Final Report Group 9

Statistics and Financial Data Analysis
Master in Computational Finance, School of Computing

Anđela Bašić, Luka Basta, Dušan Ćukalović, Ratko Nikolić

26. Jun 2024.

As per final project requirements, we conducted an analysis in four phases to analyse Invesco QQQ Trust (QQQ) ETF returns. First phase begin with ideas of which assets and parameters to include in our reserch and ended with collected and prepared data. The goal of the second phase was to build a univariate time series model for forecasting. We chose the best performing ARMA model - ARMA(1,1) without an intercept - on test data according to MAE and RMSE metrics, among best performing in-sample models. In the third phase we built a multivariate regression model for forecast and compared it against to benchmarks - Fama-French model without intercept and with/witoud January dummy. Regression model significantly outperformed ARMA model. Finally, we built a volatility model after detecting the ARCH effect. Again, we compared best performing in-sample GARCH + ARMA(1,1) models. Our results showed that they have equal predictive power.

Contents

1 Introduction

For the purposes of this project, we chose to analyze the daily price of the Invesco QQQ Trust (QQQ) ETF. The Invesco QQQ Trust ETF (QQQ) is a leading exchange-traded fund issued by Invesco, launched on March 10, 1999. With an expense ratio of 0.20% and assets exceeding $200 billion, QQQ tracks and aims to replicate the performance of the NASDAQ-100 Index, focusing on technology, telecommunications, and biotechnology sectors. It includes top tech companies like Apple, Microsoft, Amazon, Alphabet, Meta Platforms, and NVIDIA. QQQ has shown strong long-term growth, with an average annual return of around 20% over the past decade. Despite higher volatility typical of tech stocks, its performance has consistently outpaced many other indices.

Some of the reasons behind our choice of QQQ include:

- Broad Tech Exposure: Comprehensive coverage of leading tech companies.
- Data Availability: Extensive historical data and high trading volume.
- Market Trends Representation: Reflects performance of major non-financial tech companies.
- Liquidity: High liquidity ensures reliable study conditions.
- Volatility Analysis: Suitable for modeling volatility, a key project component.

2 Data collection and preparation

The gathered data spans 3 years from May 1, 2021, to April 30, 2024, and includes daily observations of the dependent (QQQ adjusted closing prices) and the potential explanatory variables like the adjusted closing prices for the S&P 500 Index (SP500), the NASDAQ Composite Index (NASDAQ), as well as the Volatility Index (VIX). All of the aforementioned were gathered from Yahoo Finance using R package quantmod and provide a comprehensive view of market performance and volatility, critical for analyzing QQQ's behavior.

Additionally, we sourced the Fama-French five-factor model and the Momentum factor from Kenneth French's data library.
The factors include Market Risk Premium (MKT), Small Minus Big (SMB), High Minus Low (HML), Robust Minus Weak (RMW), Conservative Minus Aggressive (CMA), and Momentum (MOM). These factors are well-regarded in financial literature for their ability to explain asset pricing and risk, as established by Fama and French [2, 3] and Carhart [1]. Integrating these factors provides a robust analytical framework to understand and forecast QQQ's performance and volatility.

In order to assess the stationarity of our dependent variable (QQQ Adjusted Closing Price on 1), we plotted its time series, as well as performed the Augmented Dickey Fuller Unit Root Test which showed that the dependent variable time series was not stationary. To address this finding, we calculated the log returns of the dependent variable (QQQ Log Returns on 1) which passed the

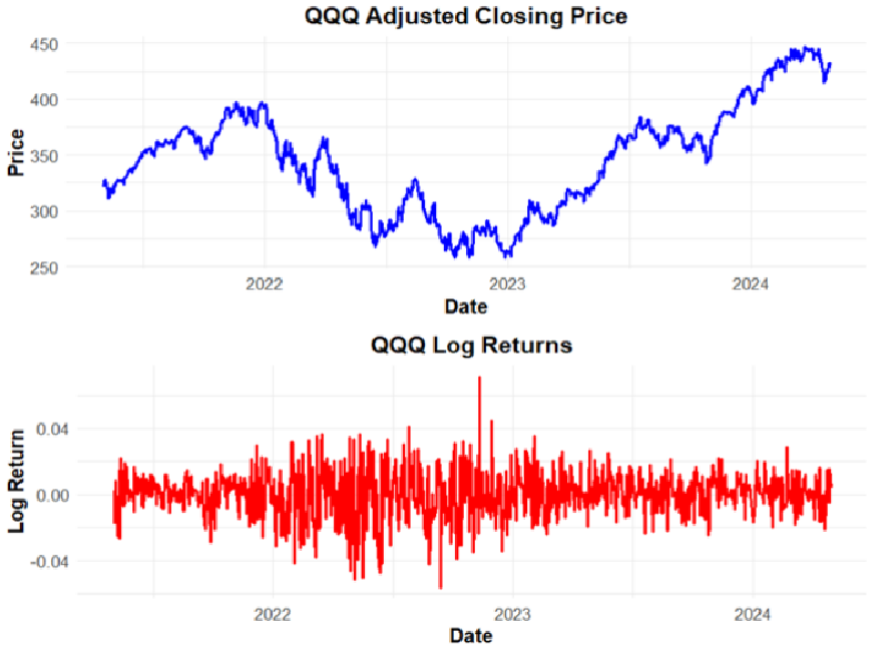


Figure 1: QQQ Adjusted Closing Price and its calculated log returns (QQQ Log Returns) between May 1st 2021 and April 30th 2024

stationarity test. Therefore, we decided to use it as our dependent variable throughout the rest of the project.

With the omission of the first row which was necessary because of the calculation of the log returns, we were left with a dataset containing 752 observations of the dependent and all of the potential explanatory variables. This dataset was then split into a training and testing set (80:20) containing 601 and 151 observations respectively. All of the steps taken in data collection and preparation phase are available in the corresponding section of the R script **(SFD_Final_Project_Group_9.R)**, while all of the data gathered is available in a CSV file **(sfd_final_project_data.csv)** that accompanies this report.

3 Building a univariate time series model for forecast

With the train/test split and the stationarity of the dependent variable evaluation dealt with in the previous phase, we went on to test autocorrelation. As it can be seen on the 2 and 3, ACF and PACF values are close to zero at most tested lags (0 to 27), with them crossing the threshold for being outside of the 95% confidence interval (which is approximately ±0.08 for a dataset with 601 observations [1]) only at lag 9 (0.089) in the context of ACF, as well as lag 9 (0.09), lag 11 (0.086) and lag 22 (-0.087) in the context of PACF. These findings gave us clues to which lags might prove particularly useful in AR, MA and ARMA models and helped us define the parameter ranges for q and r we would test.

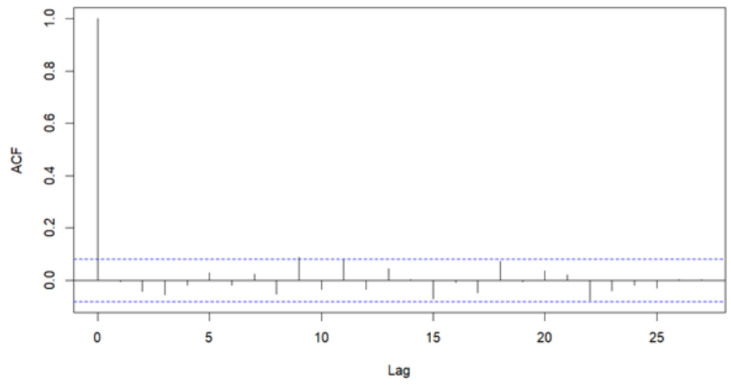


Figure 2: Autocorrelation analysis: ACF of QQQ Log Returns

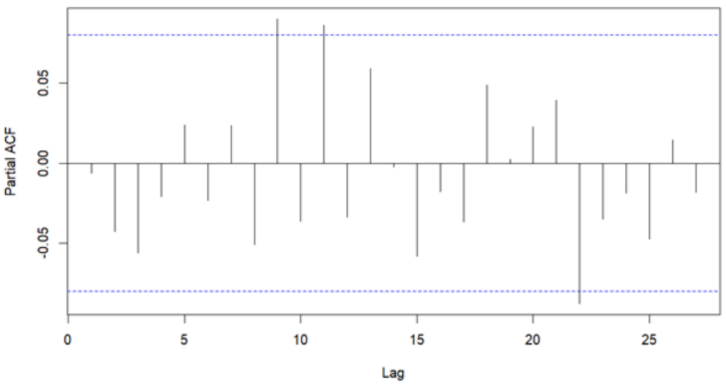


Figure 3: Autocorrelation analysis: PACF of QQQ Log Returns

[1] $\pm 1.96 \,/\, \sqrt{601} \approx \pm 0.08$

We proceeded with fitting AR(p) ($1 \leq p \leq 10$), MA(q) ($1 \leq q \leq 10$), and ARMA(p,q) ($1 \leq p \leq 10$, $1 \leq p \leq 10$) and then selecting the top models from each category ranked by AIC and BIC respectively. As it can be seen in the Table 1, within AR models, AR(1) was preferred according to both AIC (-3292.194) and BIC (-3278.998), indicating that a single lag adequately captured the data's dependencies, while more complex models didn't seem to improve the metrics. Similarly, MA(1) was the top MA model by both AIC (-3292.196) and BIC (-3279), suggesting one lag is sufficient for moving average terms. Within ARMA models, ARMA(8,6) had the lowest AIC (-3301.846), implying a complex model with five autoregressive and eight moving average terms fits best according to AIC. However, ARMA(1,1) was the top model by BIC (-3273.856), favoring a simpler model with one autoregressive and one moving average term due to stronger penalty on complexity. We went on to perform model evaluation on the top performers from each category: AR(1), MA(1), ARMA(1,1), ARMA(5,8) and ARMA(8,6) (as ARMA(5,8) and ARMA(8,6) seemed close in terms of AIC).

AR		MA		ARMA	
AIC	BIC	AIC	BIC	AIC	BIC
AR(1) −3292.194	AR(1) −3278.998	MA(1) -3292.196	MA(1) -3279.00	ARMA(8,6) -3301.846	ARMA(1,1) -3273.856
AR(2) −3291.296	AR(1) −3273.701	MA(2) -3291.396	MA(2) -3273.801	ARMA(5,8) -3301.631	ARMA(2,1) -3268.680

Table 1: Top AR, MA and ARMA models ranked by AIC and BIC respectively

After fitting the selected models, we realised that for AR(1) and MA(1) none of the coefficients are significant. Furthermore, for ARMA(5,8) insignificant coefficients were intercept, ma3, ma6, ma7 and ar3, while for ARMA(8,6) insignificant coefficients were intercept, ma2, ma3, ma6 and ar2, ar3, ar6. For ARMA(1,1) intercept was insignificant. We decided to proceed with the simplest models with most significant coefficients - **ARMA(1,1) with intercept, ARMA(1,1) without intercept and ARMA(5,8)** We performed overfitting check by estimating slightly larger models - ARMA(2,1), ARMA(1,2), ARMA(5,9), ARMA(6,8) and none of the added coefficient appeared significant.

To study if there is autocorrelation in residuals, we first performed visual check by plotting the respective ACF functions - Figure 4 . For ARMA(1,1) we noticed small visual persistence at lag 9, while for ARMA(5,8) no visual persistence was detected. To confirm our observations, we performed Ljung-Box test for each model up to lag 16 fir parameter fitdf set to be 2 for ARMA(1,1) and 13 for AMRA(5,8). In both cases null hypothesis of no serial correlation was accepted.

We performed rolling forecast to obtain 1-step-ahead and 3-step-ahead forecasts for selected models. Afterwards we calculated performance indicators (mean absolute error MAE and root mean square error RMSE) given in Table 2. Results suggest that ARMA(1,1) and ARMA(1,1)* (without an intercept) outperform ARMA(5,6), and that ARMA(1,1)* outperforms ARMA(1,1). To confirm this we performed Diebold-Mariano Test for 1 and 3 periods ahead. Tests showed that ARMA(1,1) and ARMA(1,1)* have equal predictive power, while ARMA(5,8) confirmed to be inferior to ARMA(1,1) (and ARMA(1,1)*).

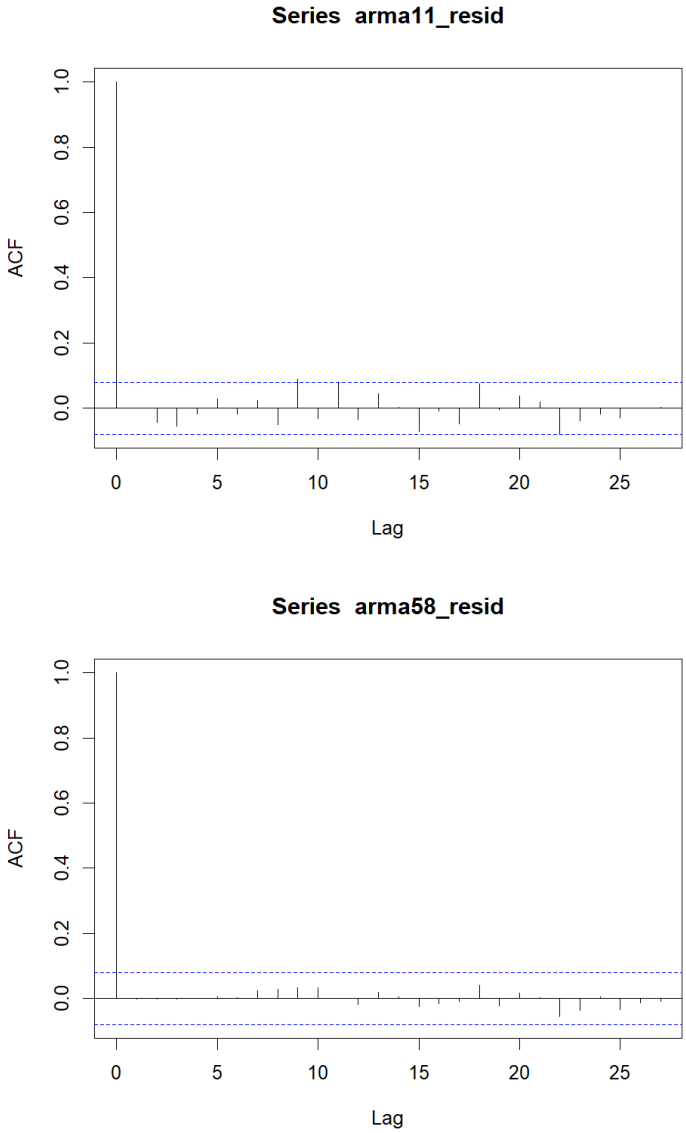


Figure 4: ACF plot for ARMA(1,1) and ARMA(5,8) residuals

Model	MAE 1	RMSE 1	MAE 3	RMSE 3
ARMA(1,1)	0.008126759	0.010300937	0.00802219	0.010154674
ARMA(1,1)*	0.008557739	0.01094061	0.008251175	0.010392905
ARMA(5,8)	0.008019006	0.010246074	0.007883778	0.01004704

Table 2: MAE and RMSE for 1 and 3 periods ahead forecast for ARMA(1,1), ARMA(1,1) without intercept and ARMA(5,8)

We conclude this section by presenting forecasts against actual values for one-period-ahead forecasts in Figure 12 a and three-periods-ahead forecasts in Figure 13.

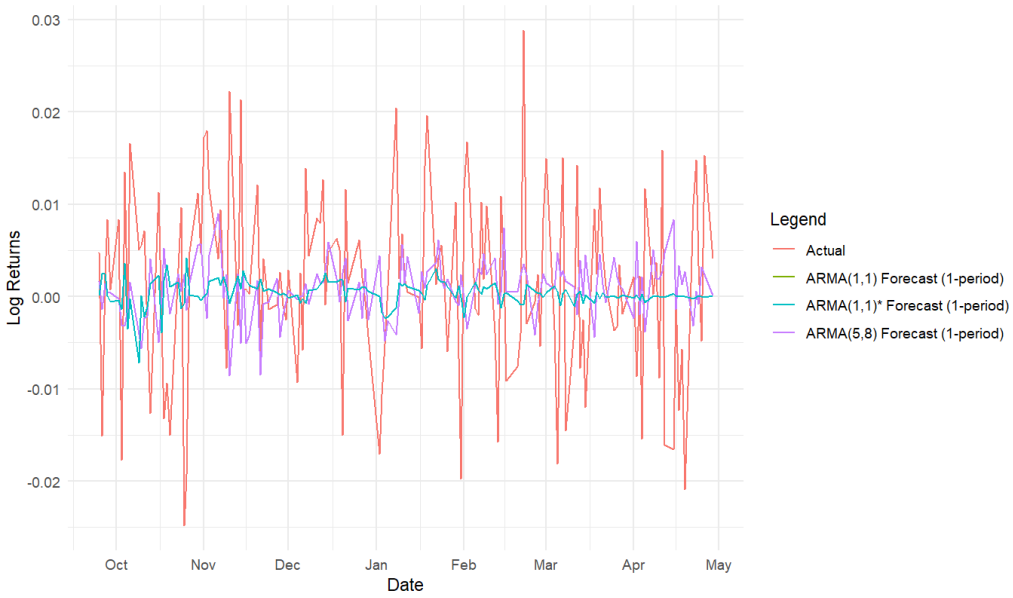


Figure 5: One-period-ahead forecasts for AMRA models

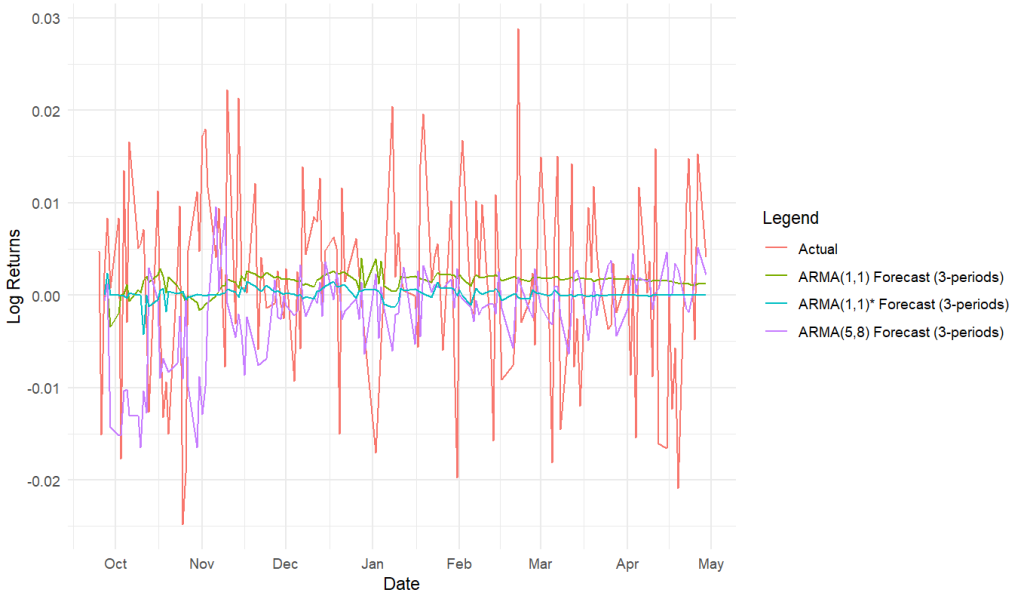


Figure 6: Three-period-ahead forecasts for ARMA models

4 Building a multivariate model for forecast

As mentioned in Section 1, the potential explanatory variables we considered for the multivariate forecast model of QQQ log returns included the adjusted closing prices for the S&P 500 Index (SP500), the NASDAQ Composite Index (NASDAQ), and the Volatility Index (VIX) aiming to provide a comprehensive view of market performance and volatility. We further expanded the list of potential explanatory variables by including the variables from the Fama-French five-factor model, as well as the Momentum factor from Kenneth French's data library. The factors include Market Risk Premium (MKT), Small Minus Big (SMB), High Minus Low (HML), Robust Minus Weak (RMW) and Conservative Minus Aggressive (CMA).

The explanatory variables plots shown in Figure 7 indicated that all of the explanatory variables apart from SP500, NASDAQ and VIX were stationary. To confirm our assumption, we performed the Augmented Dickey Fuller Unit Root Test on each regressor and selected appropriate alternative (drift for SP500, NASDAQ and VIX). As our assumption was confirmed, we calculated the log returns for the non-stationary variables (SP500, NASDAQ and VIX) and repeated the ADF test on them, which showed that they were indeed stationary.

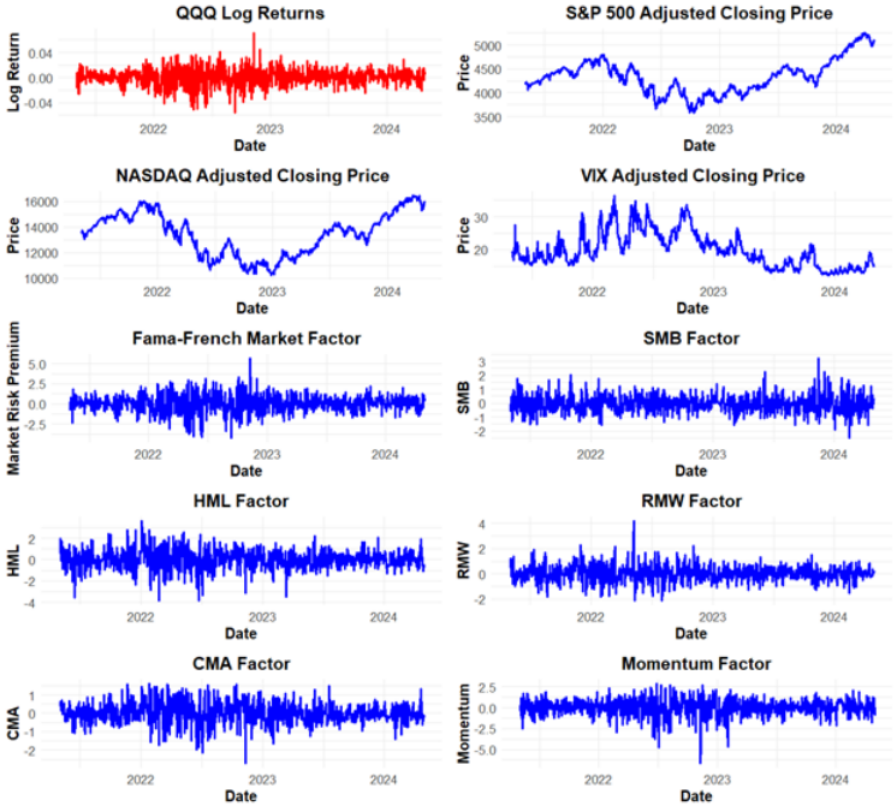


Figure 7: Plot of QQQ Log Returns and potential explanatory variables time series

We first fit a multivariate model containing all of the gathered variables that passed the stationarity test including those that were transformed to be able to do so. In order check for multicollinearity, we performed the Variance Inflation Factor test (VIF) which singled out MKTRF (395.42), SP500_log_returns (314.68) and NASDAQ_log_returns (43.19) as variables with VIF over 10 indicating presence of multicollinearity. We proceed to exclude MKTRF and refit the model which reduced multicollinearity overall and made NASDAQ_log_returns (36.57) the variable with highest VIF. After its exclusion the model had no variables with VIF over 10.

We proceeded to check for significance of the variable coefficients which showed that the intercept, VIX_log_return and RMW were not significant while MOM was borderline non-significant. We excluded all of them from the model and re-estimated parameters ending up with the model presented in Figure 8. In the final model all of the variable coefficients were significant, the adjusted R squared was 0.918 and the F-statistic was significant indicating that the model as a whole was significant as well

```
Call:
lm(formula = QQQ_log_return ~ SP500_log_return + SMB + HML +
    CMA - 1, data = train_data)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0110623 -0.0016055  0.0001875  0.0021358  0.0102472

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
SP500_log_return  1.1243789  0.0119952  93.736  < 2e-16 ***
SMB              -0.0006451  0.0002037  -3.166  0.00162 **
HML              -0.0029390  0.0001944 -15.118  < 2e-16 ***
CMA              -0.0019944  0.0003241  -6.153 1.39e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.003041 on 596 degrees of freedom
Multiple R-squared:  0.9621,    Adjusted R-squared:  0.9618
F-statistic:  3779 on 4 and 596 DF,  p-value: < 2.2e-16
```

Figure 8: Final regression model summary

Afterwards, we checked if the removed variables would become significant if returned to the model separately which proved not to be the case. Finally, we performed the stepwise regression to see how removal of the present variables would influence the model.

We reached the following conclusions:

- Model without SP500_log_return: RSE: 0.01205, R-squared: 0.4028, Adjusted R-squared: 0.3998 - Excluding SP500_log_return drastically reduces the model's explanatory power, highlighting its importance in explaining QQQ log returns.
- Model without SMB: RSE: 0.003064, R-squared: 0.9614, Adjusted R-squared: 0.9612 - Excluding SMB has a minimal impact on the model's

explanatory power, suggesting SMB contributes less compared to other variables.

- RSE: 0.003573, R-squared: 0.9475, Adjusted R-squared: 0.9473 - Excluding HML reduces the model's explanatory power, indicating HML is a significant factor in explaining QQQ log returns.
- Model without CMA: RSE: 0.003133, R-squared: 0.9597, Adjusted R-squared: 0.9595 - Excluding CMA slightly reduces the model's explanatory power, suggesting CMA is a significant but less influential factor compared to SP500_log_return and HML.

We also interpreted the models coefficients:

- SP500_log_return: A 1% increase in SP500 log returns is associated with an estimated 1.124% increase in QQQ log returns, holding other factors constant. This relationship is highly significant (p-value $<$ 2e-16).
- SMB: The coefficient is negative (-0.0006451) and significant (p-value $=$ 0.00162), suggesting that an increase in SMB (small minus big) is associated with a decrease in QQQ log returns.
- HML: The coefficient is negative (-0.0029390) and highly significant (p-value $<$ 2e-16), indicating that an increase in HML (high minus low) is associated with a decrease in QQQ log returns.
- CMA: The coefficient is negative (-0.0019944) and significant (p-value $=$ 1.39e-09), suggesting that an increase in CMA (conservative minus aggressive) is associated with a decrease in QQQ log returns.

Given that removal of any of the variables did not improve the model, we concluded that it was unnecessary and decided that the **model with SP500_log_return, SMB, HML and CMA** was the best multivariate model we could fit given the variables that we had. We refer to this model as MV3 going forward.

To further examine the MV3 model, we first tested residuals for normality using the Q-Q plot in Figure 9 which showed that the residuals were mainly normal with a slightly heavier left tail. Afterwards we performed formal tests for homoscedasticity and autocorrelation. To test for homoscedasticity we performed Breusch-Pagan test and White's test. In both cases we accepted the null hypothesis of no homoscedasticity was accepted, indicating that the variance of the errors from a regression is not dependent on the values of the independent variables. We also visualized these results as in Figure 10 via the residuals vs. fitted Q-Q plot. Finally, to test for autocorrelation we first performad visual check via ACF function of the residuals as in figure 11, which showed no signs of autocorrelation. We formally tested this claim by performing Breusch-Godfrey (LM) test for lags up to 4 and Ljung-Box test for lags up to 16. In both cases the null of no serial correlation was accepted.

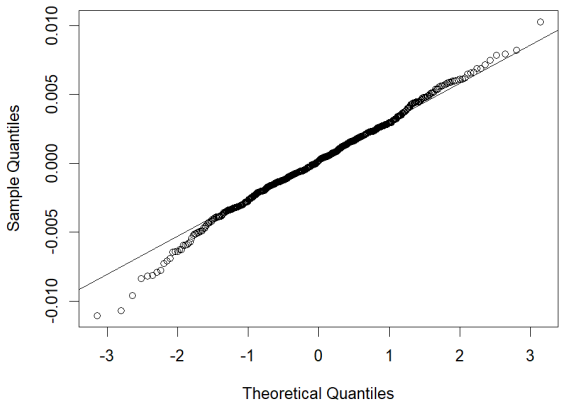


Figure 9: Normal Q-Q plot of the residuals of the MV3 model

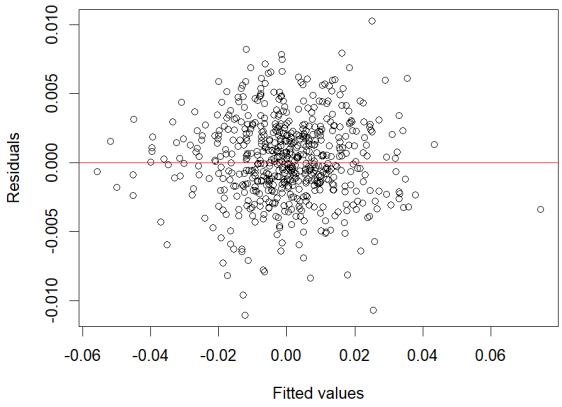


Figure 10: Q-Q plot of the residuals of the MV3 model vs. fitted

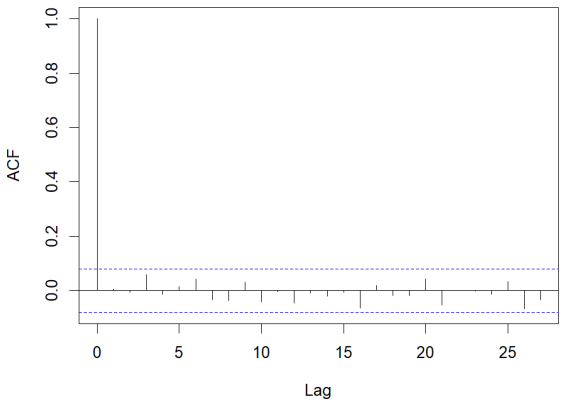


Figure 11: ACF function of residuals of the MV3 model

Final step was to conduct forecast analysis for which we selected two moodels as benchmarks. For the first benchmark we selected Fama-French model with all 5 factors (MKTRF + SMB + HML + RMW + CMA) without an intercept which proved to be non-significant when the model was fitted on the training data. For the second benchmark we selected Fama-French model with all 5 factors without an intercept and added January dummy which proved to be almost significant with p-value of 0.0534. We tested models in the similar fashion as we did in the univariate case, by performing rolling forecast and calculating MAE and RMSE. Results given in Table 3 indicate that our model outperforms both benchmarks for 1-step-ahead and 2-step-ahead forecasts. To formally test if this difference is attributed to the superior forecasting performance of our model we performed Diebold-Mariano Test which actually showed that all regression models have equal predictive power. Also, all regression models outperformed ARMA(1,1)*.

Model	MAE 1	RMSE 1	MAE 3	RMSE 3
MV3	0.002158347	0.002762714	0.002829665	0.008351524
FF	0.00219023	0.002807892	0.006715878	0.008366203
FF + Jan	0.002206757	0.002829665	0.006728503	0.008367236

Table 3: MAE and RMSE for 1 and 3 periods ahead forecast for ARMA(1,1), ARMA(1,1) without intercept and ARMA(5,8)

We conclude this section by presenting forecasts for MV3 and ARMA(1,1)* against actual values for one-period-ahead forecasts in Figure 12 a and three-periods-ahead forecasts in Figure 13.

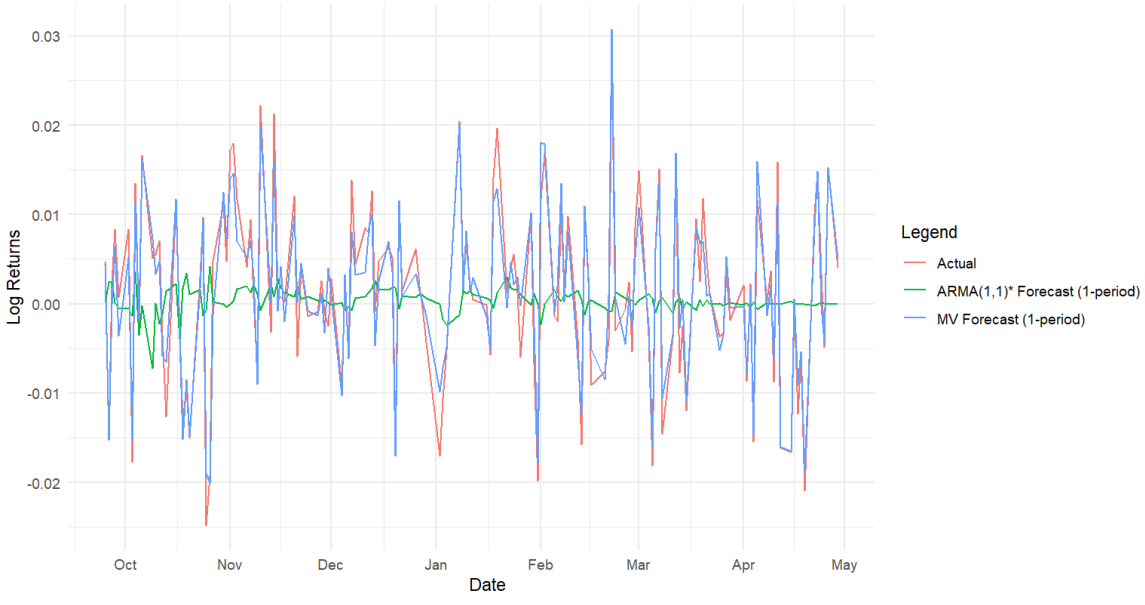


Figure 12: One-period-ahead forecasts MV3 vs ARMA(1,1)*

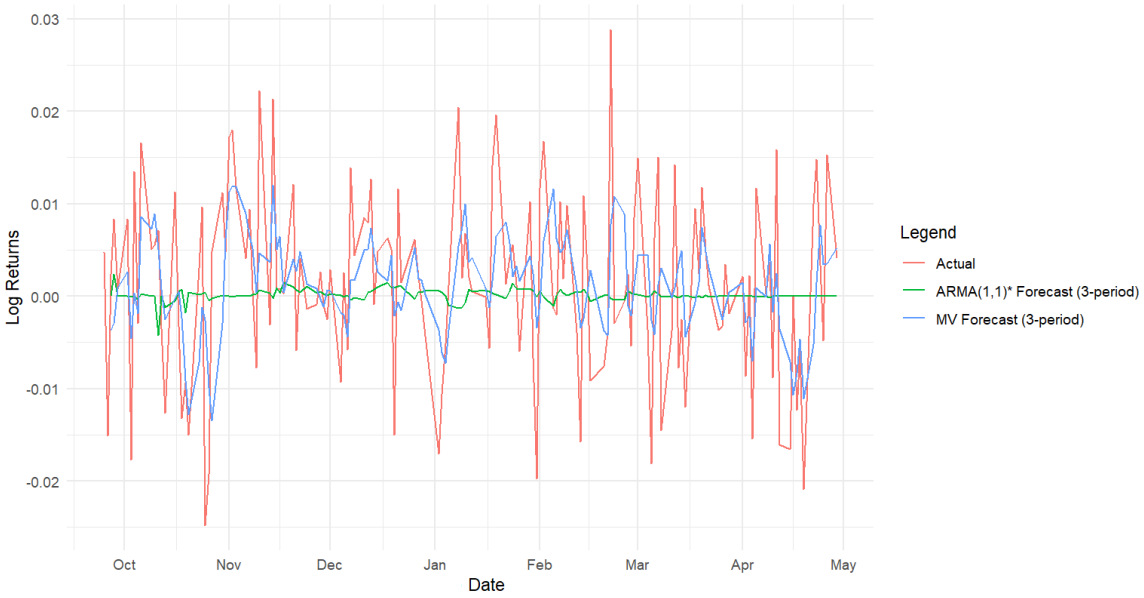


Figure 13: Three-period-ahead forecasts MV3 vs ARMA(1,1)*

AIC	BIC
GARCH(3,1)-norm −5.6764	GARCH(1,1)-norm −5.6459
GARCH(2,1)-norm −5.6752	GARCH(2,1)-norm −5.6386
GARCH(1,1)-norm −5.6752	GARCH(1,2)-norm −5.6356

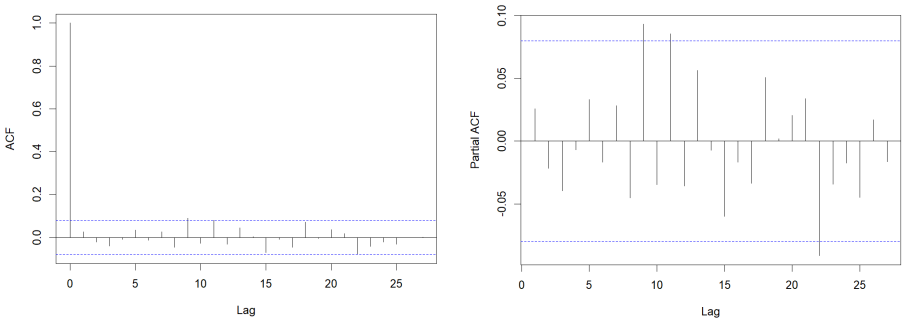
Table 4: Best performing GARCH(p,q) and ARMA(1,1)+GARCH(p,q) models under the assumption of the normal distribution of the residuals

5 Building a volatility model

For the purpose of volatility model building, we decided to use the residuals of the winning univariate model from the Phase 2, namely the ARMA(1,1) model without the intercept. ACF and PACF plots of the residuals of the aforementioned model in Figure 14 showed no major serial correlation, despite lag 9, 11 and 22 being borderline over the line of the 95% confidence interval. The lack of serial correlation was further confirmed by the Ljung-Box test.

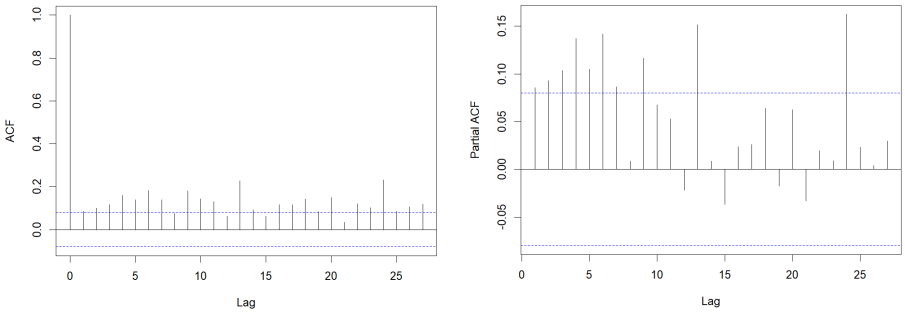
We proceeded to test for the presence of the ARCH effect in the squared residuals. ACF and PACF plots in Figure 15 confirmed the presence of serial autocorrelation which was further confirmed by the Ljung-Box test. ARCH LM-test confirmed that the null hypothesis claiming that there is no ARCH effect could be rejected with a very high level of significance.

Given these conclusions, we started working on the appropriate volatility model selection. Initially, we assumed the normal distribution of the residuals and tried fitting simple GARCH models as well as ARMA(1,1) + GARCH models with the range of autoregressive terms $1 \leq p \leq 5$ and the range of moving average terms $0 \leq q \leq 5$. We ranked the resulting models by AIC and BIC as shown in Table 4.



(a) ACF (b) PACF

Figure 14: ACF and PACF plots of the residuals of the ARMA(1,1) model without the intercept



(a) ACF (b) PACF

Figure 15: ACF and PACF plots of the squared residuals of the ARMA(1,1) model without the intercept

We proceeded to fit the two best performing models ranked both by AIC and by BIC: GARCH(1,1)-norm, GARCH(2,1)-norm and GARCH(3,1)-norm. In GARCH(1,1)-norm both α_1 and β_1 coefficients were highly significant, indicating that both the short-term shock and the long-term volatility persist over time, while in GARCH(2,1), only β_1 coefficient was significant, as was the case with GARCH(3,1).

Jarque-Bera and Shapiro-Wilk tests indicated no significant distribution from normality, while Ljung-Box test for residuals and squared residuals of all three models indicated no significant autocorrelation. The LM Arch test also showed no signs of the ARCH effect in the residuals.

Although the assumption of a normal distribution of the residuals withstood the tests, we still decided to fit all three models with the Student's t distribution as well as GED. We then ranked the resulting models again by AIC, BIC and Log-Likelihood shown in Table 5.

Given that the first three models by the Log Likelihood were all versions of GARCH(1,1) - student's t, followed by normal and GED, indicating that they fit the data the best, as well as that they are the only ones which had most of their coefficients significant, we chose the GARCH(1,1)-std as the best at this

Model	AIC	AIC Rank	BIC	BIC Rank	Log-L	Log-L Rank
GARCH(3,1)-norm	-5.676	1	-5.632	4	-1711.749	8
GARCH(3,1)-ged	-5.675	2	-5.624	8	-1712.490	9
GARCH(2,1)-norm	-5.675	3	-5.639	2	-1710.399	5
GARCH(1,1)-norm	-5.675	4	-5.646	1	-1709.397	2
GARCH(2,1)-ged	-5.675	5	-5.631	6	-1711.294	7
GARCH(1,1)-ged	-5.674	6	-5.637	3	-1710.025	3
GARCH(2,1)-std	-5.671	7	-5.628	7	-1710.277	4
GARCH(3,1)-std	-5.671	8	-5.620	9	-1711.245	6
GARCH(1,1)-std	-5.669	9	-5.632	5	-1708.491	1

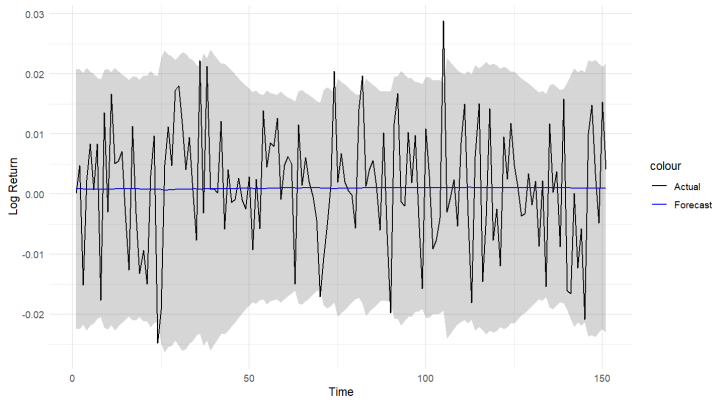
Table 5: Ranking of GARCH(1,1), GARCH(2,1) and GARCH(3,1) fitted with normal, student's t and GED distributions by AIC, BIC and Log Likelihood

point. We proceeded to evaluate all three models on the test data.

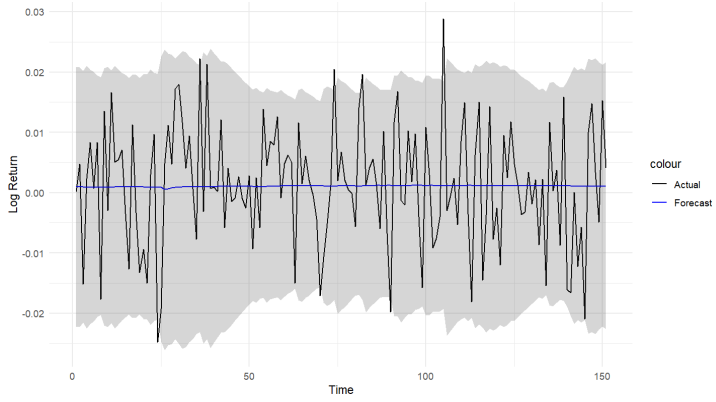
Model evaluation was performed using the rolling one-period-ahead forecast scheme and forecast intervals at 95%. It is shown in both Figure 16 and Table 6 that the models performed almost exactly the same with a hit rate of 95.36% at a 95% confidence interval indicating that they all have almost the same power for capturing the interval where the actual values will fall. However, further evaluation using MAE and RMSE revealed very slight differences in their forecast accuracy, giving advantage to the GARCH(1,1)-std, followed by GARCH(1,1)-ged and finally GARCH(1,1)-norm. To confirm if these differences in accuracies were indeed due to model superiority, we performed Diebold-Mariano Test which indicated, as expected, that all models have equal predictive power, i.e. that differences in MAE (and RMSE) were likely due to sampling variability.

Model	Hit Rate	MAE	RMSE
GARCH(1,1)-std	95.36%	0.007769225	0.009960505
GARCH(1,1)-ged	95.36%	0.007773214	0.009962512
GARCH(1,1)-norm	95.36%	0.007775599	0.009963876

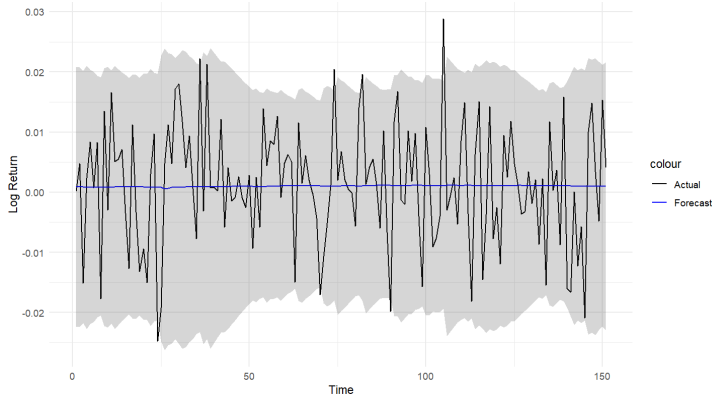
Table 6: Evaluated GARCH models and their corresponding hit rate at 95% confidence interval, MAE and RMSE



(a) GARCH(1,1)-norm



(b) GARCH(1,1)-std



(c) GARCH(1,1)-ged

Figure 16: Forecast intervals for GARCH(1,1) with normal distribution, student's t distribution and GED

6 Conclusion

For our final project for the course Statistics and Financial Data Analysis at Master in Computational Finance we chose to analyse returns of Invesco QQQ Trust (QQQ) ETF due to its broad tech exposure, data availability, liquidity and convenience regarding volatility analysis. After data collection and preparation phase, we modeled its log-returns with an univariate ARMA(1,1) model without an intercept, which proved to have the best out-of-sample performance. Afterwards, we modeled QQQ log returns with regression model having explanatory variables as SP500 og return, SMB, HML and CMA. We chose explanatory variables according to stepwise, general-to-specific selection. We compared our model to two benchmarks - Fama-French model without intercept and with/witoud January dummy, as January dummy appear to be almost-significant. Our analysis showed that all three regression models have equal predictive power, and they all outperform ARMA(1,1) without an intercept. Finally, we conducted volatility model building, after detecting the ARCH effect in QQQ squared residuals. Our in-sample analysis left us with GARCH(1,1) - Student's t, normal and generalized error distribution, which we tested on our test data by performing rolling forecasts. Our results showed that all three models have equal predictive power.

Pease note that the forecasts, accuracy metrics and forecasting errors for the section 3 - Building a univariate time series model for forecast are saved in folder Phase 2. These might be useful if you wish to run some parts of this code and skip the lengthy rolling forecast for ARMA(5,8).

References

[1] Mark M. Carhart. On persistence in mutual fund performance. *The Journal of Finance*, 52(1):57–82, 1997.

[2] Eugene F. Fama and Kenneth R. French. The cross-section of expected stock returns. *The Journal of Finance*, 47(2):427–465, 1992.

[3] Eugene F. Fama and Kenneth R. French. A five-factor asset pricing model. *Journal of Financial Economics*, 116(1):1–22, 2015.