Modeling Volatility in the S&P 500 and its Futures Market

Alexander Lerner
Department of Statistics
alerner1@stanford.edu

Isaac Schaider
Department of Sociology
schaider@stanford.edu

Abstract

This paper explores various models for forecasting the S&P 500 index as well as its corresponding futures market. We start by covering the concepts behind GARCH and E-GARCH modeling, then determine the optimal models for our data, and finally analyze the performance of these models for forecasting the volatility of the S&P 500 and its futures market. Modeling the volatility of this index is of great interest, and we explore the connection between the variance of the index with its futures market, a largely speculative market on its underlying performance. We find that the optimal models for forecasting volatility are the same for both indexes, but have lower predictive power on the inherently more volatile futures market.

1 Introduction

1.1 The S&P 500

The Standard and Poor's 500, often denoted as the S&P 500, is a stock market index covering the performance of 500 of the largest public companies in the United States. As one of the most followed indexes, it is seen as a proxy for the performance of the United States equity market as a whole. The index's value is capitalization weighted, which indicates that companies with the largest market capitilizations have the greatest influence on the pricing of the index.

1.2 The Futures Market

Futures are a financial derivative that are seen as an agreement to purchase or sell a specific asset at a date in the future for a set price. In general, futures are used for speculation on the underlying asset or to hedge an investment position. Futures allow investors to speculate by not having to actually own the underlying asset (by buying and selling the contract up to its expiration time). Investors can also use futures contracts to hedge by protecting against either upside or downside risk by their current position on the asset.

Futures markets are inherently more speculative, as the margin requirements for trading are generally lower, which allow investors to take greater risks (higher leverage). This leverage can then increase volatility in the futures markets, and as a result provide an interesting area of inquiry.

We will thus use the example of the S&P 500 the S&P 500 futures in our project. We will explore the performance of certain time series models to generalize to both a futures and asset market, and explain some of our methods below.

2 Methods

2.1 Model Formulations

Financial markets, especially broader indexes such as the S&P 500, experience periods of volatility, where day-to-day volatility is closely linked to the market's reception of macroeconomic news or other shocks. This renders certain models such as ARMA less useful, as they rely on the assumption of a constant conditional variance, or a variance that does not depend on time. In reality, the variance of a period of trading may very well be influenced by the time period before it, especially in response to news such as increasing interest rates or developing international news. As such, simple ARMA models may not be able to capture this form of financial market behavior without consideration of non-constant variance. For this reason we will introduce ARCH/GARCH modeling, which treat heteroskedasticity as a phenomenon to be modeled. [1]

Engle's introduction of ARCH, or Autoregressive Conditional Heteroscedasticity, processes became a significant model for the forecast of volatility in financial markets (initially inflation rates).[2] For example, let us denote the price of an asset on the end of a day in trading (in this case our asset is the value of the S&P 500 index) as X_t. If we are interested in our day-to-day returns, we let $Y_t = \frac{(X_t - X_{t-1})}{X_{t-1}}$.

Assuming our process Y_t is zero mean, an ARCH(1) process would be defined as:

$$y_t = \sigma_t \epsilon_t$$

with

$$Var(y_t|y_{t-1}) = \sigma_t^2 = \alpha_0 + \alpha_1 y_{t-1}^2$$

with $\epsilon_t \sim WN(0, \sigma_w^2)$ which implies that the variance of our process at time t is defined by the value at the previous time step. This would imply that a large value of volatility at step $t-1$ implies a large value of volatility for step t in our process. Note that the ARCH model can be extended to p autoregressive terms.

A generalization of ARCH models parameterized by Bollerslev (1986) known as GARCH introduced further complexity similar to the extension of an AR process to that of an ARMA process.[3] As such, a GARCH(p,q) process is defined as:

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i y_{t-i}^2 + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2$$

which is to say that the variance in a time period can be modeled by a weighted average of variances and past observations. The parameters of interest can be estimated using Maximum Likelihood Estimation (MLE). [1]

We fit several GARCH models to our data, while conducting important statistical tests to evaluate our models (explained further under the Evaluation section). Engel (2001) points out asymmetric GARCH models may be useful since basic GARCH models currently only factor in the magnitude of returns rather than their direction. [1] Thus, we also fit an Exponential GARCH – or E-GARCH – model to our data.

The E-GARCH model captures a quality that is not included in the GARCH model: the empirically observed fact that negative shocks at time $t-1$ have a stronger impact in the variance at time t than positive shocks. This asymmetry used to be called leverage effect because the increase in risk was believed to come from the increased leverage induced by a negative shock. The EGARCH model was proposed by Nelson (1991) to overcome the weakness in GARCH handling of asymmetric effects in financial time series. [4] Formally, an E-GARCH (p, q) can be formulated as:

$$x_t = \mu + a_t$$

$$\ln \sigma_t^2 = \alpha_o + \sum_{i=1}^{p} \alpha_i \left(|\epsilon_{t-i}| + \gamma_i \epsilon_{t-i}\right) + \sum_{j=1}^{q} \beta_j \ln \sigma_{t-j}^2$$

$$a_t = \sigma_t \times \epsilon_t$$

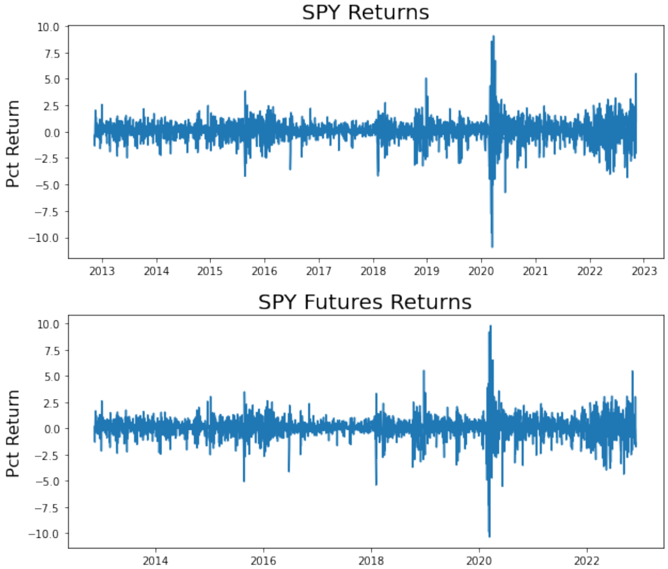
$$\epsilon_t \sim P_v(0, 1)$$

Where x_t is the time series value at time t; μ is the mean of the GARCH model; a_t is the model's residual at time t; σ_t is the conditional standard deviation (i.e. volatility) at time t; p is the order of the ARCH component model; $\alpha_o, \alpha_1, \alpha_2, \ldots, \alpha_p$ are the parameters of the ARCH component model; q is the order of the GARCH component model; and $\beta_1, \beta_2, \ldots, \beta_q$ are the parameters of the GARCH component model. $[\epsilon_t] \sim i.i.d.$ are the standardized residuals following the the probability distribution function P_v with mean 0 and variance of 1.

2.2 Data

We acquired our data on the historical prices of the S&P 500 (SPY) from MarketWatch.com and the historical prices of the S&P 500 Futures from Investing.com. Both datasets ran from 11/11/2012 to 11/11/2022. We calculated the returns by dividing the change in price by the price on the previous day then multiplying by 100. Next, we converted the data in the Date column from type string to DateTime then used the Date column as our index. Finally, we checked that the returns data was indeed stationary by graphing the time series:

Figure 1: Returns SPY and Futures



2.3 Model Selection

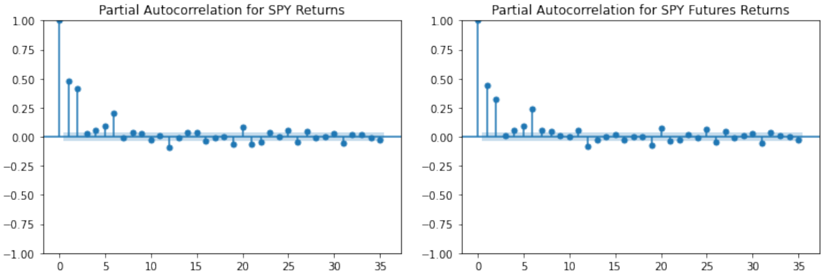
A GARCH model assumes a perfect fit for the conditional variance equation. This feature is due the construction of the GARCH model since there is no error term in the conditional variance equation in a GARCH model.

Measuring the goodness of fit of conditional variance models is problematic because the conditional variance is unobserved. Thus conventional techniques such as running a regression of the dependent variable (the conditional variance) on the regressors (e.g. lagged conditional variances and lagged squared error terms) and taking the R^2 as a measure of fit do not work.

You can assess how "good" a GARCH model is by looking at model residuals and checking how well they match the model assumptions. The AIC and BIC gives you model likelihood adjusted for the number of parameters so as to penalize for over-fitting. The primary difference between the two metrics is that BIC penalizes a model if it is trained on a higher number of samples.

In order to select the optimal GARCH (and E-GARCH models), we first plotted the PACF graph for both the SPY returns and SPY futures returns. The number of significant lags informed the range of values that we tested to find the optimal GARCH models. We can see in Figure 2 that the first 7 lags from the PACFs for the SPY and futures data were significant:

Figure 2: Autocorrelation for SPY and Futures



Thus, we measured the AIC and BIC for GARCH and E-GARCH models with the p and q parameters ranging from 1 to 7.

2.4 Forecasts

After selecting for the optimal model using AIC and BIC, we split our data into train and test sets with the last 21 trading days (roughly equivalent to one month) comprising the test set. After training our models, we ran two forecasts with differing time horizons for the last 21 trading days. The first forecast had a 21-day time horizon while the second forecast was rolling. Using the root mean squared error (RMSE), we compared these forecasts with the realized volatility of the test set. Realized volatility is the square root of the realized variance and measures the standard deviation in the daily return of an underlying index over a given period.

3 Results

3.1 SPY Model Selection

We first ran GARCH models on the SPY data for $p = [1, 7]$ and $q = [1, 7]$ and calculated the AIC and BIC for each model. We can see in Figure 3 that the calculations for AIC and BIC both indicate that the GARCH(1,1) is the optimal model for the SPY data. As shown in the Table 1, the AIC of the GARCH(1,1) was 6261.5 while the BIC was 6284.8. The p-values for the α_1 and β_1 parameters were both highly significant.

Figure 3: AIC & BIC Heatmaps (GARCH, SPY)

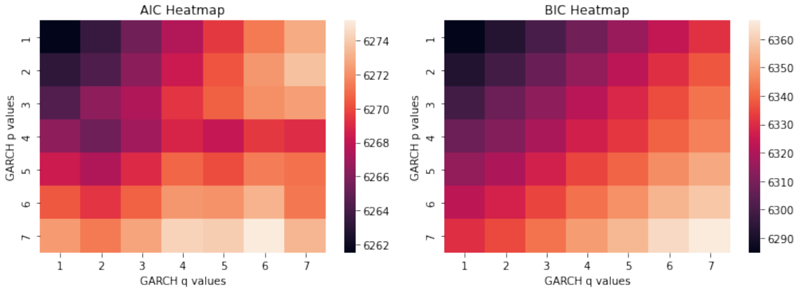
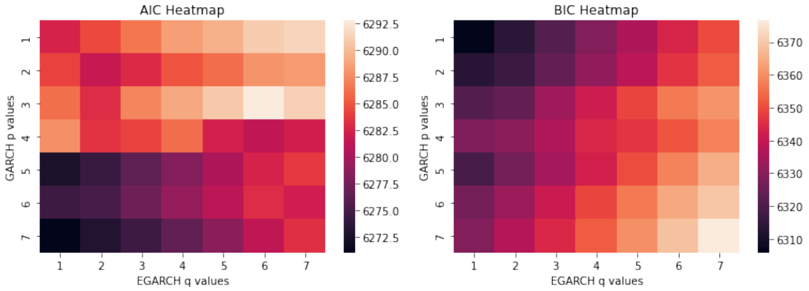


Table 1

	GARCH(1,1)		
	coefficient	std error	p-value
ω	0.0421	9.214e-03	4.945e-06
α_1	0.2233	3.079e-02	4.042e-13
β_1	0.7468	2.724e-02	1.822e-165
AIC	6261.52		
BIC	6284.84		
n	2518		

When we repeated our model selection process using the E-GARCH model, the AIC and BIC analyses identified different optimal models. We can see in Figure 4 that AIC identifies the GARCH(7,1) as the optimal model while BIC identifies GARCH(1,1):

Figure 4: AIC & BIC Heatmaps (E-GARCH, SPY)



When examining the model statistics of the E-GARCH(7,1) and the E-GARCH(1,1) shown in the Table 2, we noticed two things. First, 5 of the 7 α-parameters from the E-GARCH(7,1) model were not significant. Second, when we compare the E-GARCH(7,1) and the E-GARCH(1,1), the AIC improved by 11.5 points in E-GARCH(7,1) while BIC improved by 23.5 in the E-GARCH(1,1). Therefore, given the lack of significance in the E-GARCH(7,1) and the greater improvement in BIC for the E-GARCH(1,1), we concluded that the E-GARCH(1,1) was the optimal model for the SPY data, which corroborates with the selected GARCH model discussed previously.

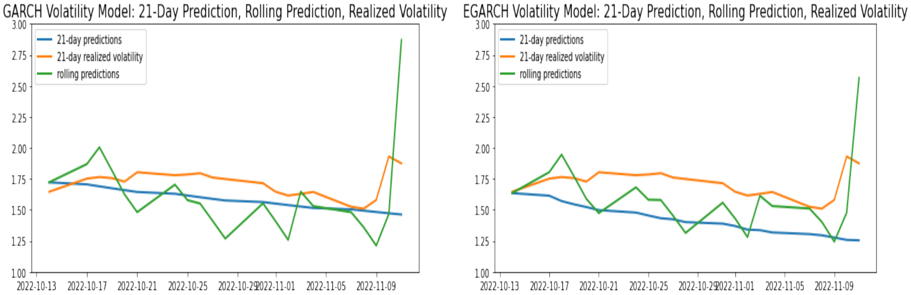
Table 2

	E-GARCH(1,1)			E-GARCH(7,1)		
	coefficient	std error	p-value	coefficient	std error	p-value
ω	-9.215e-04	8.658e-03	4.945e-06	2.415e-03	4.666e-03	0.605
α_1	0.4015	4.286e-02	7.386e-21	0.3900	5.734e-02	1.031e-11
α_2				3.181e-03	6.404e-02	0.960
α_3				-0.0143	6.912e-02	0.837
α_4				0.0757	7.865e-02	0.336
α_5				-0.1979	8.109e-02	1.465e-02
α_6				0.0737	6.599e-02	0.264
α_7				-0.1017	5.660e-02	7.243e-02
β_1	0.9360	1.324e-02	0.000	0.9728	1.339e-02	0.000
AIC	6282.47			6270.97		
BIC	6305.79			6329.28		
n	2518			2518		

3.2 SPY Forecasts

We used our GARCH(1,1) and E-GARCH(1,1) models to forecast the volatility of SPY for the final month of our dataset. We implemented a rolling forecast with a time horizon of one day and a longer forecast with a time horizon of 21 days (the number of trading days in a given month).

Figure 5: SPY Volatility Forecasts



To gauge the performance of our forecasts, we compared them with the realized volatility measure for the final month of SPY data using RMSE. As detailed in Table 3, the GARCH(1,1) performed better on the 21-day forecast while the E-GARCH performed better on the rolling forecast.

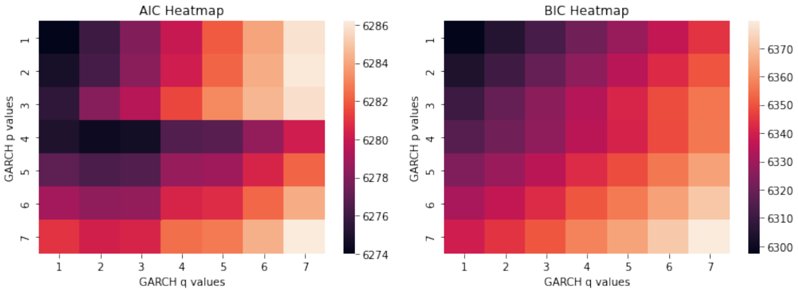
Table 3

	GARCH(1,1)	E-GARCH(1,1)
Forecast	RMSE	RMSE
Rolling	0.33	0.27
21-day	0.18	0.33

3.3 Futures Model Selection

We ran GARCH models on the futures data for $p = [1, 7]$ and $q = [1, 7]$ and calculated the AIC and BIC for each model. We can see in Figure 6 that the calculations for AIC and BIC both indicate that the GARCH(1,1) is the optimal model for the futures data.

Figure 6: AIC & BIC Heatmaps (GARCH, Futures)



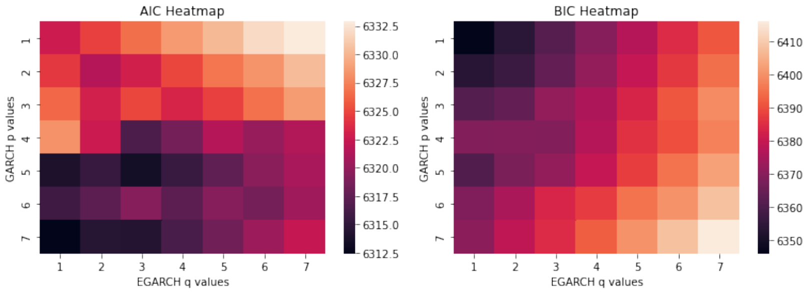
As shown in the Table 4, the AIC of the GARCH(1,1) was 6274.03 while the BIC was 6297.40. The p-values for the α_1 and β_1 parameters were both highly significant.

Table 4

	GARCH(1,1)		
	coefficient	std error	p-value
ω	0.0429	9.835e-03	1.279e-05
α_1	0.2422	3.511e-02	5.223e-12
β_1	0.7325	2.953e-02	7.097e-136
AIC	6274.03		
BIC	6297.40		
n	2518		

When we repeated our model selection process using the E-GARCH model, the AIC and BIC analyses identified different optimal models. We can see in Figure 7 that AIC identifies the GARCH(7,1) as the optimal model while BIC identifies GARCH(1,1):

Figure 7: AIC & BIC Heatmaps (E-GARCH, Futures)



When examining the model statistics of the E-GARCH(7,1) and the E-GARCH(1,1) shown in the Table 5, we noticed two things. First, 5 of the 7 α-parameters from the E-GARCH(7,1) model were not significant. Second, when we compare the E-GARCH(7,1) and the E-GARCH(1,1), the AIC improved by 18.9 points in E-GARCH(7,1) while BIC improved by 53.94 in the E-GARCH(1,1). Therefore, given the lack of significance in the E-GARCH(7,1) and the greater improvement in BIC for the E-GARCH(1,1), we concluded that the E-GARCH(1,1) is the optimal model for the futures data, which corroborates with the selected GARCH model discussed previously.

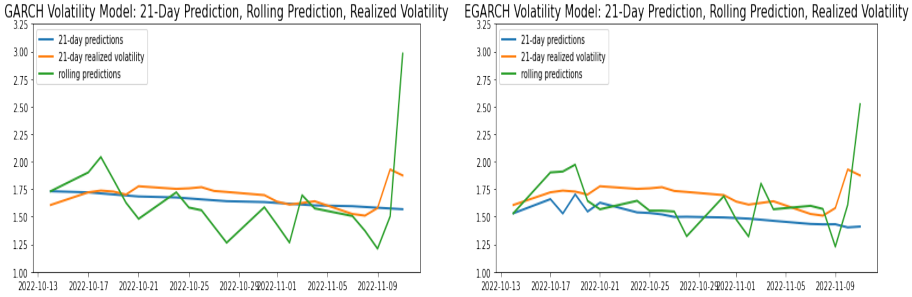
Table 5

	E-GARCH(1,1)			E-GARCH(7,1)		
	coefficient	std error	p-value	coefficient	std error	p-value
ω	1.099e-03	9.910e-03	0.912	3.321e-03	5.369e-03	0.536
α_1	0.4296	5.166e-02	9.177e-17	0.4153	7.215e-02	8.603e-09
α_2				1.584e-04	7.220e-02	0.998
α_3				-0.0441	6.987e-02	0.5287
α_4				0.108	9.140e-02	0.236
α_5				-0.195	9.211e-02	3.472e-02
α_6				0.069	6.821e-02	0.313
α_7				-0.109	6.958e-02	0.119
β_1	0.9292	1.472e-02	0.000	0.9728	1.339e-02	0.000
AIC	6293.54			6312.42		
BIC	6316.91			6370.85		
n	2518			2518		

3.4 Futures Forecasts

We used our GARCH(1,1) and E-GARCH(1,1) models to forecast the volatility of the SPY futures for the final month of our dataset. We implemented a rolling forecast with a time horizon of one day and a longer forecast with a time horizon of 21 days (the number of trading days in a given month).

Figure 8: Volatility Forecasts for Futures


GARCH Volatility Model: 21-Day Prediction, Rolling Prediction, Realized Volatility

EGARCH Volatility Model: 21-Day Prediction, Rolling Prediction, Realized Volatility

To gauge the performance of our forecasts, we compared them with the realized volatility measure for the final month of the futures data using RMSE. As detailed in the Table 6, the GARCH(1,1) performed better on the 21-day forecast while the E-GARCH performed better on the rolling forecast.

Table 6

	GARCH(1,1)	E-GARCH(1,1)
Forecast	RMSE	RMSE
Rolling	0.34	0.25
21-day	0.12	0.22

4 Conclusion

Our analysis reveals two findings regarding modeling volatility of the SPY and its futures market. First, our volatility models performed better on the SPY data than the futures data when comparing the AIC of the optimal models. The GARCH(1,1) and E-GARCH(1,1) have lower AICs for the SPY data than for the futures data. The increased speculation in the futures market due to the lower margin requirements likely contributes to the amplified volatility of the SPY futures market and the increased difficulty in modeling.

Second, for both the SPY and futures data, the GARCH(1,1) models had more accurate forecasts for the 21-day time horizon than the E-GARCH(1,1) models (using realized volatility over the 21-day period as the benchmark). Conversely, the E-GARCH(1,1) models had more accurate forecasts for the one day ahead rolling time horizon than the GARCH(1,1). The E-GARCH model's improved short-term forecasting power could arise from its ability to model asymmetric positive and negative effects, which could allow the model to better capture the day-to-day swings in returns. Further research is needed to understand whether GARCH and E-GARCH models perform better for differing forecasting time horizons, and future work could explore novel measures of volatility to benchmark forecasting ability.

References

[1] Robert Engle. Garch 101: The use of arch/garch models in applied econometrics. *Journal of Economic Perspectives*, 15(4):157–168, December 2001.

[2] Robert F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50(4):987–1007, 1982.

[3] Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327, 1986.

[4] Daniel B Nelson. Conditional heteroskedasticity in asset returns: A new approach. *Econometrica: Journal of the econometric society*, pages 347–370, 1991.