the skewed tGARCH provided best fit the data. The results from the backtesting are now presented,

Table 3: Violation Ratio Results

Period	Model	Violation Ratio
2016-2018	GARCH	0.3952569
2016-2018	t-GARCH	1.1857708
2016-2018	Skew t-GARCH	1.1857708
2016-2018	EWMA	1.5810277
2019-2021	GARCH	1.1764706
2019-2021	t-GARCH	1.1764706
2019-2021	Skew t-GARCH	1.1764706
2019-2021	EWMA	4.7058824

Table 4: Coverage Test Results

Model	Period	Test Statistic	Exceeds Critical
GARCH	2016-2018	1.2128885	FALSE
GARCH	2019-2021	0.0759162	FALSE
t_GARCH	2016-2018	0.0832404	FALSE
t_GARCH	2019-2021	0.0759162	FALSE
Skew_t_GARCH	2016-2018	0.0832404	FALSE
Skew_t_GARCH	2019-2021	0.0759162	FALSE
EWMA	2016-2018	0.7332448	FALSE
EWMA	2019-2021	18.6297607	TRUE

As stated previously we can assess a model to be good if the violation ratio lies between approximately 0.8 and 1.2, we can clearly see that EWMA fails under both market conditions and when further investigating with the coverage tests EWMA once again fails during the period of high volatility.

The GARCH models nearly all fit well and do not violate the coverage test critical value, however GARCH appears to greatly underestimate the expected losses in periods of low volatility.

The results of this testing leans towards the tGARCH and skewed tGARCH models being the most optimal in both high and low periods of market volatility.

The results from the evaluation of the volatility models offer valuable insights into their relative performance under varying market conditions. By examining the violation ratios and conducting the Bernoulli Coverage Test, the performance of GARCH, t-GARCH, Skew-t GARCH, and EWMA models can be contextualised within the broader scope of financial risk management.

The violation ratios indicate how well each model captures extreme events, as expected by the 1% Value-at-Risk (VaR). For instance:

- The GARCH model demonstrated reasonably consistent performance despite its simplicity. However, in times of low market volatility it failed to accurately represent financial losses.
- The tGARCH model, which incorporates heavier tails from the Student's t-distribution improved upon GARCH in all metrics and performed optimally in all market conditions.
- The skewed tGARCH model, which differs from tGARCH by incorporating skewness, a feature present in not only the returns data for this analysis, but more generally in returns data, outperformed tGARCH. However, to further see the difference it would be valuable to conduct more research, within different markets and over varying estimation windows.

- The EWMA model consistently misrepresented the realised market risks and was not able to accurately capture the clustering or distributional asymmetry - likely due to the exponential smoothing. Seemingly, the only benefit to this model would be the computational efficiency and simplicity.

These results do align with what we would expect to see in theory.

Compared to studies such as Ding (2011), which emphasised APARCH models, and Felletter (2017), which used alternative performance metrics, this research provides a complementary perspective by focusing on backtesting methods. The use of violation ratios and Bernoulli Coverage Test offers a practical and robust approach to evaluating risk models, bridging gaps in existing research.

Conclusion

The results of this analysis show that GARCH and EWMA were the poorest performing. They had the lowest likelihoods scores for the whole data range and for the market condition specific analysis the EWMA model failed the Violation Ratio and Coverage Test. Similarly, the GARCH model performed poorly during the period of low volatility, with a violation ratio of 0.4. Both tGARCH and the skewed tGARCH models performed well, meeting the requirements of the backtesting analysis to be considered a good model.

There are limitations of this research. Originally the plan also considered the APARCH(p,q) model, defined as:

$$\sigma_t^\delta = \omega + \sum_{i=1}^{p} \alpha_i (|\varepsilon_{t-i}| - \gamma_i \varepsilon_{t-i})^\delta + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^\delta$$

(Bollerslev, 2008; Ding, Granger, & Engle, 1991)

The GARCH model can be obtained from this by setting $\delta = 2$ and $\gamma = 0$.

Initial analysis gave a likelihood score greater than all the other models, which suggests it to have fit the whole data range the best. However, the backtesting of this model required far greater computational power. Libraries such as `parallel` were used to run the backtesting on multiple CPU cores, however this still would have taken over 10 minutes on a high-power computer. For reproducibility this was removed, and more focus was placed on the other GARCH and EWMA models.

Similarly, the issue of computing power meant a more limited testing window was used during the backtesting analysis. This could potentially be addressed using machine learning/neural network based backtesting methods which may be more efficient for GARCH backtesting. Further analysis on this topic with varying estimation windows, both shorter and longer, could provide valuable insight into how the models perform in varying market conditions.

Expanding the code used in this research to accommodate for a more thorough analysis is relatively easily done, however as stated this would require far more computational power.

Finally, the applications of this research to the context of risk management would be to use the skewed tGARCH or tGARCH for volatility forecasting. These models were shown to be highly effective and produce reliable results which allow for forecasting potential losses in financial portfolios.

References

[1] Bollerslev, T. (2008). Glossary to ARCH (GARCH). *Handbook of Financial Time Series*, 1–16.

[2] Ding, D. (2011). Modeling of Market Volatility with APARCH Model,

https://www.diva-portal.org/smash/get/diva2:417608/FULLTEXT01.pdf

[3] Ding, Z., Granger, C. W. J., & Engle, R. F. (1991). A Long Memory Property of Stock Market Returns and a New Model. *Journal of Empirical Finance, 1*, 83–106.

[4] Felletter, J. P. (2017) Evaluating Volatility Forecasts in Various Equity Market Regimes,

https://digitalcommons.sacredheart.edu/cgi/viewcontent.cgi?article=1008&context=wcob_theses

[All links accessed: 2025-01-19]

Appendix

Table 5: GARCH

p	q	AIC
1	1	-6.5326
2	1	-6.5339
3	1	-6.5334
1	2	-6.5322
2	2	-6.5341
3	2	-6.5330
1	3	-6.5316
2	3	-6.5336
3	3	-6.5334

Table 6: tGARCH

p	q	df	AIC
1	1	3.0	-6.5571
2	1	3.0	-6.5608
3	1	3.0	-6.5604
1	2	3.0	-6.5566
2	2	3.0	-6.5604
3	2	3.0	-6.5602
1	3	3.0	-6.5562
2	3	3.0	-6.5601
3	3	3.0	-6.5598
1	1	3.5	-6.5743
2	1	3.5	-6.5776
3	1	3.5	-6.5772
1	2	3.5	-6.5739
2	2	3.5	-6.5773
3	2	3.5	-6.5769
1	3	3.5	-6.5735
2	3	3.5	-6.5769
3	3	3.5	-6.5766
1	1	4.0	-6.5823
1	2	4.5	-6.5856
2	2	4.5	-6.5883
3	2	4.5	-6.5879
1	3	4.5	-6.5852
2	3	4.5	-6.5880
3	3	4.5	-6.5876
1	1	5.0	-6.5877
2	1	5.0	-6.5901
3	1	5.0	-6.5897
1	2	5.0	-6.5873
2	2	5.0	-6.5898
3	2	5.0	-6.5893
1	3	5.0	-6.5868
2	3	5.0	-6.5894
3	3	5.0	-6.5890

Table 7: Skew tGARCH

p	q	df	AIC
1	1	3.0	-6.5598
2	1	3.0	-6.5636
3	1	3.0	-6.5632
1	2	3.0	-6.5594
2	2	3.0	-6.5632
3	2	3.0	-6.5629
1	3	3.0	-6.5589
2	3	3.0	-6.5628
3	3	3.0	-6.5626
1	1	3.5	-6.5781
2	1	3.5	-6.5814
3	1	3.5	-6.5810
1	2	3.5	-6.5777
2	2	3.5	-6.5810
3	2	3.5	-6.5807
1	3	3.5	-6.5772
2	3	3.5	-6.5806
3	3	3.5	-6.5804
1	1	4.0	-6.5869
1	2	4.5	-6.5910
2	2	4.5	-6.5937
3	2	4.5	-6.5933
1	3	4.5	-6.5905
2	3	4.5	-6.5934
3	3	4.5	-6.5930
1	1	5.0	-6.5937
2	1	5.0	-6.5962
3	1	5.0	-6.5958
1	2	5.0	-6.5933
2	2	5.0	-6.5958
3	2	5.0	-6.5954
1	3	5.0	-6.5928
2	3	5.0	-6.5954
3	3	5.0	-6.5951