Final Project

Zihao Lin (zl293)

Nov. 15th. 2020

Part 1 Summary

This project aims to analyze the time series characteristics of every day profit and forecast the real volatility of S&P 500 index. Using ARMA model, I found that AR(2) model could fit the profit of stock price best however the diagnostics shows that the AR(2) is not good enough. Then I use ARIMA(2, 0, 0)-GARCH(1,1) model to fit the error term of profit and then use this model to predict the volatility of profit. The final result shows that the correlation between predicted value of volatility and realized volatility is 79.6% which is a good result.

Part 2 Introduction

Stock price changes with time, therefore people always consider how to use time series model to analyze the stock market data such as close price and profit. Many people hope to use time series model to predict the stock price however most of them failed. That's not surprised because the stock market is fundamentally a zero-sum game. In addition to predicting the close price, there is another more meaningful topic, predicting the real volatility of the stock everyday profit. Many financial derivatives has implied volatility which is based on the volatility of subject matter which may often be a stock price. Predicting the volatility of stock price will help dealers more easily to price the financial derivatives and also may find some arbitrage opportunity hidden deeply.

This project aims to analyze the time series characteristics of every day profit and forecast the real volatility of SP 500 index. The S&P 500 is a stock market index that measures the stock performance of 500 large companies listed on stock exchanges in the United States. I will use ARMA model to analyze the time series characteristics of profits and use GARCH model to predict the volatility of profit. Because we do not cover the GARCH model in our class, I will use one part to indroduce the GARCH model.

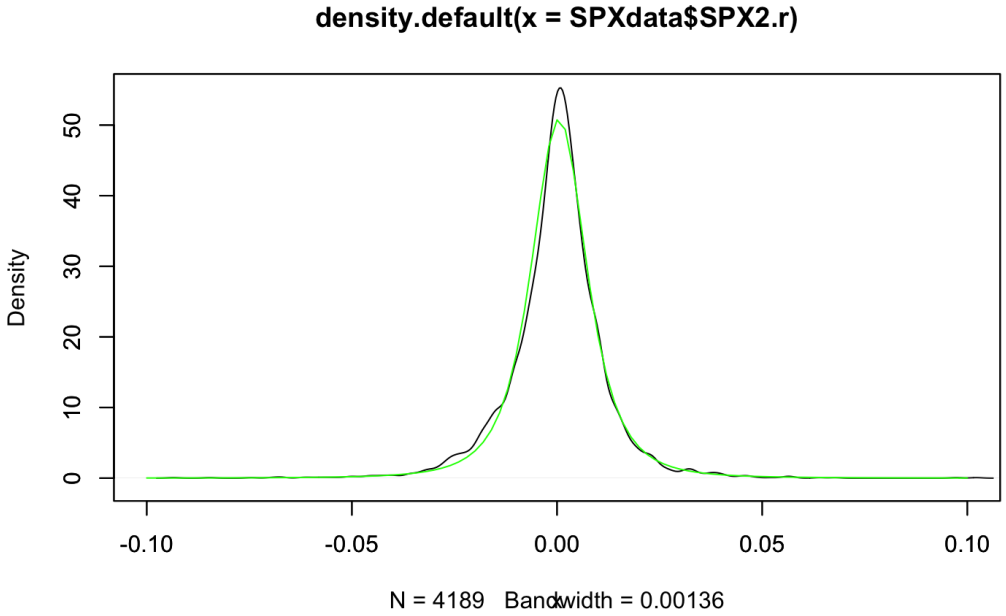
Part 3 Data

I gained the data (SPXdata) from website which has been processed well. The data includes the date (DATE) close price (SPX2.closeprice), open price (SPX2.openprice), profit (SPX2.r) and realized volatility (SPX2.rvol) of S&P 500 index from Jan. 03rd 2020 to Oct. 06th 2016. The data set contains 4189 observations. Here I have to mention that the original dataset also contains some other data which I did not figure out what they are, therefore I only show the data that I will use in the project.

We should notice that actually the volatility of profit is invisible because we do not know exactly this time what the volatility is. The only way to get the volatility is to estimate it which is the goal of my project. I still need a standard value of volatility, which is the SPX2.rvol in my dataset. It is realized volatility that estimated by HEAVY model which is a model applied to estimating the volatility. HEAVY model is not the goal of my report so I would not talk about it. Now I just assumed that the estimated volatility from HEAVY model is believable, unbiased and efficient. I would use it as the realized volatility and compare my predicted valatility to it to check the property of my prediction model.

Before selecting the model, the first thing I should do is to check the time series characteristics. I plot the acf and pacf plots, finding that the there is no value in lags larger than 0.05 in

acf plot. However, in pacf plot, when lag equals to 1 and 2, the value of pacf is obviously larger than 0.05. This means that I can use AR(2) model to fit the profit of S&P 500. After checking the lags, I checked the stationary characteristics of profit. The p-value of Dickey-Fuller Test is 0.01, which means that I should accept the alternative hypothesis – the time series is stationary. The p-value of KPSS Test is 0.1, which means that I should accept null hypothesis – the time series is stationary. Both test shows that the time series of profit is stationary then we can safely fit the time series model. However, something happens when I check the normality of time series of profit. The p-value of Shapiro-Wilk normality test is smaller than 0.05 which means that the distribution of profit is not normality. The value of kurtosis is 10.56322, which indicates that the plot of density of profit is kind of leptokurtosis and heavy tails. Then I plot the density of profit (black one) and the t-distribution density line (green one) and found that they are very similar. This characteristics will be useful in the following work.



Part 4 Model

I. ARIMA Model of profit

I fit the model using the function auto.arima in R. The fitted model is shown below:

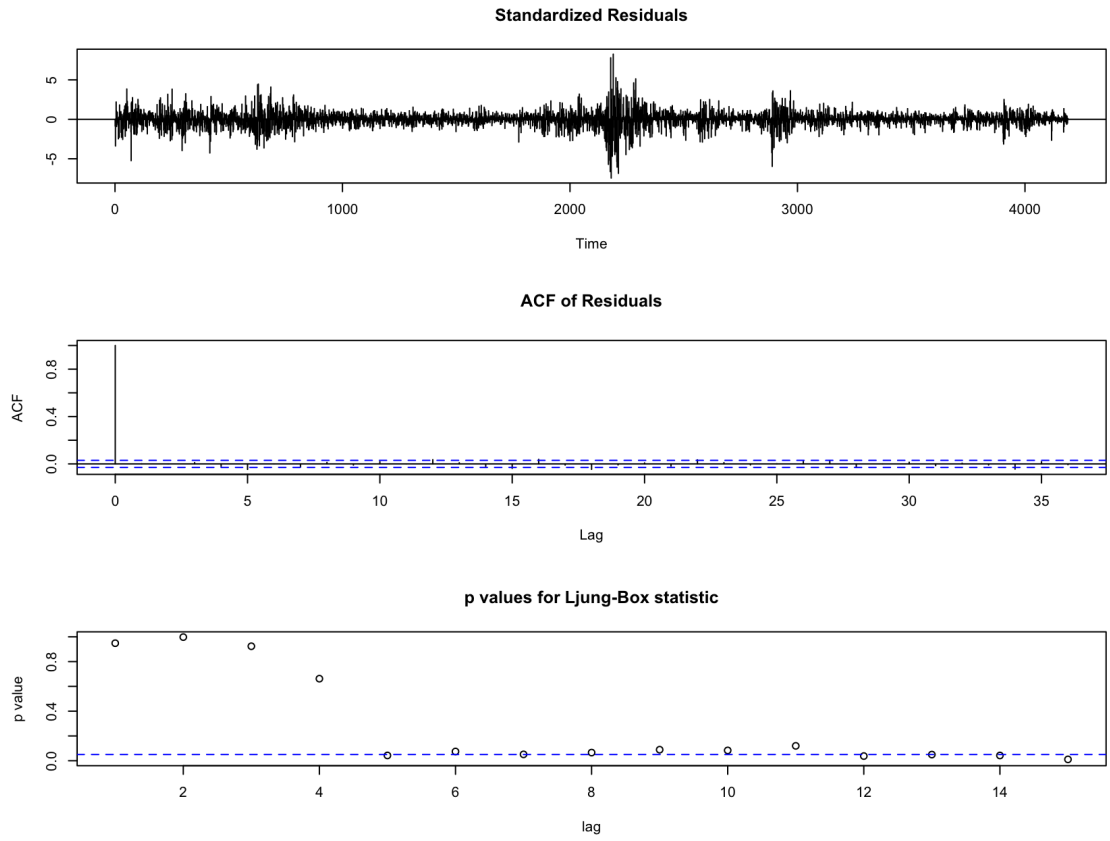
$$y_t = 0.0001 - 0.0839y_{t-1} - 0.0633y_{t-2} + \epsilon_t \tag{1}$$

which satisfies what I guessed in the previous analysis.

After I get the model, I did some model diagnostics. I drew the several plots using tsdiag function in R which used to model diagnostics. The command tsdiag displays three of our diagnostic tools in one display – a sequence plot of the standardized residuals, the sample ACF of the residuals, and p-value for the Ljung-Box test statistic for a whole range of values of lags from 1 to 15.

In plot of standardized residuals, there are many residuals (more than 10) in the series with magnitudes larger than 3 which is very unusual in a standard normal distribution. This means the model is not good enough. According to the ACF of residuals, there is no evidence of autocorrelation in the residuals of this model, which is good! We can see that when lag is more than 4, the p-value of Ljung-Box statistic is near 0, which means that we should reject the null hypo, and there exists self-correlation.

I also draw the qq-norm plot to see the normality of the residuals. QQ norm plot shows that the residuals does not obey the normal distribution. Also, according to the Shapiro-Wilk normality test, the p-value is smaller than 0.05, which means that the residuals do not obey normal distribution.



According to the result of diagnostics, we can see that the AR(2) model is not good enough. However, because the AR(2) model is given by the auto.arima function in R which using AIC to select the best model, and because my project aims to predict the volatility of profit, I just kept the AR(2) model of profit and use it in the follwing analysis to predict the volatility of profit. If the model diagnostic of the GARCH model which applied to predict the volatility is good, I believed the final results is good.

II. Introduction to GARCH model

Our goal is to estimate the variance of profit, however simple ARMA model could not help us. Therefore, we introduce the GARCH model, generalized autoregressive conditional heteroskedasticity model which is based on ARCH model, autoregressive conditional heteroscedasticity model. ARCH model is a statistical model for time series data that describes the variance of the current error term or innovation as a function of the actual sizes of the previous time periods' error terms; often the variance is related to the squares of the previous innovations. The ARCH model is appropriate when the error variance in a time series follows an AR model. If an ARMA model is assumed for the error variance, the model is a GARCH model.

The GARCH(p, q) model (where p is the order of GARCH terms σ^2 and q is the order of ARCH terms ϵ), is given by:

$$y_t = x_t^{'} + \epsilon_t \tag{2}$$

$$\epsilon_t|\psi_{t-1} \sim N(0, \sigma_t^2) \tag{3}$$

$$\sigma_t^2 = \omega + \alpha_1\epsilon_{t-1}^2 + ... + \alpha_q\epsilon_{t-q}^2 + \beta_1\sigma_{t-1}^2 + ... + \beta_p\sigma_{t-p}^2 = \omega + \sum_{i=1}^{q}\alpha_i\epsilon_{t-i}^2 + \sum_{i=1}^{p}\beta_i\sigma_{t-i}^2 \tag{4}$$

ARCH models are commonly employed in modeling financial time series that exhibit time-varying volatility and volatility clustering, i.e. periods of swings interspersed with periods of relative calm.

III. Fit GARCH model and Diagnostics

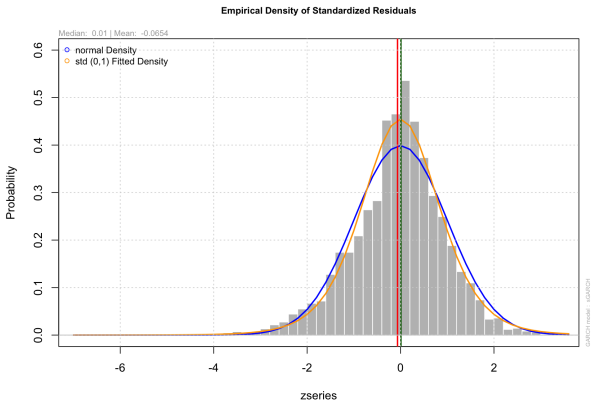
I was not quite sure how to select the garch model, therefore, I searched on the internet and saw that in general case, people will use GARCH(1,1) model as a base line of prediction. Now I just used the experienced ARIMA(2,0,0)-GARCH(1,1) model as my prediction model.

First, I used rugarch package to fit the GARCH model, ugarchspec function could do that.

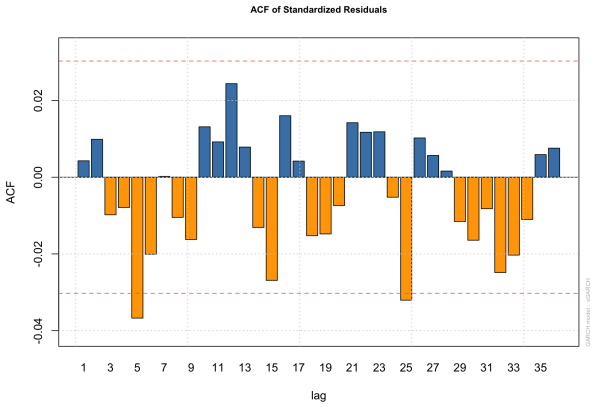
	Estimate	Std. Error	z value	$Pr(>\|z\|)$
mu	0.000591	0.000105	5.63914	0.000000
ar1	-0.060876	0.015650	-3.88995	0.000100
ar2	-0.044445	0.015862	-2.80200	0.005079
omega	0.000001	0.000001	0.96658	0.333755
alpha1	0.102885	0.021758	4.72858	0.000002
beta1	0.890753	0.020647	43.14138	0.000000
shape	7.090451	0.404279	17.53851	0.000000

I need expalin some values in the above table. I have *mu* parameter estimated since I have selected $include.mean = TRUE$ in the functino of ugarchspec, which means that my final model includes the ARIMA(2,0,0). The $ar1$ and $ar2$ is the coefficients of the first lag and second lag in ARIMA(2,0,0) part. The p-values of above three parameters are all smaller than 0.05 which means that they are significant. The parameter *omega* in your model is the variance intercept parameter. Here the p-value of *omega* is larger than 0.05 which means it is insignificant. $Alpha1$ is the ARCH(q) parameter. In my case, q is 1. $Beta1$ is the GARCH(p) parameter. In my case, p is 1. To understand parameter *shape*, please note that in Part3, I mentioned that the profit obeys t-distribution but not normal distribution. Here, the *shape* is one of the parameters to define the t-distribution. These several parameters all have p-values smaller than 0.05 which means they are significant.

According to above, the model looks good. Then I have to do some diagnostics. According to the following left plot, I can see that the residuals looks normal. The blue line is a normal distribution and the green line is the residual. That looks good! According to the following right plot, I can see that there exists no obvious autocorrelation because the value of acf is not larger than 0.05, which is pretty good!

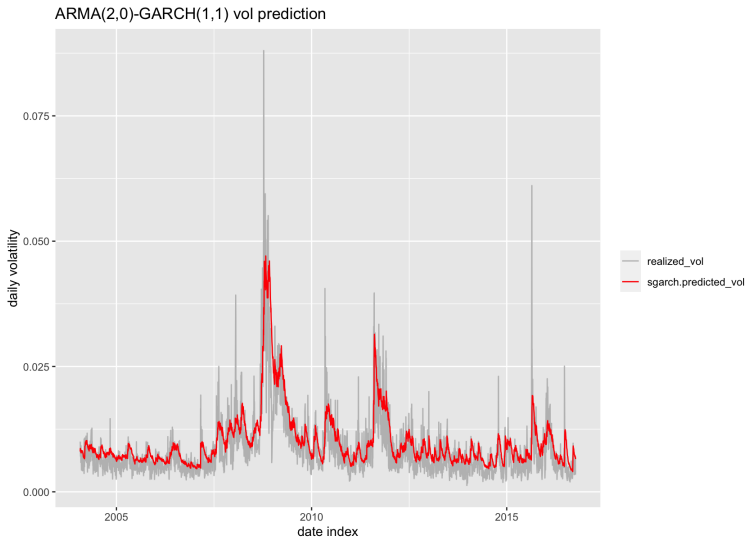


I also checked the residuals plot. Most of the residuals is not more than 0.05 and also not more than two times standard variance which is vary good. According to these diagnostics, we can see that the moel is good enough. And I can use this model to predict the volatility.



IV. Predict Volatility

My goal is to predict the volatility, therefore, I used 4189 observations to do backtest (from Jan. 3rd 2000 to Oct. 06th 2016). I used the first 1000 observations to train the model and each time rolled to predict the next value and refitted the model every five observations. The result is shown below:



I used two way to evaluate the model. The MSE of the result is 1.767135e-05 which is good. I also checked the correlation between the predicted value and the estimated value (or in other words the realized value of volatility). The correlation is 0.7960698, which is also great. It means that my result could predict the volatility in a good way.

This result is not good enough because I in every day, the predicted value is not exactly the same as the realized value. Therefore, next step is to find a new model or in other words a more complicated and better model to predict the volatility.

Part 5 Conclusion

Predicting the volatility of stock price will help dealers more easily to price the financial derivatives and also may find some arbitrage opportunity hidden deeply. It is meaningful for financial market. I used MA(2) model to fit the profit of S&P 500 index, although the fitted model is not good enough, we can still use it in our final prediction model. Our final prediction model is ARIMA(2,0,0)-GARCH(1,1) model. I predicted the volatility in more than 3000 days and the final result is good. The correlation between my predicted value and the realized value is 79.6%. There is still some work to improve the final model and also improve the fitted model of the profit.