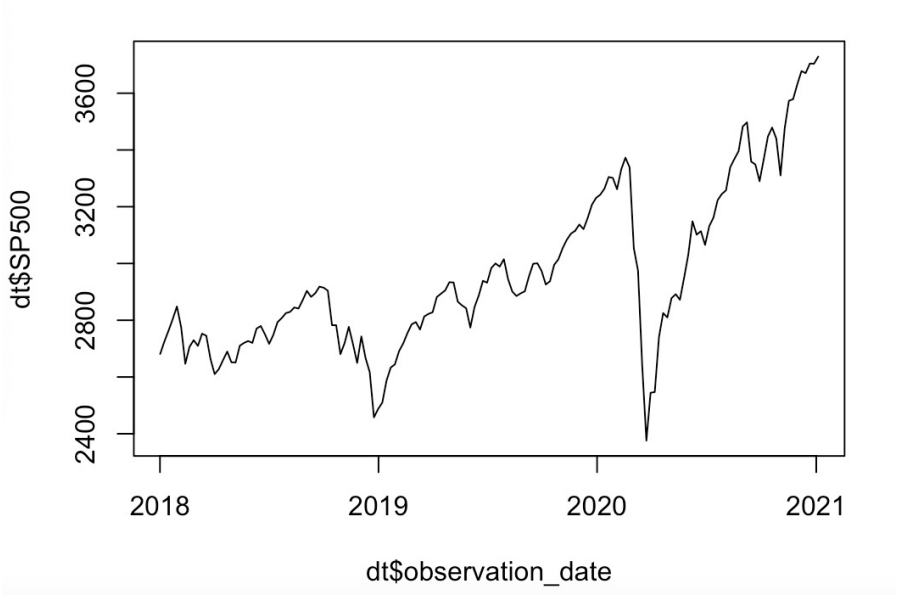
Econ 421 Final Project

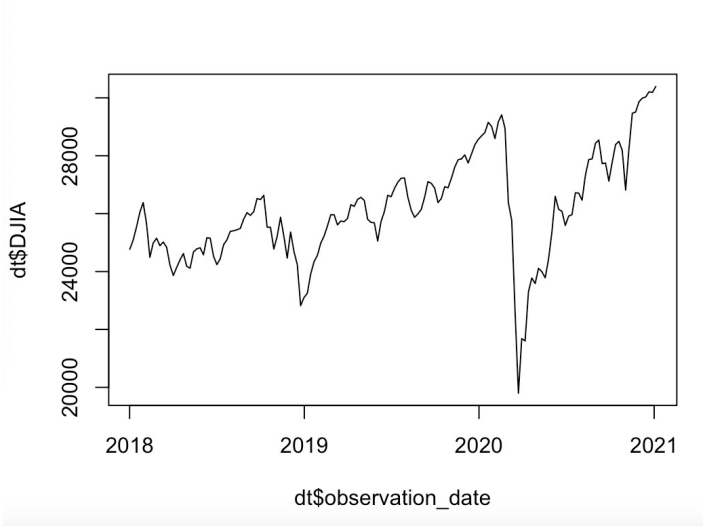
Q1：Solution

The five variables I have chosen are: SP500 index, Dow Jones Industrial Average Index, 10-year break inflation rate, Nasdaq Composite Index, CBOE Volatility Index. Four of them are stock indexes which are macroeconomically correlated with each other, and I used 10-year break inflation rate to have a glance of the whole economy. I used weekly estimation and ending with every Monday starts from 2018-01-01 to 2020-12-28. Below I will provide the plot of the data and the basic stats.

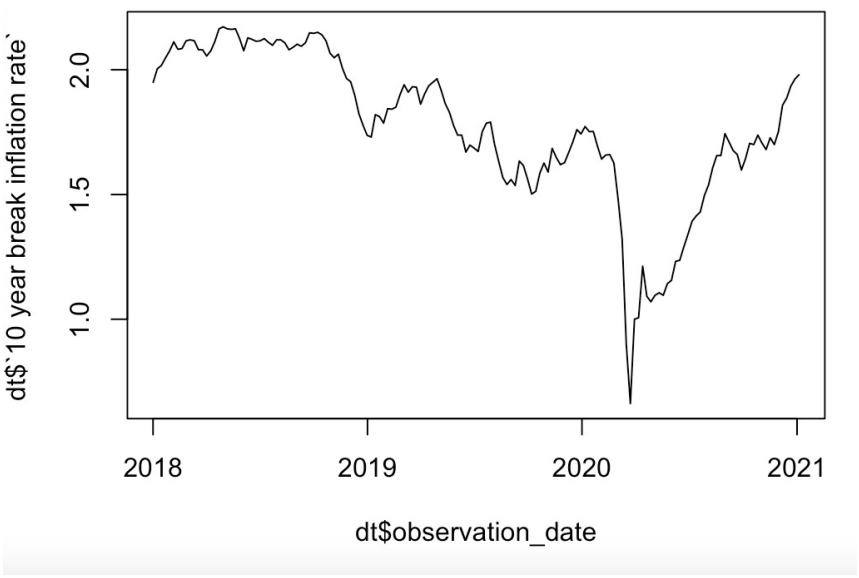
SP500:



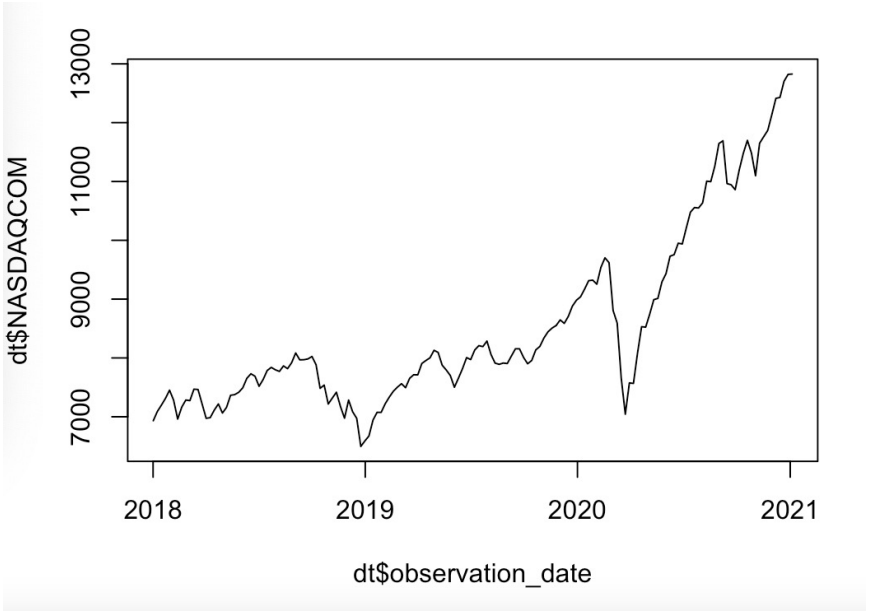
Dow Jones Industrial Average:



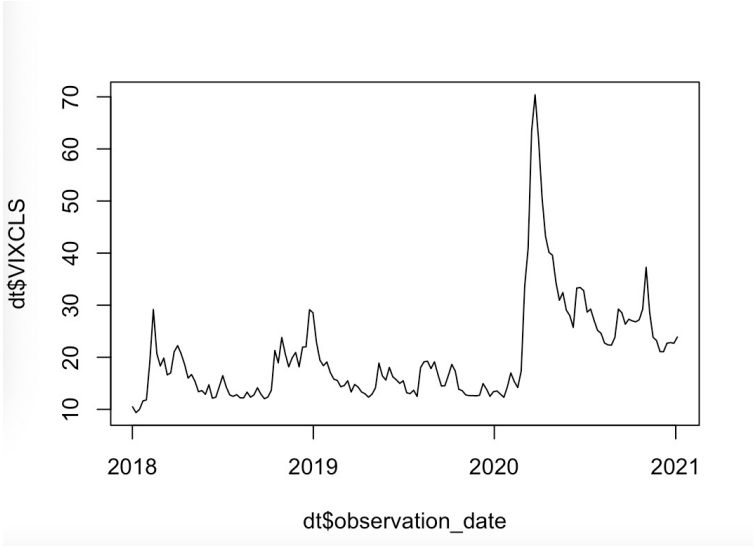
10-year break inflation rate:



Nasdaq Composite Index:



CBOE Volatility Index:



The mean, variance and covariance

Mean:

```
 observation_date                  SP500           DJIA
 Min.   :2018-01-01 00:00:00   Min.   :2376   Min.   :19798
 1st Qu.:2018-10-02 18:00:00   1st Qu.:2749   1st Qu.:24939
 Median :2019-07-04 12:00:00   Median :2893   Median :25963
 Mean   :2019-07-04 12:00:00   Mean   :2960   Mean   :26114
 3rd Qu.:2020-04-04 06:00:00   3rd Qu.:3129   3rd Qu.:27120
 Max.   :2021-01-04 00:00:00   Max.   :3729   Max.   :30394
 10 year break inflation rate   NASDAQCOM        VIXCLS
 Min.   :0.662                Min.   : 6495   Min.   : 9.376
 1st Qu.:1.633                1st Qu.: 7473   1st Qu.:13.614
 Median :1.775                Median : 7962   Median :17.352
 Mean   :1.771                Mean   : 8524   Mean   :20.354
 3rd Qu.:2.061                3rd Qu.: 9138   3rd Qu.:23.654
 Max.   :2.172                Max.   :12827   Max.   :70.398
>
```

Variance:

```
> var(dt$SP500)
[1] 84479.06
> var(dt$DJIA)
[1] 3324569
> var(dt$`10 year break inflation rate`)
[1] 0.09494098
> var(dt$NASDAQCOM)
[1] 2353218
> var(dt$VIXCLS)
[1] 95.20016
>
```

Covariance, where my main variable is chosen as SP500, so I will present the covariance between SP500 and others here.

```
> cov(dt$SP500,dt$DJIA)
[1] 481213.9
> cov(dt$SP500,dt$`10 year break inflation rate`)
[1] -15.9535
> cov(dt$SP500,dt$NASDAQCOM)
[1] 422103.8
> cov(dt$SP500,dt$VIXCLS)
[1] -41.77216
```

Q2: Solution

Since I have used 3-year dataset for this project, T is the whole dataset which contains the data from 2018-01-01 to 2020-12-28, I need to divide it into 3 samples: R, P1, P2. I would like to use the first two-year data sample as my initial forecast sample, and the rest one year will be divided into 2 6-month dataset for P1 and P2 respectively. The first 6 month of 2020 will be the sample for P1.

Q3: Solution

By constructing the regression analysis between SP500 and other variables, I got the following observations, in such case, I will regression on SP500 with DJIA, 10-year break inflation rate and Nasdaq Composite Index for my alternative autoregressive model since they give higher R^2 which means they are more realizable to be used to forecast.

```
Call:
lm(formula = dt$SP500 ~ dt$DJIA)

Residuals:
    Min      1Q  Median      3Q     Max
-152.30  -90.00  -52.21  128.60  330.40

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.202e+02  1.400e+02   -5.86 2.66e-08 ***
dt$DJIA      1.447e-01  5.347e-03   27.07  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 122.2 on 156 degrees of freedom
Multiple R-squared:  0.8245,    Adjusted R-squared:  0.8234
F-statistic: 732.9 on 1 and 156 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = dt$SP500 ~ dt$`10 year break inflation rate`)

Residuals:
    Min      1Q  Median      3Q     Max
-770.26 -184.66  -66.15  128.76  804.36

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                       3257.30     133.59  24.383   <2e-16 *
dt$`10 year break inflation rate`  -168.04      74.32  -2.261   0.0251 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 286.9 on 156 degrees of freedom
Multiple R-squared:  0.03173,   Adjusted R-squared:  0.02553
F-statistic: 5.113 on 1 and 156 DF,  p-value: 0.02514
```

```
Call:
lm(formula = dt$SP500 ~ dt$NASDAQCOM)

Residuals:
    Min      1Q  Median      3Q     Max
-318.25  -51.30   -1.80   49.13  203.32

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.431e+03  4.232e+01   33.81   <2e-16 ***
dt$NASDAQCOM 1.794e-01  4.886e-03   36.71   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 93.92 on 156 degrees of freedom
Multiple R-squared:  0.8962,    Adjusted R-squared:  0.8956
F-statistic:  1348 on 1 and 156 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = dt$SP500 ~ dt$VIXCLS)

Residuals:
    Min      1Q  Median      3Q     Max
-561.93 -211.35  -68.69  170.62  770.81

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2968.6201    53.7966  55.182   <2e-16 ***
dt$VIXCLS     -0.4388     2.3848  -0.184    0.854
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 291.6 on 156 degrees of freedom
Multiple R-squared:  0.000217,  Adjusted R-squared:  -0.006192
F-statistic: 0.03385 on 1 and 156 DF,  p-value: 0.8543
```

Using P1 to call for t-test, I got the following results:

```
        One Sample t-test

data:  dtp1$SP500
t = 55.329, df = 25, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 2887.593 3110.877
sample estimates:
mean of x
 2999.235
```

```
        One Sample t-test

data:  dtp1$DJIA
t = 47.445, df = 25, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 24518.52 26743.77
sample estimates:
mean of x
 25631.14
```

```
        One Sample t-test

data:  dtp1$`10 year break inflation rate`
t = 21.872, df = 25, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 1.184624 1.430915
sample estimates:
mean of x
 1.307769
```

```
        One Sample t-test

data:  dtp1$NASDAQCOM
t = 57.317, df = 25, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 8609.292 9251.062
sample estimates:
mean of x
 8930.177
```

```
        One Sample t-test

data:  dtp1$VIXCLS
t = 10.19, df = 25, p-value = 2.191e-10
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 25.80642 38.88081
sample estimates:
mean of x
 32.34362
```

Q4: Solution

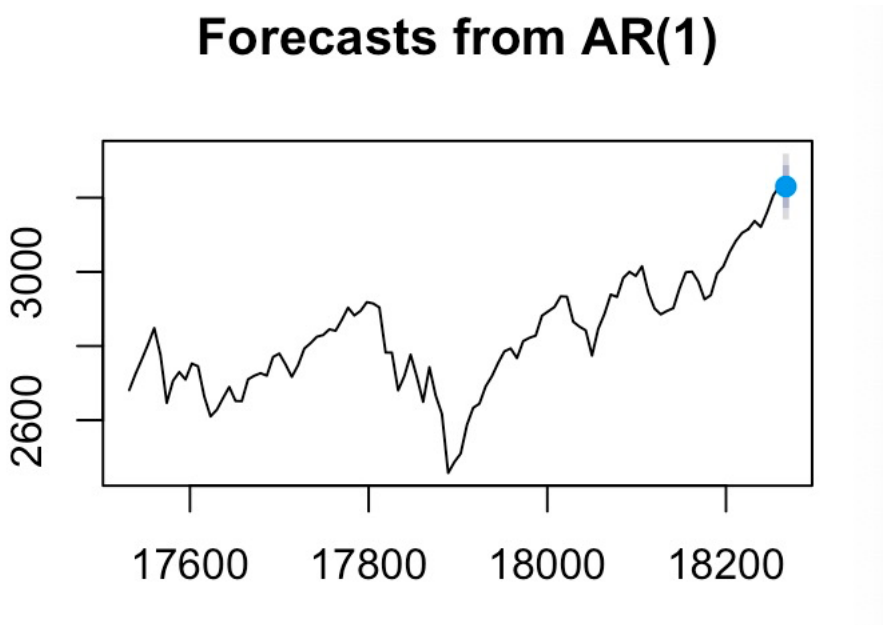
(a) Firstly, I tried to figure out the regression of the variables and to find out the most correlated variants to do the autoregressive forecast. Then I calculate the t-stats of the data sample for the P1 sample in order to figure out the correction to my stats I have to choose. Lastly, I calculate the Bayesian Information Criterion in order to determine the lags I used in the forecast model, which is 1.

(b) I have calculated the mean square forecast error at the end of each forecast model, to choose the most accurate model.

(c) Model Analysis

Model1 Benchmark Model

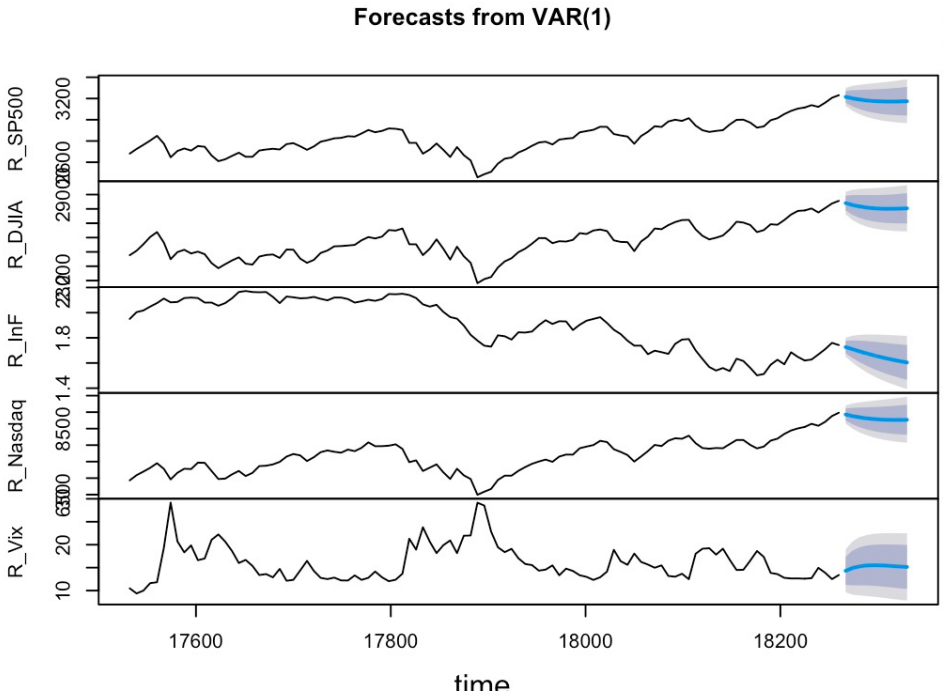
```
> forecast_model1
      Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
18267       3230.226 3172.323 3288.128 3141.672 3318.780
18274       3229.315 3148.049 3310.582 3105.029 3353.601
18281       3228.419 3129.638 3327.200 3077.347 3379.491
18288       3227.536 3114.329 3340.743 3054.401 3400.672
18295       3226.667 3101.041 3352.293 3034.538 3418.795
18302       3225.811 3089.215 3362.406 3016.905 3434.716
18309       3224.967 3078.516 3371.419 3000.989 3448.946
18316       3224.137 3068.723 3379.552 2986.451 3461.823
18323       3223.319 3059.681 3386.958 2973.056 3473.583
18330       3222.514 3051.276 3393.752 2960.629 3484.400
```




```
> mse(3242.440, 3230.226)
[1] 149.1818
```

Model2 Simple Autoregssive Model

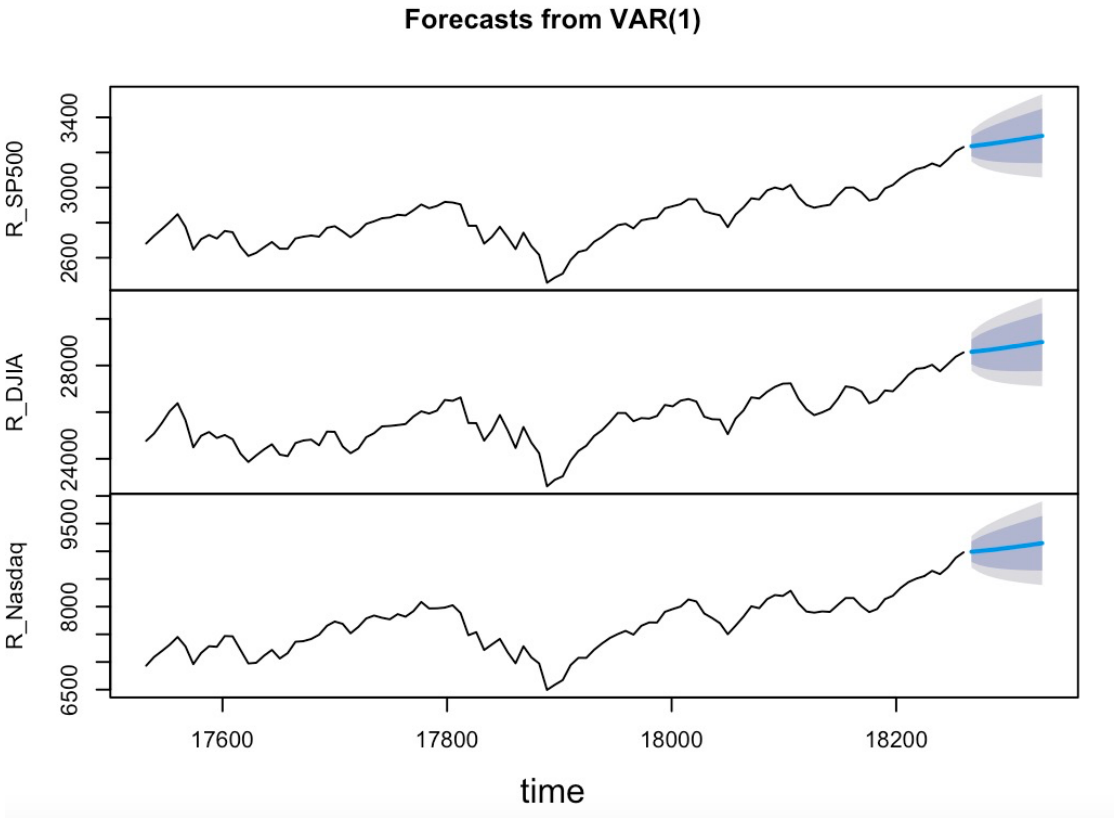
```
R_SP500
      Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
18267       3215.245 3159.539 3270.952 3130.050 3300.441
18274       3201.620 3125.272 3277.967 3084.856 3318.383
18281       3190.615 3100.305 3280.926 3052.497 3328.733
18288       3182.269 3081.635 3282.903 3028.362 3336.175
18295       3176.418 3067.722 3285.114 3010.181 3342.655
18302       3172.788 3057.487 3288.090 2996.450 3349.127
18309       3171.052 3050.100 3292.004 2986.071 3356.032
18316       3170.872 3044.909 3296.835 2978.228 3363.516
18323       3171.929 3041.403 3302.454 2972.307 3371.550
18330       3173.938 3039.185 3308.691 2967.851 3380.024
```




```
> mse(3242.440,3215.245)
[1] 739.568
>
```

Model3 Alternative Autoregressive Model

```
> forecast_result_model3_p1
$R_SP500
          fcst    lower    upper        CI
 [1,] 3235.794 3147.562 3324.025  88.23159
 [2,] 3241.108 3121.066 3361.150 120.04162
 [3,] 3246.932 3104.356 3389.507 142.57570
 [4,] 3253.141 3092.454 3413.828 160.68715
 [5,] 3259.641 3083.391 3435.892 176.25068
 [6,] 3266.360 3076.173 3456.547 190.18675
 [7,] 3273.241 3070.236 3476.246 203.00489
 [8,] 3280.243 3065.233 3495.253 215.00966
 [9,] 3287.334 3060.940 3513.728 226.39388
[10,] 3294.489 3057.204 3531.775 237.28554
```




```
> mse(3242.440,3235.794)
[1] 44.16932
>
```

Concluding from the above forecast error, we see model 3 fit the forecast well since it gives the least mean forecast error.

(d) The accuracy stats, Bayesian Information criterion is given as follow for the P2 sample.

```
> ##model1
> BIC(lm(dtp2$SP500~dtp2$DJIA))
[1] 250.6
>
> ##model 2
> BIC(lm(dtp2$SP500~dtp2$DJIA+dtp2$`10 year break inflation rate`+dtp2$NASDAQCOM+dtp
XCLS))
[1] 225.1372
>
> ##model 3
> BIC(lm(dtp2$SP500~dtp2$DJIA+dtp2$`10 year break inflation rate`+dtp2$NASDAQCOM))
[1] 221.9853
>
```

Model 3 returns the smallest BIC and we conclude the lag is using correctly.

(e) By the discussion above, I think model 3 is the most accurate forecast since the forecast return the smallest mean square forecast error. Generally, model 3 is chosen by regression with the most relevant 2 variants that we calculated at first. So those regression could be used to forecast the data and return the most accurate forecasting data.

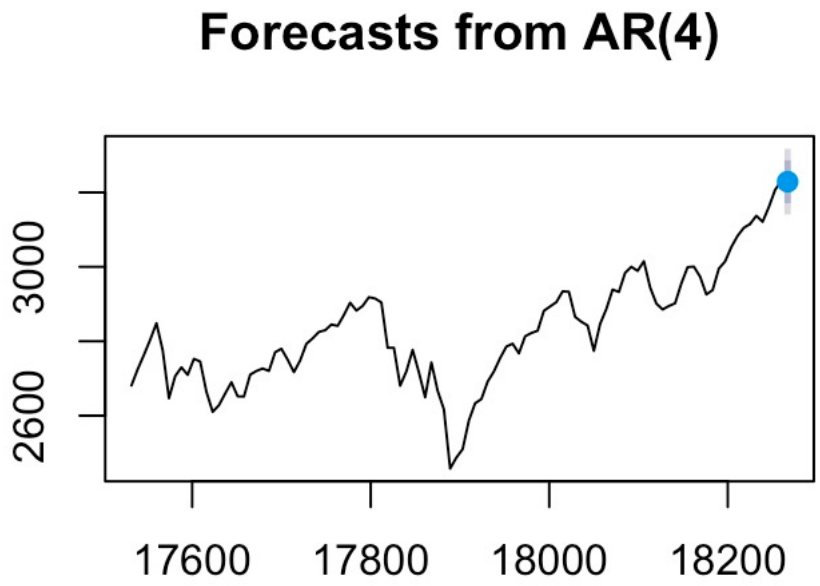
Q5: Solution

(a) Firstly, I tried to figure out the regression of the variables and to find out the most correlated variants to do the autoregressive forecast. Then I calculate the t-stats of the data sample for the P1 sample in order to figure out the correction to my stats I have to choose. Lastly, I calculate the Bayesian Information Criterion in order to determine the lags I used in the forecast model, which is 1. In the second forecast problem set, I change the forecast horizon into h2 =4 since I am using a weekly data.

(b) I am using mean square forecast error to reflect the accuracy of my forecast, and the Bayesian Information Criterion to summarize the statistics.

(c) Model 1

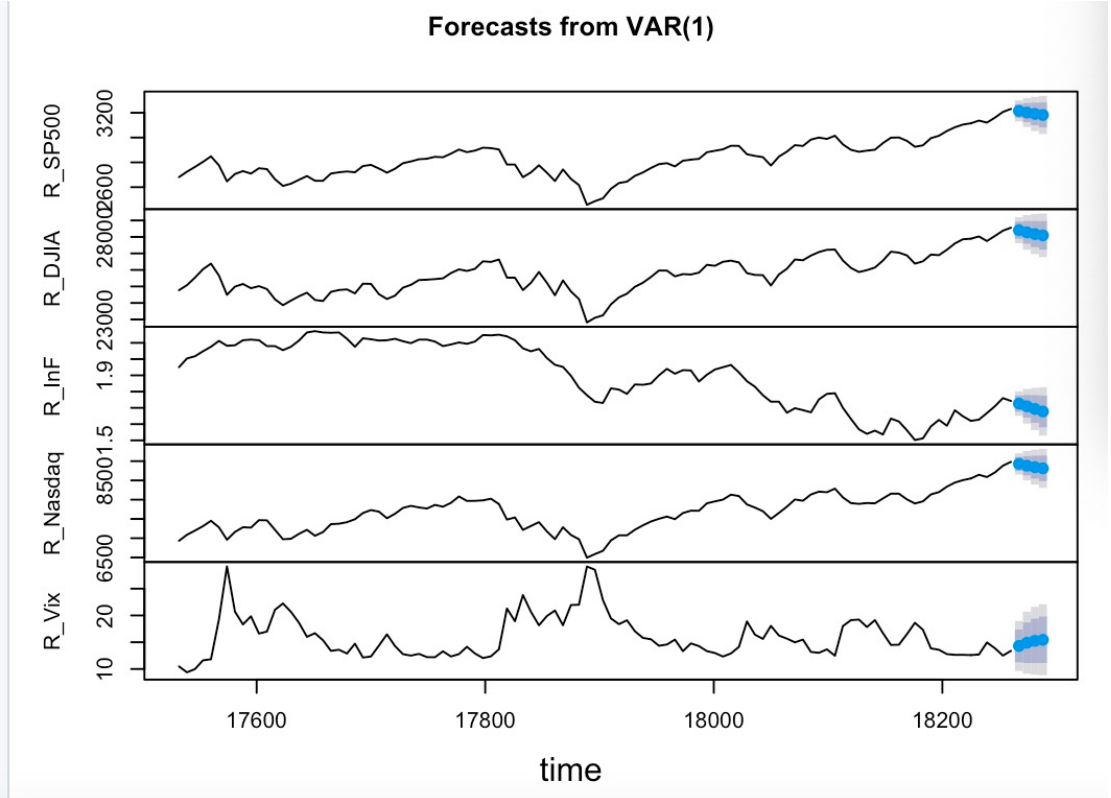
```
> forecast_model1_h2
      Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
18267       3228.815 3171.191 3286.438 3140.687 3316.942
18274       3223.775 3137.443 3310.106 3091.742 3355.808
18281       3220.841 3114.802 3326.880 3058.668 3383.013
18288       3219.712 3099.259 3340.166 3035.494 3403.930
18295       3218.870 3086.203 3351.538 3015.973 3421.768
18302       3217.883 3074.257 3361.508 2998.227 3437.539
18309       3216.783 3063.156 3370.410 2981.831 3451.735
18316       3215.676 3052.884 3378.467 2966.708 3464.644
18323       3214.595 3043.359 3385.830 2952.713 3476.477
18330       3213.537 3034.472 3392.603 2939.680 3487.394
```




```
> mse(valueh2_actual,(3228.815+3223.775+3220.841+3219.712)/4)
[1] 3001.204
>
```

Model2

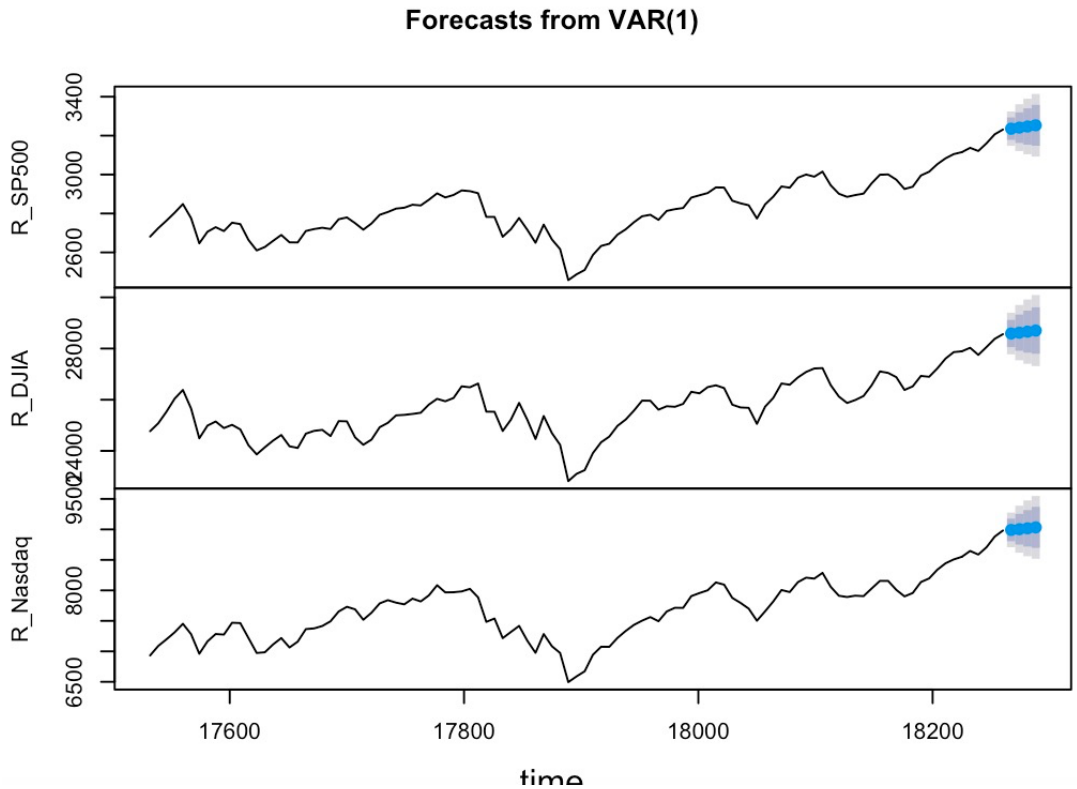
```
> forecast_model2_p1_h2
R_SP500
      Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
18267       3215.245 3159.539 3270.952 3130.050 3300.441
18274       3201.620 3125.272 3277.967 3084.856 3318.383
18281       3190.615 3100.305 3280.926 3052.497 3328.733
18288       3182.269 3081.635 3282.903 3028.362 3336.175
```




```
> mse(valueh2_actual,(3215.245+3201.620+3190.615+3182.269)/4)
[1] 6501.479
>
```

Model 3

```
R_SP500
      Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
18267       3235.794 3178.102 3293.485 3147.562 3324.025
18274       3241.108 3162.617 3319.599 3121.066 3361.150
18281       3246.932 3153.706 3340.157 3104.356 3389.507
18288       3253.141 3148.073 3358.209 3092.454 3413.828
```




```
> mse(valueh2_actual,(3235.794+3241.108+3246.932+3253.141)/4)
[1] 1144.148
```

(d) Bayesian Information Criterion is concluded as follow:

```
> plot(Forecast_result_model3_p1_h2)
> ##model1
> BIC(lm(dtp2$SP500~dtp2$DJIA))
[1] 250.6
>
> ##model 2
> BIC(lm(dtp2$SP500~dtp2$DJIA+dtp2$`10 year break inflation rate`+
IXCLS))
[1] 225.1372
>
> ##model 3
> BIC(lm(dtp2$SP500~dtp2$DJIA+dtp2$`10 year break inflation rate`+
[1] 221.9853
```

(e) By the argument statistic and forecast error listed above, model 3 also gives the most accurate forecasting data. But generally, all of the 3 model increased their forecast error which means the forecast became less accurate as the forecast horizon getting larger. This might because the insight of our forecast cannot support such far future forecast, and could be more useful when doing forecast for

only h = 1.

(f) The best model is model 3 as well by the argument from (e).

(g) I think the model selected might be correlated with the forecast horizon. Since when we are estimating the forecast error, we will take more weight on each of the forecast we choose into consideration. Specifically, with larger horizons, we will get more actual data into consideration which will return a more accurate forecast error instead of only the most recent horizon. The influence of the horizon is pretty low since there are potential correlation with the forecast and the actual data, which means if we get a really closed h = 1 forecast, for h = 12 forecast we may still get closed forecasting data to the actual value. There might be some marginal correlation which leads to the change of model selection. To sum up, model selection is dependent by the variant and horizon, with more correlated variant, the forecast error is lower.

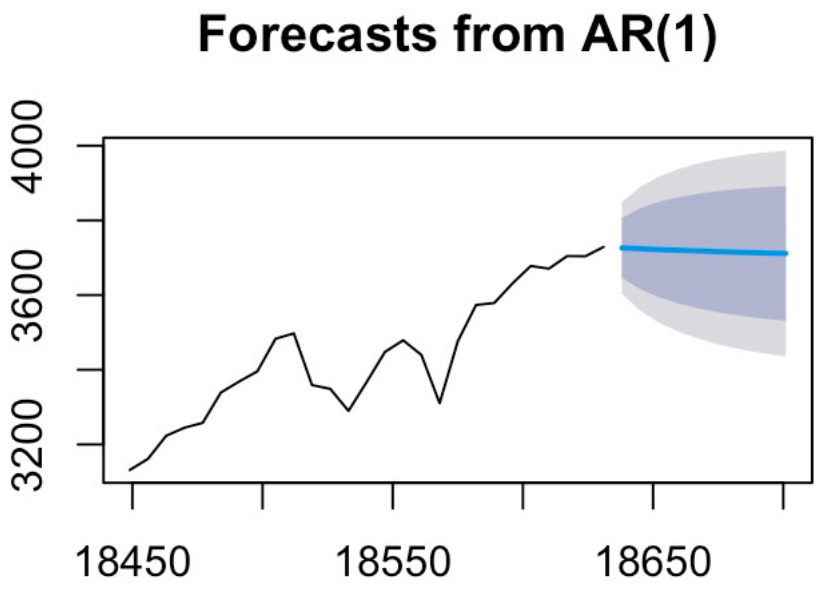
Q6 Using P2 as data sample to forecast

(a) Firstly, I tried to figure out the regression of the variables and to find out the most correlated variants to do the autoregressive forecast. Then I calculate the t-stats of the data sample for the P1 sample in order to figure out the correction to my stats I have to choose. Lastly, I calculate the Bayesian Information Criterion in order to determine the lags I used in the forecast model, which is 1. In the second forecast problem set, I change the forecast horizon into h2 =4 since I am using a weekly data. Here we doubled check if our forecast is optimal for the later time assumption and take the P2 sample into consideration.

(b) I am using mean square forecast error to reflect the accuracy of my forecast, and the Bayesian Information Criterion to summarize the statistics.

(c) Model 1 h = 1

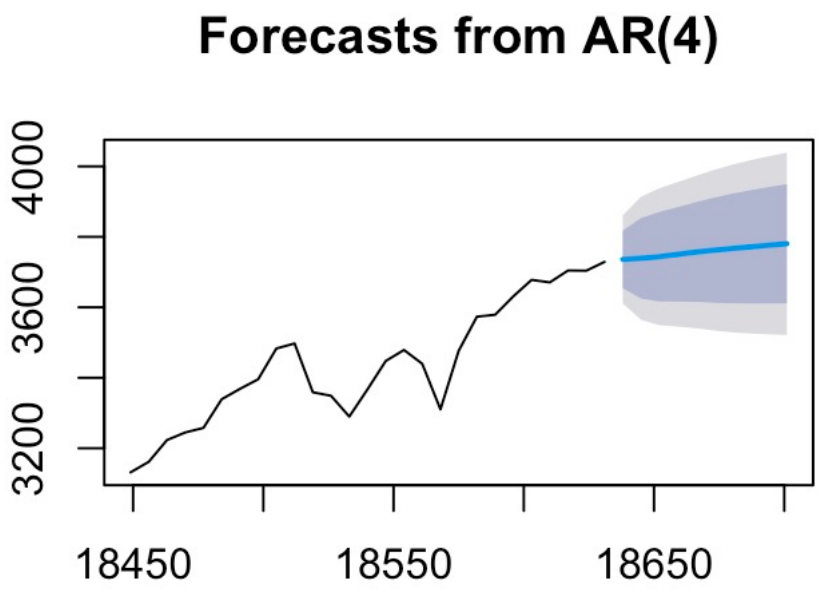
```
> forecast_model1_h1_P2
      Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
18638       3726.421 3646.611 3806.232 3604.362 3848.481
```




```
> mse(3780.62,3726.421)
[1] 2937.532
```

Model 1 with h = h2 =4

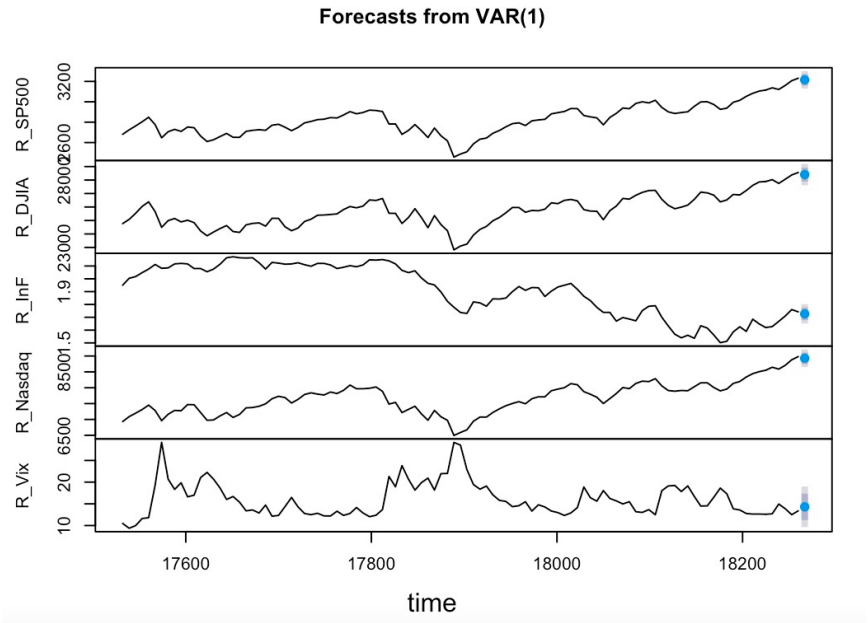
```
> forecast_model1_h2_P2
      Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
18638       3736.068 3654.603 3817.532 3611.479 3860.657
18645       3739.185 3625.044 3853.327 3564.621 3913.750
18652       3743.514 3616.587 3870.442 3549.395 3937.634
18659       3750.254 3616.299 3884.209 3545.388 3955.120
```




```
> mse(valueactual_p2,(3736.068+3739.185+3743.514+3750.254)/4)
[1] 3041.771
>
```

Model 2 when h = h1 = 1

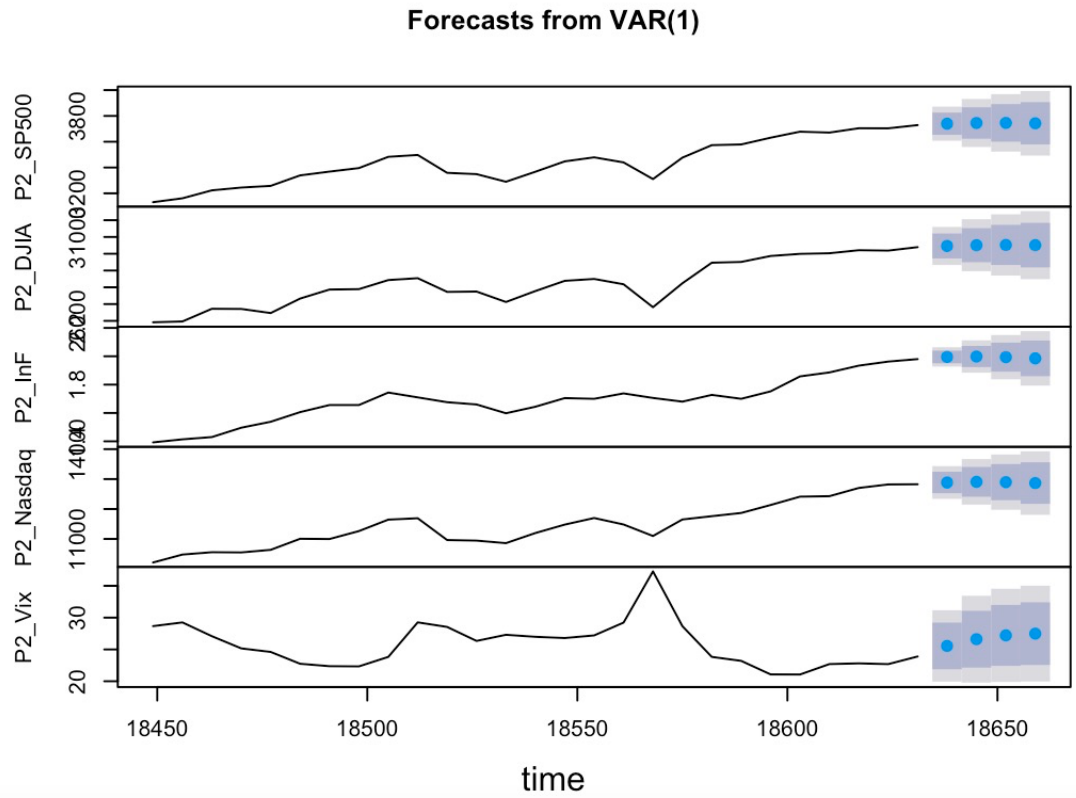
```
P2_SP500
      Point Forecast    Lo 80    Hi 80    Lo 95   Hi 95
18638       3739.607 3653.224 3825.991 3607.495 3871.72
```



```
> mse(3780.62,3739.607)
[1] 1682.066
>
```

Model 2 when h = h2 =4

```
P2_SP500
      Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
18638       3739.607 3653.224 3825.991 3607.495 3871.720
18645       3745.214 3623.628 3866.800 3559.264 3931.164
18652       3746.014 3600.402 3891.626 3523.320 3968.708
18659       3742.838 3579.052 3906.625 3492.348 3993.329
```

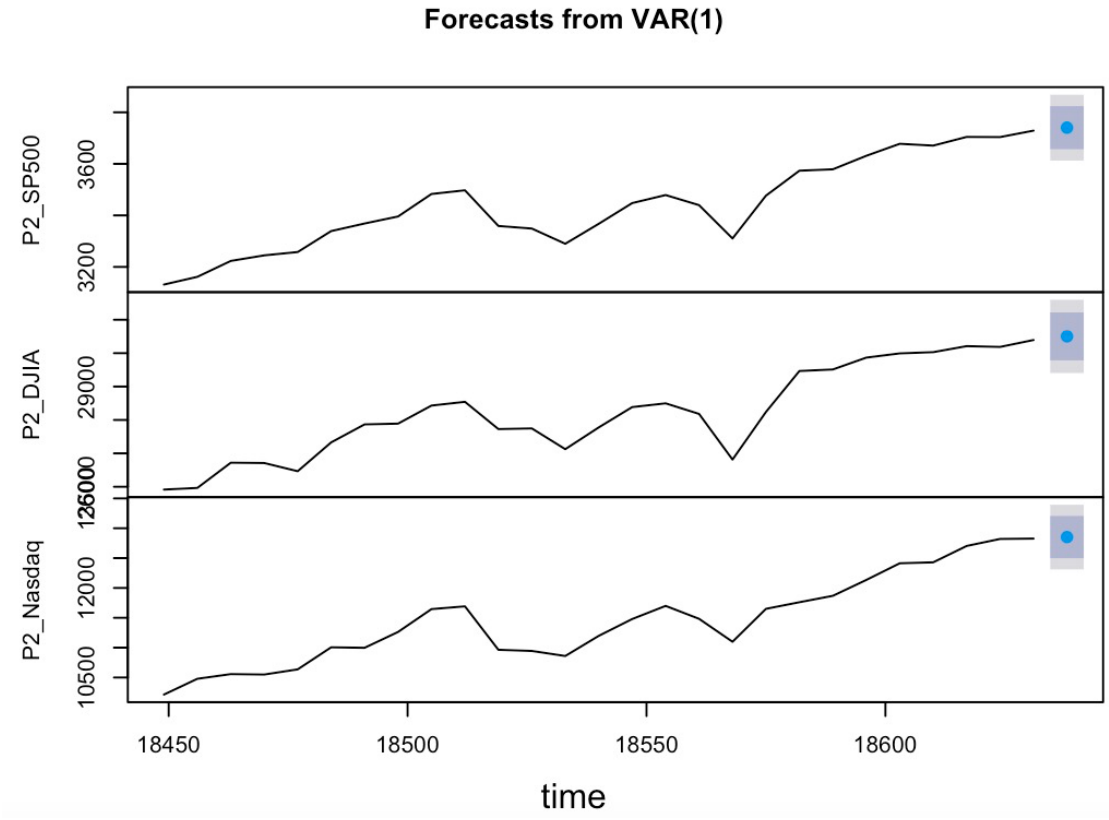



```
> mse(valueactual_p2,(3739.607+3745.214+3746.014+3742.838)/4)
[1] 2914.839
>
```

Model3 when h = h1 = 1

```
> Forecast_result_model3_p2
P2_SP500
      Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
18638        3740.45 3656.892 3824.009 3612.658 3868.242
```

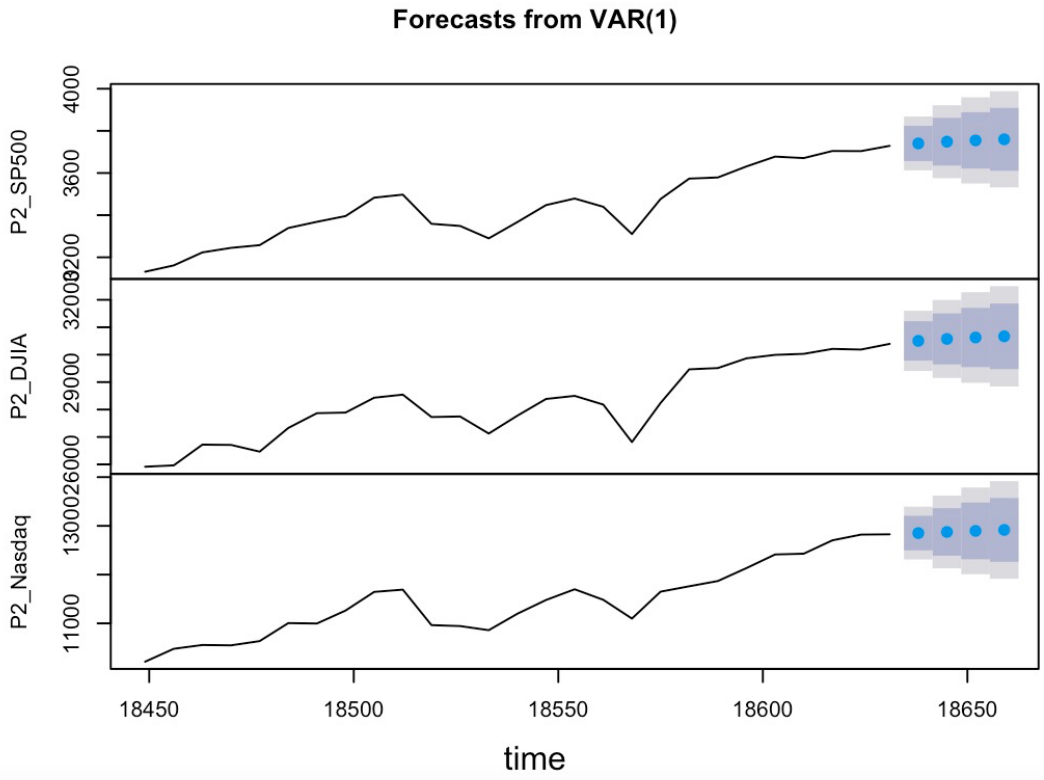
```
> mse(3780.62,3740.45)
[1] 1613.629
```



Model 3 when h = h2 = 4

```
P2_SP500
      Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
18638       3740.450 3656.892 3824.009 3612.658 3868.242
18645       3748.522 3635.405 3861.639 3575.525 3921.520
18652       3754.759 3621.103 3888.415 3550.350 3959.169
18659       3759.936 3610.678 3909.194 3531.665 3988.207
```

```
> mse(valueactual_p2,(3740.450+3748.522+3754.759+3759.936)/4)
[1] 2161.39
>
```



(d) The forecast accuracy statistic, BIC is listed as below:

```
[1] 1613.629
> BIC(lm(dtp2$SP500~dtp2$DJIA))
[1] 250.6
>
> ##model 2
> BIC(lm(dtp2$SP500~dtp2$DJIA+dtp2$`10 year break inflation rate`+dtp2$NASDAQCOM+dtp2
IXCLS))
[1] 225.1372
>
> ##model 3
> BIC(lm(dtp2$SP500~dtp2$DJIA+dtp2$`10 year break inflation rate`+dtp2$NASDAQCOM))
[1] 221.9853
>
```

(e) Based on the BIC and the mean square forecast error, model 3 is also the best model I will choose to make forecast more correct and reliable. Since it has the lowest forecasting error and the most accurate forecasting statistic among all three models.

(f) When we were considering h = h2, we will also choose model 3 as the best model although all 3 will return higher mean square forecast error as the horizon goes up. Model 3 still return the lowest MSE, so we assume it's the most accurate one.

(g) We get the same result as we use R sample to do forecast as we used P2 sample, we could see from the mean square forecast error that model 2 became better relatively to it was in the previous sample. Since the data we choose is economically related to each other so we will get different forecast accuracy in the different sample period. But overall, model 3 which has the most related data will return the best forecast.