Modeling S&P Composite using GARCH model

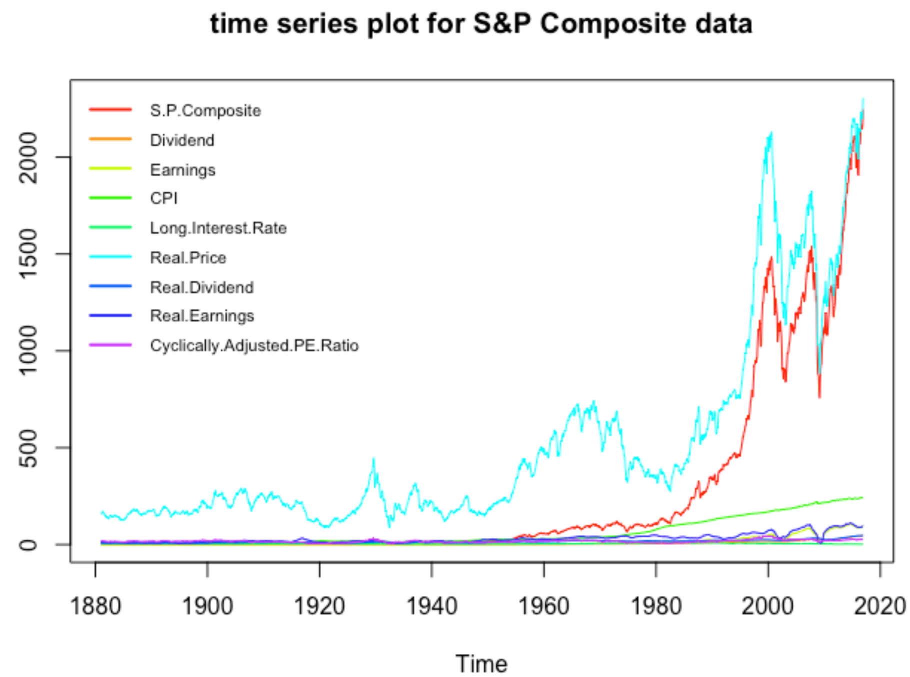
Haining Zhang, Qinqing Liu

1. Introduction

The volatility of this S&P 500 stock index returns can be seen as a measurement of the risk for investment and provides essential information for the investors to make the correct decisions.

The S&P Composite data set is collected by Yale Department of Economics (https://www.quandl.com/data/YALE/SPCOMP-S-P-Composite). This data set consists of monthly stock price, dividends, and earnings data and the consumer price index (to allow conversion to real values), etc, all starting January 1871. We delete NA values in first 10 years and get the data over the period Jan, 1881 through Dec, 2016. Time series plot for all 9 variables shows as follows.

In this dataset, CPI is the Consumer Price Index; Dividend is a distribution of a portion of a company's earnings; Earnings is an after-tax net income of the company; Long interest Rates refer to government bonds maturing in ten years; Real Price are adjusted for general price level changes over time; Cyclically Adjusted PE. Ratio is defined as price divided by the average of ten years of earnings, adjusted for inflation.



Following is the part of the dataset:

Year	S&P Compo	Dividend	Earnings	CPI	Long Interes	Real Price	Real Divider	Real Earning	Cyclically Adjusted PE Ratio
2016/12/31	2246.63	45.7	94.55	241.432	2.49	2305.83118	46.9042455	97.041497	27.8650982
2016/11/30	2164.99	45.4766667	92.73	241.353	2.14	2222.7672	46.6903048	95.2046903	26.8509535
2016/10/31	2143.02	45.2533333	90.91	241.729	1.76	2196.78854	46.3887431	93.1909392	26.5251431
2016/9/30	2157.69	45.03	89.09	241.428	1.63	2214.58421	46.217356	91.4391349	26.7278733
2016/8/31	2170.95	44.84	88.3666667	240.849	1.56	2233.55042	46.1329836	90.9147632	26.9488724
2016/7/31	2148.9	44.65	87.6433333	240.628	1.5	2212.89512	45.9796952	90.2533876	26.6940033
2016/6/30	2083.89	44.46	86.92	241.018	1.64	2142.47666	45.7099521	89.3636761	25.8403729
2016/5/31	2065.55	44.2666667	86.76	240.229	1.81	2130.59579	45.6606588	89.4921406	25.6947099
2016/4/30	2075.54	44.0733333	86.6	239.261	1.81	2149.56202	45.6451639	89.6885008	25.9223375
2016/3/31	2021.95	43.88	86.44	238.132	1.89	2103.98887	45.6603931	89.9472283	25.3722986
2016/2/29	1904.42	43.7166667	86.47	237.111	1.78	1990.22335	45.6863144	90.3658927	24.0026068
2016/1/31	1918.6	43.5533333	86.5	236.916	2.09	2006.69253	45.553085	90.4716482	24.2061672

2. Goal of Analysis

In order to follow the bond market, it is important to learn about the S&P Composite index of stocks because the volatility of this S&P Composite stock index returns can be seen as a measurement of the risk for investment and provides essential information for the investors to make the correct decisions.

The S&P C ... rgest
corporatio ... tock
Exchange o ... to have a
price that p ... ny. Also,
the S&P Co ... for the
economy d ... ll areas of
the United ... me series
analysis, w ... es data.

In this case ... RCH
model and ... s as
following, v ... ck Index
Returns in

$$
\begin{cases}
X_t = \boldsymbol{\beta}'\boldsymbol{z_t} + y_t^* & (1) \\
\phi(B)y_t^* = \theta(B)y_t & (2) \\
y_t = \sigma_t \varepsilon_t & (3) \\
\sigma_t^2 = \alpha_0 + \sum_{j=1}^{m} \alpha_j y_{t-j}^2 + \sum_{j=1}^{s} \beta_j \sigma_{t-j}^2 & (4)
\end{cases}
$$

where $\boldsymbol{\beta}'\boldsymbol{z_t}$ is a function of exogenous predictors.

3. Comprehensive Data Analysis.

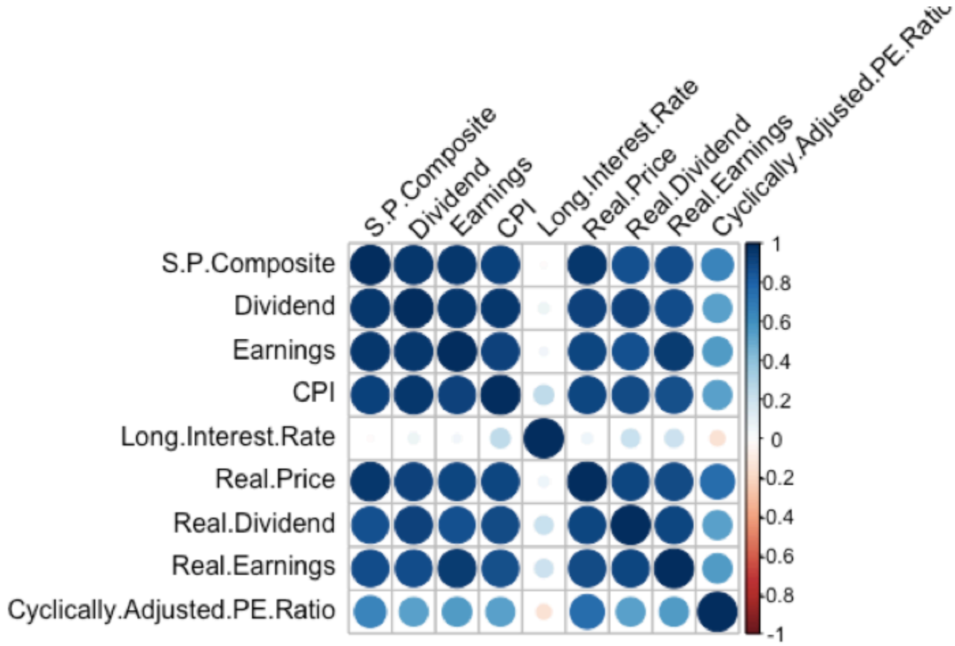
(1) Linear Regression.

Firstly, we fit a full linear regression model with Dividend, Earnings, Real Dividend, Real Earnings, CPI (Consumer Price Index), Long Interest

Rate, Real Price and Cyclically Adjusted PE Ratio and obtain regression residuals.

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -84.312 -18.255  -2.383  20.032  73.674
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)

## (Intercept)                 157.412098   4.405286  35.733  < 2e-16
***
## Dividend                     30.566873   0.757761  40.338  < 2e-16
***
## Earnings                      4.050867   0.292287  13.859  < 2e-16
***
## CPI                          -0.757145   0.047086 -16.080  < 2e-16
***
## Long.Interest.Rate           -4.489523   0.469905  -9.554  < 2e-16
***
## Real.Price                    0.708420   0.007627  92.885  < 2e-16
***
## Real.Dividend               -18.037391   0.533901 -33.784  < 2e-16
***
## Real.Earnings                -1.682304   0.230640  -7.294 4.68e-13
***
## Cyclically.Adjusted.PE.Ratio  -5.829591   0.243546 -23.936  < 2e-16
***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.15 on 1623 degrees of freedom
## Multiple R-squared:  0.9963, Adjusted R-squared:  0.9963
## F-statistic: 5.532e+04 on 8 and 1623 DF,  p-value: < 2.2e-16
```

We also draw correlation plot for all variables and find most variables have significant high positive correlations (>0.8) with S&P Composite. So use these variables to fit a regression model is reasonable.



Obviously, there are collinearity between dividend and real dividend, earnings and real earnings. Therefore, we turn to obtain the VIF values of these variables.

VIF values:

```
##                    Dividend                     Earnings
##                   90.630964                    88.901893
##                         CPI            Long.Interest.Rate
##                   20.931561                     2.375851
##                  Real.Price                Real.Dividend
##                   29.378285                    32.956594
##               Real.Earnings Cyclically.Adjusted.PE.Ratio
##                   54.280547                     5.004143
```

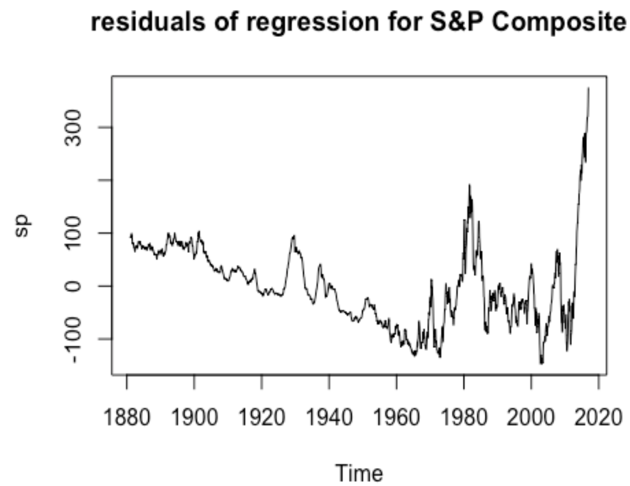
So, after drop two variables Dividend and Earnings who have larger VIF, we go on to fit the reduced model as following (let $Z_{1t}, Z_{2t}, Z_{3t}, Z_{4t}, Z_{5t}$):

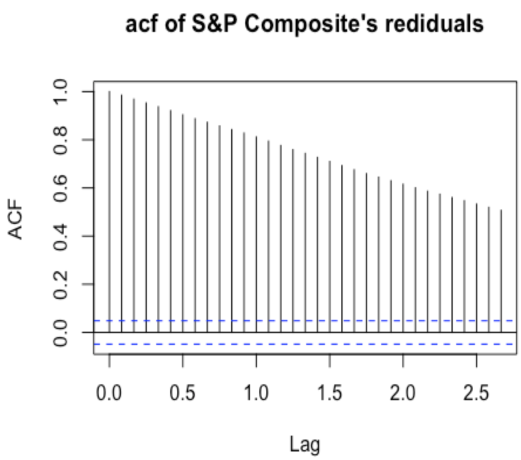
```
## Call:
## lm(formula = S.P.Composite ~ ., data = data3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -147.12  -51.71   -7.44   55.64  375.45
##
## Coefficients:
##                                Estimate Std. Error  t value Pr(>|t|)

## (Intercept)                   184.89841   10.71348   17.26  < 2e-16 *
**
## CPI                             2.36787    0.08186   28.93  < 2e-16 *
**
## Long.Interest.Rate            -32.39503    0.90167  -35.93  < 2e-16 *
**
## Real.Price                      0.86657    0.01890   45.86  < 2e-16 *
**
## Real.Dividend                 -10.79104    0.69952  -15.43  < 2e-16 *
**
## Real.Earnings                   0.85080    0.22568    3.77 0.000169 *


## Cyclically.Adjusted.PE.Ratio -13.61284    0.58071  -23.44  < 2e-16 *
**
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 74.28 on 1625 degrees of freedom
## Multiple R-squared:  0.9762, Adjusted R-squared:  0.9762
## F-statistic: 1.113e+04 on 6 and 1625 DF,  p-value: < 2.2e-16
```

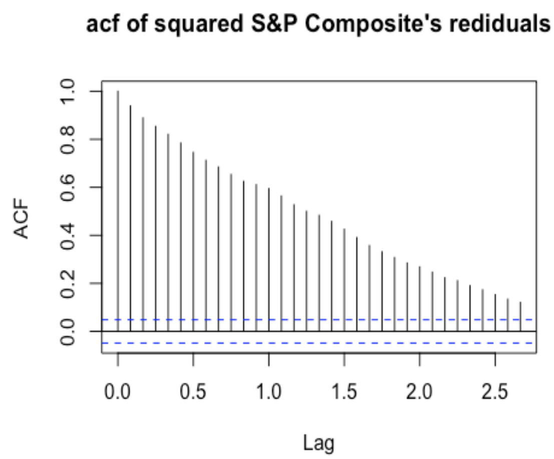
Therefore, (1) can be:

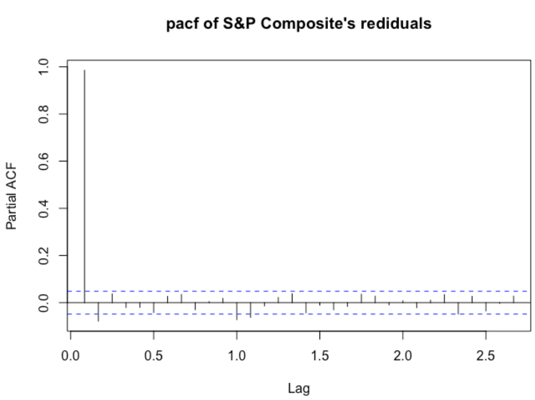
$$X_t = 184.89841 + 2.3679 Z_{1t} - 32.395 Z_{2t} + 0.8666 Z_{3t} - 10.7910 Z_{4t} + 0.8508 Z_{5t} + y_t^*$$

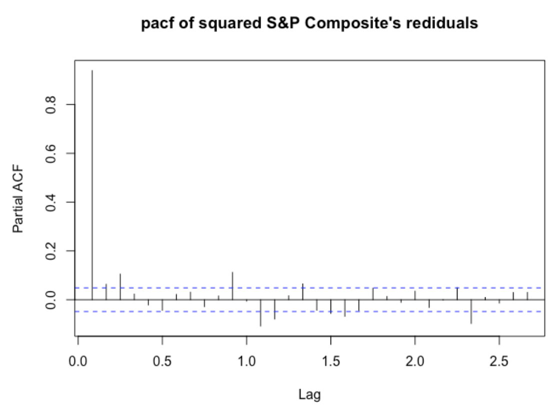
Next, we need to fit the ARMA+GARCH model to the residuals (y_t^*) of this linear regression. Before fitting this final model, it is necessary to check the time series plot of y_t^* as well as ACF and PACF plots of both y_t^* and y_t^{*2}. The plots are as following:











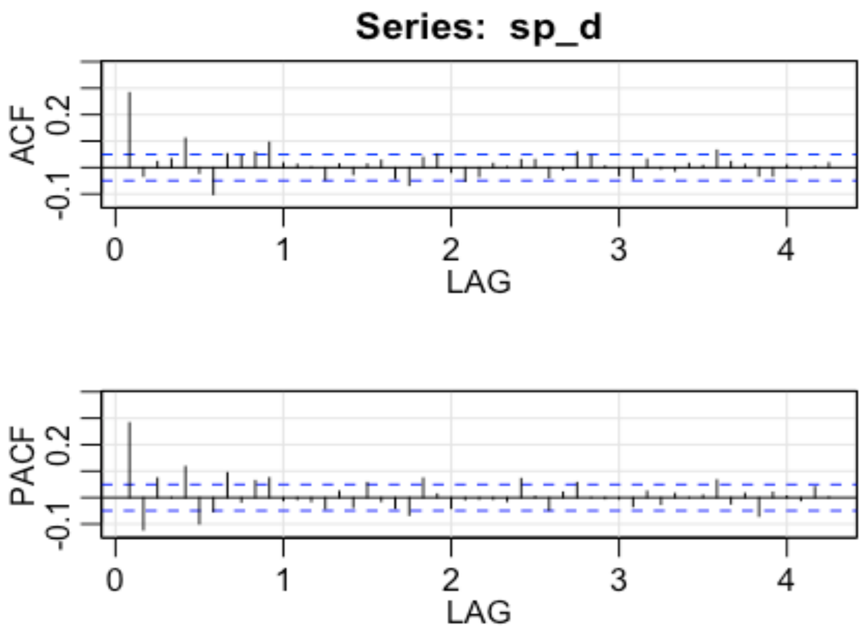
From the plots, we find obvious trend in the time series plot of the y_t^*. Also, the ACF and PACF plots are not good enough.

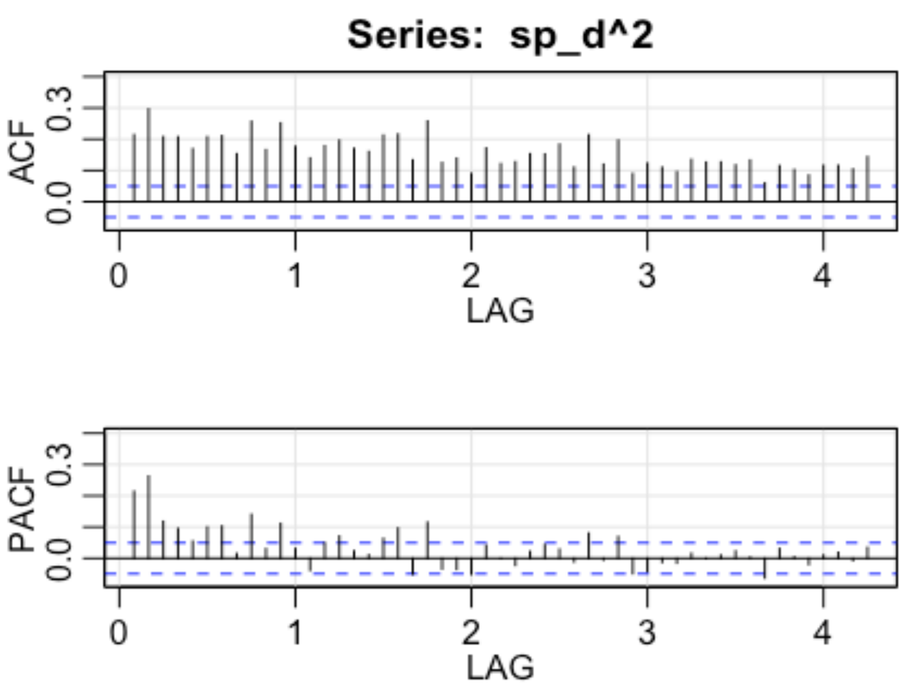
In addition, we need to use the Augmented Dickey-Fuller Test and Phillips-Perron test to check the stationarity of the y_t^* and y_t^{*2} as following:.

```
##
## Augmented Dickey-Fuller Test
##
## data:  sp
## Dickey-Fuller = -0.89051, Lag order = 11, p-value = 0.9535
## alternative hypothesis: stationary


##  Phillips-Perron Unit Root Test
##
## data:  sp
## Dickey-Fuller Z(alpha) = 0.10731, Truncation lag parameter = 8,
## p-value = 0.99
## alternative hypothesis: stationary

##
##  Augmented Dickey-Fuller Test
##
## data:  sp^2
## Dickey-Fuller = 5.9939, Lag order = 11, p-value = 0.99
## alternative hypothesis: stationary

##
##  Phillips-Perron Unit Root Test
##
## data:  sp^2
## Dickey-Fuller Z(alpha) = 57.284, Truncation lag parameter = 8,
## p-value = 0.99
## alternative hypothesis: stationary
```

From all p-values we obtained above, we can conclude that the residuals y_t^* and its square are non-stationary. So, in order to remove the trend, we try to do difference of the y_t^* and mark it as ‘sp_d’.

Following are the ACF and PACF plots of both ‘sp_d’ and the squre of the ‘sp_d’:





From these plots, we find that the ACF and PACF of 'sp_d' have some patterns and decay into blue dotted lines with the lag values increasing. In addition, the ACF and PACF plots of the square of 'sp_d' have obvious patterns. Therefore, these all results show that we need to fit ARMA+GARCH model to the dataset.

It is also necessary to check the stationarity again. From the ADF test and PP test following, we can reject the null hypothesis (non-stationary) and conclude that the series are stationary.

```
##
##  Augmented Dickey-Fuller Test
##
## data:  sp_d
## Dickey-Fuller = -9.6165, Lag order = 11, p-value = 0.01
## alternative hypothesis: stationary

## Warning in pp.test(sp_d): p-value smaller than printed p-value

##
##  Phillips-Perron Unit Root Test
##
## data:  sp_d
## Dickey-Fuller Z(alpha) = -1120.5, Truncation lag parameter = 8,
## p-value = 0.01
## alternative hypothesis: stationary
```

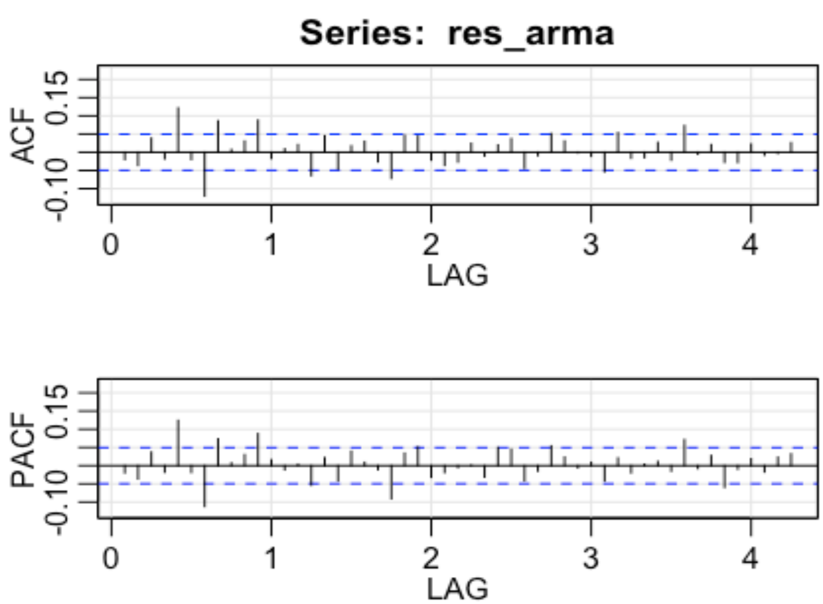
(2) ARMA Model.

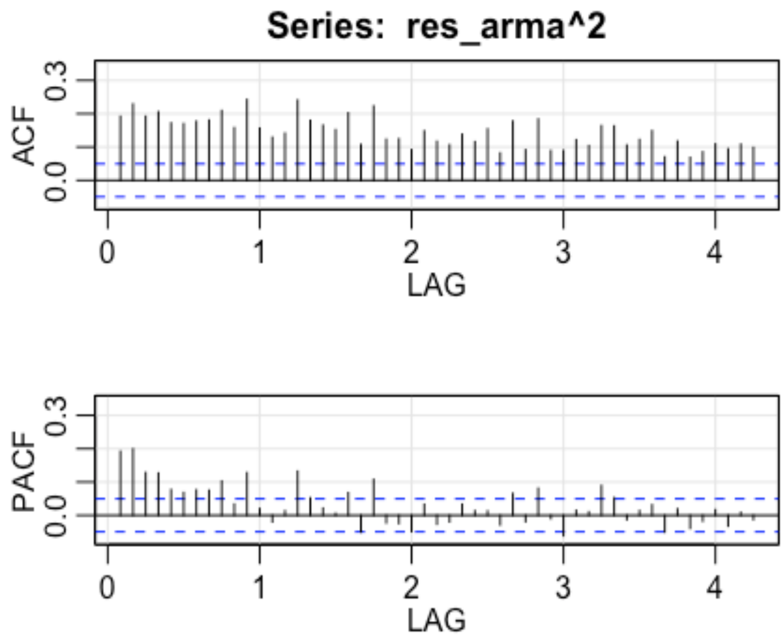
Before fitting the ARMA+GARCH model by garchfit {fGarch}, we are supposed to decide a best order for the ARMA model. So, we set loops to choose a model with the smallest BIC automatically. At last, we decide to use MA(1), whose BIC is 4.148424, to fit ‘sp_d’ as the ARMA part of the final model we will fit next.

```
##   p q      BIC
## 2 0 1 4.148424
```

Moreover, we also want to decide the order for the GARCH part. Due to the patterns showed in the ACF and PACF plots for the residuals ($\sigma_t \varepsilon_t$) of the ARMA model, we decide to use GARCH(1, 1) in the GARCH part.

```
## 
## Call:
## arima(x = sp_d, order = c(0, 0, 1), method = "CSS")
## 
## Coefficients:
##          ma1  intercept
##       0.3510     0.1737
## s.e.  0.0253     0.2650
## 
## sigma^2 estimated as 62.76:  part log likelihood = -5689.94
```





(3) ARMA+GARCH Model.

Finally, we use the garchFit {fGarch} to fit the final model to the 'sp_d' as following:

Model 1: We assume that the distribution of ε_t is standard normal.

From the result, we can see that the Jarque-Bera test and Shapiro-Wilk test can show that the normal assumption is not suitable. The skewness and excess kurtosis exist in the model distribution assumption because the p-value is small enough.

And we also use the LM Arch test to do the diagnostic of the model. The LM Arch test (p-value=0.9763>0.05) shows that the ε_t is uncorrelated, which conforms to the assumption of the GARCH-type model.

So this model is not a good fit for the data.

```
## Title:
##  GARCH Modelling
##
## Call:
##  garchFit(formula = ~arma(0, 1) + garch(1, 1), data = sp_d, trace =
FALSE)
##
## Conditional Distribution:
##  norm
##
## Coefficient(s):
##         mu        ma1      omega     alpha1      beta1
## -0.242206   0.354688   0.056163   0.162626   0.857883
##
## Std. Errors:
##  based on Hessian
##
## Error Analysis:
##         Estimate  Std. Error  t value Pr(>|t|)
## mu      -0.24221     0.08889   -2.725  0.00643 **
## ma1      0.35469     0.02386   14.864  < 2e-16 ***
## omega    0.05616     0.01183    4.746 2.07e-06 ***
## alpha1   0.16263     0.02022    8.044 8.88e-16 ***
## beta1    0.85788     0.01435   59.776  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
##
## Standardised Residuals Tests:
##                                 Statistic p-Value
##  Jarque-Bera Test   R    Chi^2  177.3047  0
##  Shapiro-Wilk Test  R    W      0.9877136 1.540478e-10
##  Ljung-Box Test     R    Q(10)  52.58574  8.887877e-08
##  Ljung-Box Test     R    Q(15)  56.4072   1.034172e-06
##  Ljung-Box Test     R    Q(20)  57.88576  1.50502e-05
##  Ljung-Box Test     R^2  Q(10)  4.152994  0.9401813
##  Ljung-Box Test     R^2  Q(15)  5.422665  0.9879026
##  Ljung-Box Test     R^2  Q(20)  8.028445  0.9916779
##  LM Arch Test       R    TR^2   4.348687  0.9762913
##
```

```
## Information Criterion Statistics:
##      AIC      BIC      SIC     HQIC
## 5.902742 5.919287 5.902724 5.908880
```

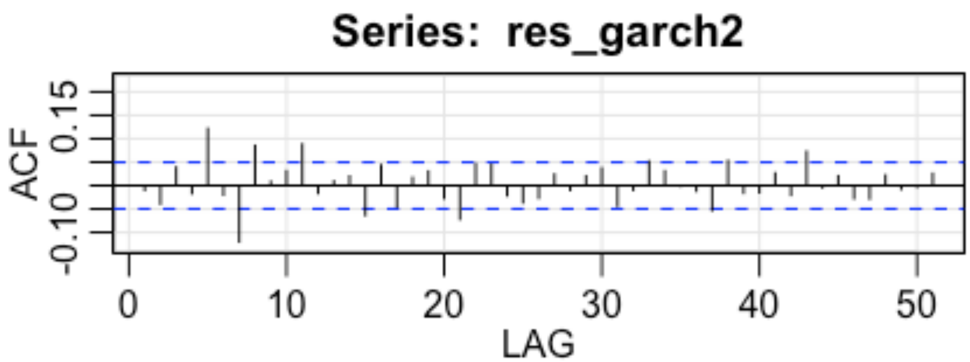
So we try to use to use non normal conditional distribution: standard t distribution and skewed t distribution.

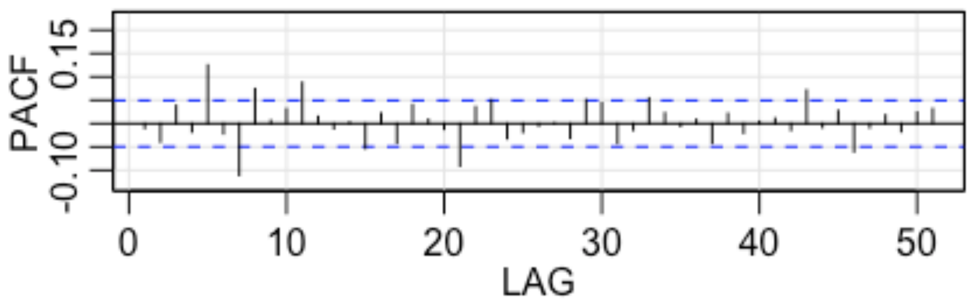
Model 2: Assume that the distribution of ε_t is standard Student's t with 5 d.f, mean=0 and SD=1. We can see the estimations of the model parameters are all significant for standard t distribution.

The Ljung-Box statistics indicate quite significant autocorrelations in standardized residuals since p-values are below 0.05, and no autocorrelations in squared standardized residuals. However, since this model is not fitted to the raw data, we use the LM Arch test to do the diagnostic of the model. The LM Arch test (p-value=0.9823>0.05) shows that the residuals are uncorrelated, which conforms to the assumption of the GARCH-type model. So we still conclude that the model does not exhibit significant lack of fit.

```
##
## Title:
##  GARCH Modelling
##
## Call:
##  garchFit(formula = ~arma(0, 1) + garch(1, 1), data = sp_d, cond.dis
t = "std",
##     trace = FALSE)
##
## Mean and Variance Equation:
```

```
##  data ~ arma(0, 1) + garch(1, 1)
## <environment: 0x7fbcb3068df0>
##  [data = sp_d]
##
## Conditional Distribution:
##  std
##
## Coefficient(s):
##        mu        ma1      omega     alpha1      beta1      shape
## -0.208595   0.340602   0.042987   0.155234   0.865298   7.127081
##
## Std. Errors:
##  based on Hessian
##
## Error Analysis:
##         Estimate  Std. Error  t value Pr(>|t|)
## mu      -0.20859     0.08391   -2.486   0.0129 *
## ma1      0.34060     0.02317   14.697  < 2e-16 ***
## omega    0.04299     0.01775    2.422   0.0154 *
## alpha1   0.15523     0.02308    6.725 1.76e-11 ***
## beta1    0.86530     0.01643   52.662  < 2e-16 ***
## shape    7.12708     1.17928    6.044 1.51e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log Likelihood:
##  -4777.167    normalized:  -2.92898
##
##
## Standardised Residuals Tests:
##                                 Statistic p-Value
##  Jarque-Bera Test   R    Chi^2  193.0316  0
##  Shapiro-Wilk Test  R    W      0.987196  7.828725e-11
##  Ljung-Box Test     R    Q(10)  54.21589  4.423043e-08
##  Ljung-Box Test     R    Q(15)  57.98403  5.582691e-07
##  Ljung-Box Test     R    Q(20)  59.37682  8.88872e-06
##  Ljung-Box Test     R^2  Q(10)  3.872842  0.952901
##  Ljung-Box Test     R^2  Q(15)  5.004112  0.9920915
##  Ljung-Box Test     R^2  Q(20)  7.926538  0.9923427
##  LM Arch Test       R    TR^2   4.059584  0.982337
##
## Information Criterion Statistics:
##      AIC      BIC      SIC     HQIC
## 5.865318 5.885172 5.865291 5.872684
```



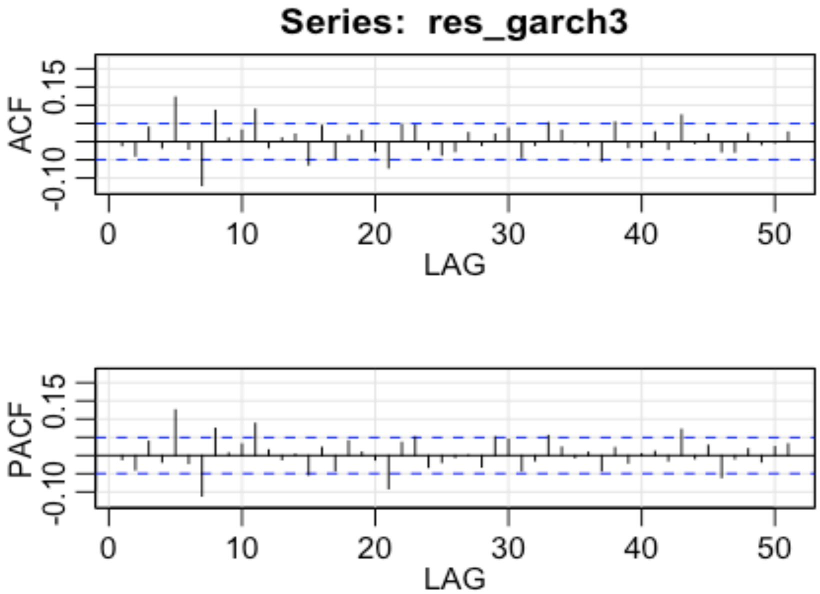


Model 3: Assume that the distribution of ε_t is skew-standard Student's t with 5 d.f, mean=0 and SD=1. The Ljung-Box statistics indicate quite significant autocorrelations in standardized residuals since p-values are below 0.05, and no autocorrelations in squared standardized residuals. However, since this model is not fitted to the raw data, we use the LM Arch test to do the diagnostic of the model. The LM Arch test (p-value=0.9825>0.05) shows that the ε_t is uncorrelated, which conforms to the assumption of the GARCH-type model.

```
##
## Title:
##  GARCH Modelling
```

```
##
## Call:
##  garchFit(formula = ~arma(0, 1) + garch(1, 1), data = sp_d, cond.dis
t = "sstd",
##     trace = FALSE)
##
## Mean and Variance Equation:
##  data ~ arma(0, 1) + garch(1, 1)
## <environment: 0x7fbcb17b54a0>
##  [data = sp_d]
##
## Conditional Distribution:
##  sstd
##
## Coefficient(s):
##        mu        ma1      omega     alpha1      beta1       skew
## -0.196536   0.341319   0.042294   0.154148   0.865940   1.016825
##     shape
##  7.179054
##
## Std. Errors:
##  based on Hessian
##
## Error Analysis:
##         Estimate  Std. Error  t value Pr(>|t|)
## mu      -0.19654     0.08799   -2.234   0.0255 *
## ma1      0.34132     0.02324   14.687  < 2e-16 ***
## omega    0.04229     0.01887    2.242   0.0250 *
## alpha1   0.15415     0.02297    6.711 1.93e-11 ***
## beta1    0.86594     0.01640   52.811  < 2e-16 ***
## skew     1.01682     0.03652   27.842  < 2e-16 ***
## shape    7.17905     1.19810    5.992 2.07e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log Likelihood:
##  -4777.059    normalized:  -2.928914
##
##
## Standardised Residuals Tests:
##                                 Statistic p-Value
##  Jarque-Bera Test   R    Chi^2  194.646   0
##  Shapiro-Wilk Test  R    W      0.9871448 7.32846e-11
##  Ljung-Box Test     R    Q(10)  54.40046  4.086138e-08
##  Ljung-Box Test     R    Q(15)  58.17188  5.185822e-07
##  Ljung-Box Test     R    Q(20)  59.5596   8.330052e-06
##  Ljung-Box Test     R^2  Q(10)  3.860603  0.9534164
##  Ljung-Box Test     R^2  Q(15)  4.982057  0.9922772
##  Ljung-Box Test     R^2  Q(20)  7.935103  0.9922885
##  LM Arch Test       R    TR^2   4.048839  0.9825389
```

##
##
##
##



Fi

st

th t

bo

T

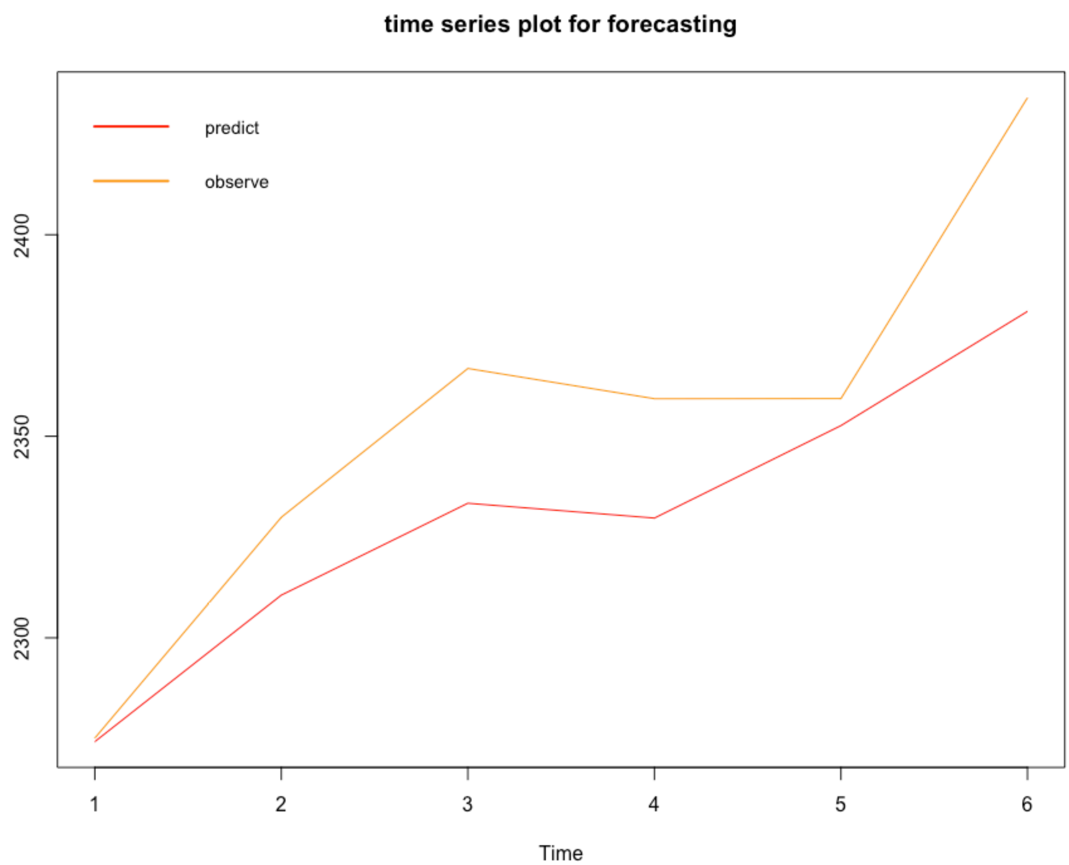
$$\begin{cases} X_t = 184.8984 + 2.3679z_{1t} - 32.395z_{2t} + 0.8666z_{3t} - 10.791z_{4t} \\ \qquad + 0.8508z_{5t} + y_t^* \\ y_t^* = (1 + 0.3406B)(y_t - 0.20859) \\ y_t = \sigma_t \varepsilon_t \\ \sigma_t^2 = 0.0430 + 0.1552y_{t-1}^2 + 0.8653\sigma_{t-1}^2 \end{cases}$$

4. Forecasting.

Finally, we use our final GARCH-M model to forecast the S&P Composite Index from January 2017 to June 2017 as following:

2274.243, 2310.593, 2333.384, 2329.704, 2352.646, 2380.969.

Also, we compare the real observed S&P Composite Index data with our forecasting values. Their overlay time-series plot is as following:



5

Ir

g(

tl

fc

$$
\begin{cases}
X_t = 184.8984 + 2.3679z_{1t} - 32.395z_{2t} + 0.8666z_{3t} - 10.791z_{4t} \\
\qquad +0.8508z_{5t} + y_t^* \\
y_t^* = (1 + 0.3406B)(y_t - 0.20859) \\
y_t = \sigma_t \varepsilon_t \\
\sigma_t^2 = 0.0430 + 0.1552y_{t-1}^2 + 0.8653\sigma_{t-1}^2
\end{cases}
$$

With the final fitted model, we predicted the S&P Composite values from 01/2017 to 06/2017. From the plot in Part 4, we can see that the prediction is approximately same with the observed data. So, the model performs well.

In the next step, we can try more complex model such as APARCH, TGARCH and EGARCH, etc.

Reference

[1] https://www.investopedia.com/ask/answers/040215/what-does-sp-500-index-measure-and-how-it-calculated.asp

[2] https://stats.stackexchange.com/questions/202526/garch-diagnostics-autocorrelation-in-standardized-residuals-but-not-in-their-sq

[3] Modeling S&P 500 STOCK INDEX using ARMA-ASYMMETRIC POWER ARCH models, Jia Zhou, Chanli He[June 2009]