PROBLEM STATEMENT

This dataset contains weekly data for the Dow Jones Industrial Index. It has been used in computational investing research. In this dataset, each record (row) is data for a week. Each record also has the percentage of return that stock has in the following week (percent_change_next_weeks_price).

Ideally, this could be used to determine which stock will produce the greatest rate of return in the following week.

We want to know:

- Which variables are significant in predicting the percent change in price of a stock for following week.
- How well those variables describe the percent change in price of a stock.

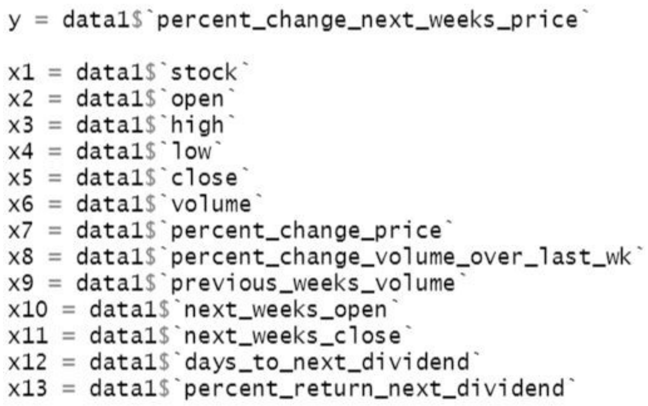
GIVEN DATA

The dataset contains 14 columns with the "percent change next week's price" column, taken as response variable and 13 predictor variables for n = 720 stocks. The data dictionary for each of the variables can be found here:

#put data description link

DEFINING VARIABLES

Let "percent_change_next_weeks_price" of the stock be the response variable Y and predictor variables are as mentioned in the data dictionary. These predictors (as mentioned in the data dictionary) are respectively denoted by X_1, X_2,..., X_{13}. We observe data {(y_i, x_1, x_2,..., x_{13}) : 1≤ i ≤ 720}.



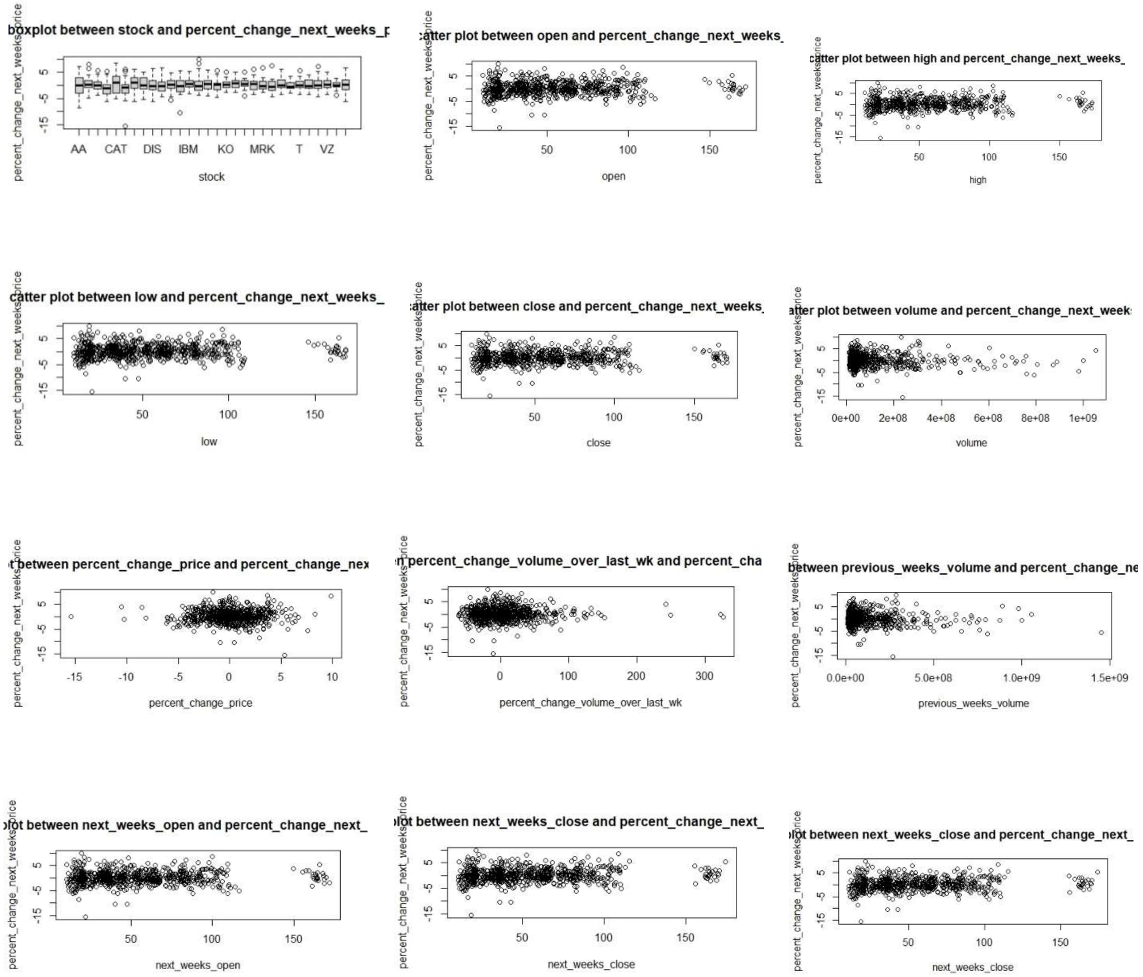
```
y = data1$`percent_change_next_weeks_price`

x1 = data1$`stock`
x2 = data1$`open`
x3 = data1$`high`
x4 = data1$`low`
x5 = data1$`close`
x6 = data1$`volume`
x7 = data1$`percent_change_price`
x8 = data1$`percent_change_volume_over_last_wk`
x9 = data1$`previous_weeks_volume`
x10 = data1$`next_weeks_open`
x11 = data1$`next_weeks_close`
x12 = data1$`days_to_next_dividend`
x13 = data1$`percent_return_next_dividend`
```

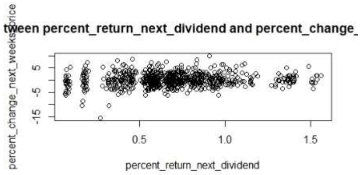
NOTATION

Denote the observed response vector y = (y_1, y_2,.., y_{720})’, the observed vector $\boldsymbol{X_j}$= ($x_{1j}, x_{2j}, \ldots, x_{720j}$)’ of the j-th predictor, $\mathbf{1_n}$ is the vector of length n with all entries 1, and the design matrix X = [$\mathbf{1_n}\ \mathbf{X}_1, \mathbf{X}_2, .., \mathbf{X}_{13}$]. β = ($\beta_0$, β_1,..., β_{13})’ is the 14 x 1 unknown parameter vector of the model. E= (ε_1, ε_2 ,..., ε_{720})’ is the 720 x 1 vector corresponding to the random error. Also let $\mathbf{0}_n$ be the null vector of length n and $\boldsymbol{I_n}$ be the identity matrix of size n.

Scatter Plots:

From the data, the pairwise scatter plots of the response and predictors are found out and visualised:





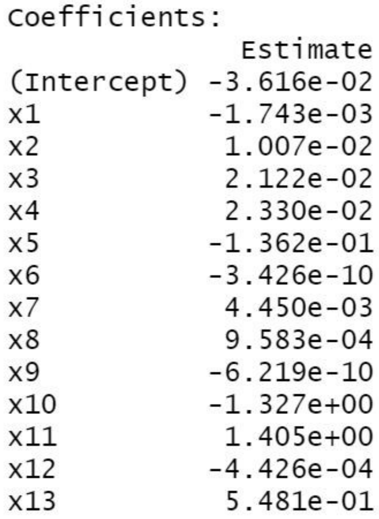
From the scatter plots, we can assume a linear relationship visible between percent_change_price and percent_change_next_weeks_price. For rest of the predictor variables, judging by scatter plots, there seems no/little linearity with the response variable.

Multiple linear regression model: Multiple linear regression model is given by

$$Y=X\ \beta + E$$

where E ~ $N_n(0_n, \sigma^2 I_n)$ and the predictor variables $x_1, x_2, \ldots, x_p$ are assumed to be non-stochastic.

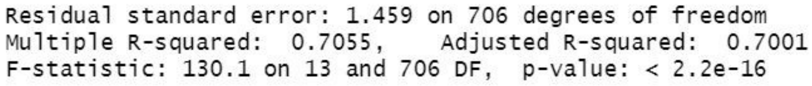
Fitting: Here we use least squares estimator of β which is given by $\widehat{\beta}$ = $(X'X)^{-1}X'y$. For the given data, it turns out to be as shown in the table below.



```
Coefficients:
              Estimate
(Intercept) -3.616e-02
x1          -1.743e-03
x2           1.007e-02
x3           2.122e-02
x4           2.330e-02
x5          -1.362e-01
x6          -3.426e-10
x7           4.450e-03
x8           9.583e-04
x9          -6.219e-10
x10         -1.327e+00
x11          1.405e+00
x12         -4.426e-04
x13          5.481e-01
```

And the fitted multiple linear regression model is: $\hat{y} = \boldsymbol{X}\widehat{\beta}$

R^2 and Adjusted R^2: For the given data and model, the value of R^2 and adjusted R^2 turn out to be 0.7055 and 0.7001 respectively. Therefore, 70.55% of the total variation of the response variable is explained by the above least-squares fitted multiple linear regression model.



```
Residual standard error: 1.459 on 706 degrees of freedom
Multiple R-squared:  0.7055,    Adjusted R-squared:  0.7001
F-statistic: 130.1 on 13 and 706 DF,  p-value: < 2.2e-16
```

Checking Significance of all Predictor variables: Here we wish to test H_0: $\beta_i = 0$ for all i = 0, 1, 2,…, 13 against H_1: $\beta_i \neq 0$ for at least one i.

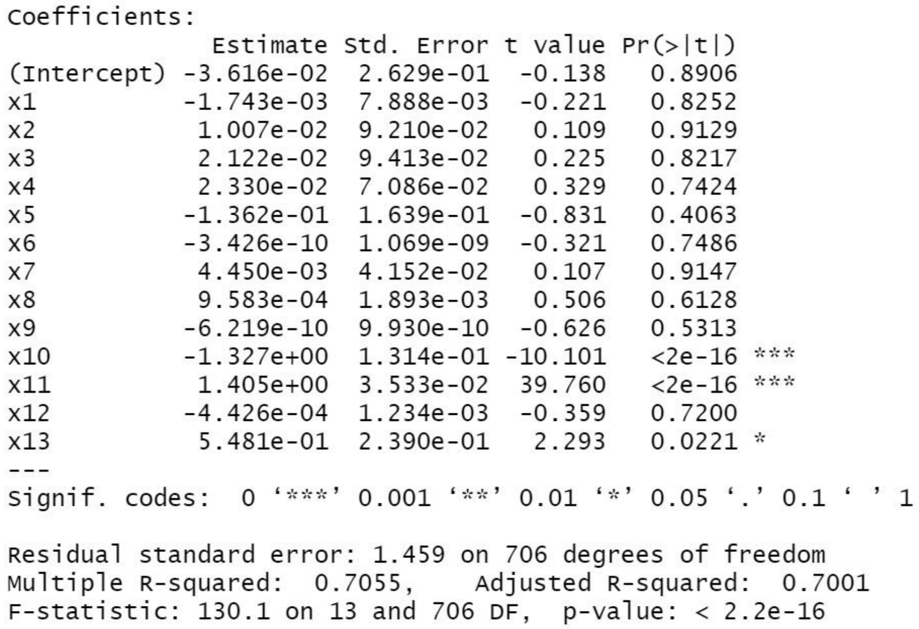
Test statistic: $F_1 = MS_R/MSE$

Since P($F_{p,n-p-1} > (F_1)_{observed}$) = p-value: < 2.2e-16, we reject the null hypothesis at 5% level of significance, i.e., at least one of the $\beta_i \neq 0$.

We now move to checking significance of individual regression coefficients.

H_0: $\beta_j = 0$ against H_1: $\beta_j \neq 0$; for all j

Test statistic: $T_j = \frac{\widehat{\beta_j}-0}{\sqrt{MSE.C_{jj}}}$ where $C = ((C_{ij})) = (X'X)^{-1}$



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -3.616e-02  2.629e-01  -0.138   0.8906    
x1          -1.743e-03  7.888e-03  -0.221   0.8252    
x2           1.007e-02  9.210e-02   0.109   0.9129    
x3           2.122e-02  9.413e-02   0.225   0.8217    
x4           2.330e-02  7.086e-02   0.329   0.7424    
x5          -1.362e-01  1.639e-01  -0.831   0.4063    
x6          -3.426e-10  1.069e-09  -0.321   0.7486    
x7           4.450e-03  4.152e-02   0.107   0.9147    
x8           9.583e-04  1.893e-03   0.506   0.6128    
x9          -6.219e-10  9.930e-10  -0.626   0.5313    
x10         -1.327e+00  1.314e-01 -10.101   <2e-16 ***
x11          1.405e+00  3.533e-02  39.760   <2e-16 ***
x12         -4.426e-04  1.234e-03  -0.359   0.7200    
x13          5.481e-01  2.390e-01   2.293   0.0221 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.459 on 706 degrees of freedom
Multiple R-squared:  0.7055,	Adjusted R-squared:  0.7001 
F-statistic: 130.1 on 13 and 706 DF,  p-value: < 2.2e-16
```

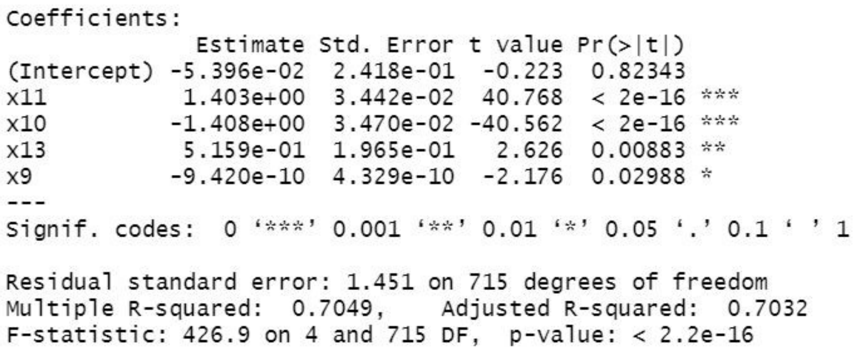
We can observe from the table given above that the variables ‘next_weeks_open’, ‘next_weeks_close’ and ‘percent_return_next_dividend’ are significant at 5% level of significance.

Since all predictor variables in the model are not significant, we need to run variable selection algorithm to get a reasonable set of significant predictors.

Variable Selection Algorithm:

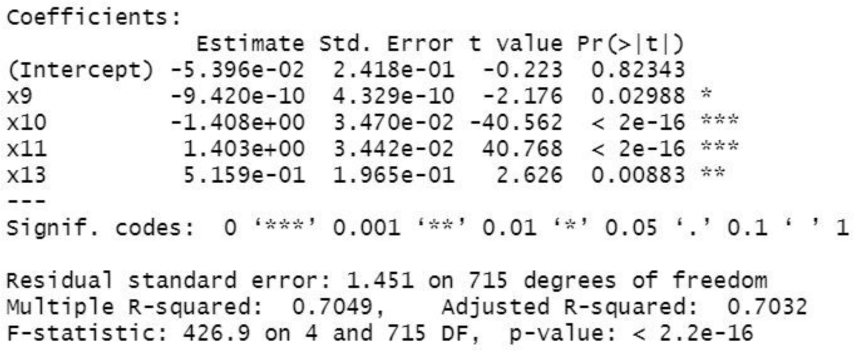
We apply all three variable selection algorithms which are:

1. Forward selection method



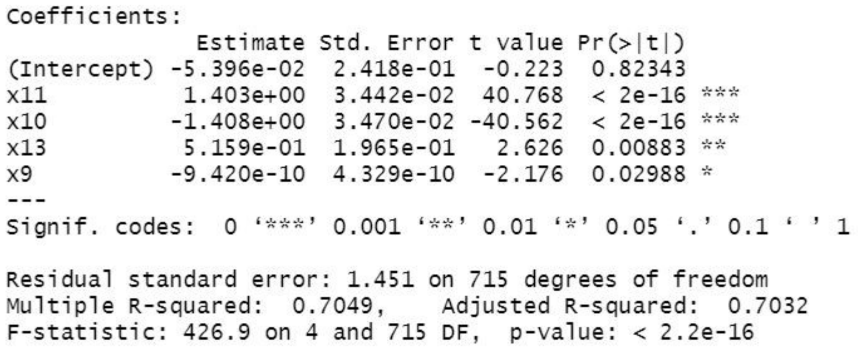
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -5.396e-02  2.418e-01  -0.223  0.82343    
x11          1.403e+00  3.442e-02  40.768  < 2e-16 ***
x10         -1.408e+00  3.470e-02 -40.562  < 2e-16 ***
x13          5.159e-01  1.965e-01   2.626  0.00883 ** 
x9          -9.420e-10  4.329e-10  -2.176  0.02988 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.451 on 715 degrees of freedom
Multiple R-squared:  0.7049,	Adjusted R-squared:  0.7032 
F-statistic: 426.9 on 4 and 715 DF,  p-value: < 2.2e-16
```

2. Backward elimination method



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -5.396e-02  2.418e-01  -0.223  0.82343    
x9          -9.420e-10  4.329e-10  -2.176  0.02988 *  
x10         -1.408e+00  3.470e-02 -40.562  < 2e-16 ***
x11          1.403e+00  3.442e-02  40.768  < 2e-16 ***
x13          5.159e-01  1.965e-01   2.626  0.00883 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.451 on 715 degrees of freedom
Multiple R-squared:  0.7049,	Adjusted R-squared:  0.7032 
F-statistic: 426.9 on 4 and 715 DF,  p-value: < 2.2e-16
```

3. Step-wise selection method



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -5.396e-02  2.418e-01  -0.223  0.82343    
x11          1.403e+00  3.442e-02  40.768  < 2e-16 ***
x10         -1.408e+00  3.470e-02 -40.562  < 2e-16 ***
x13          5.159e-01  1.965e-01   2.626  0.00883 ** 
x9          -9.420e-10  4.329e-10  -2.176  0.02988 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.451 on 715 degrees of freedom
Multiple R-squared:  0.7049,	Adjusted R-squared:  0.7032 
F-statistic: 426.9 on 4 and 715 DF,  p-value: < 2.2e-16
```

After applying all above-mentioned algorithms and fitting the models, we observe that both forward selection method, backward elimination and Step-wise selection method give the maximum value of R^2 and adjusted R^2 which are 0.7049 and 0.7032 respectively.

All of these models retain the same set of predictors at their completion as can be seen from the output.

This means that either of these three can be used to get a reasonable set of significant predictors.

We drop the variables ‘stock’, ‘open’, ’high’, ’low’, ‘close’, ‘volume’, ‘percent change price’, ‘percent change volume over last week’, ‘days to next dividend’ from our data after applying variable selection algorithms and update the corresponding design matrix.

Check for presence of Multicollinearity in the updated model:

First, we apply centring and scaling on our updated data set. Now we again fit the MLRM to the centred and scaled data set. As the R^2 and adjusted R^2 doesn’t change for the centred and scaled dataset this proves that standardization has no impact on the MLRM.

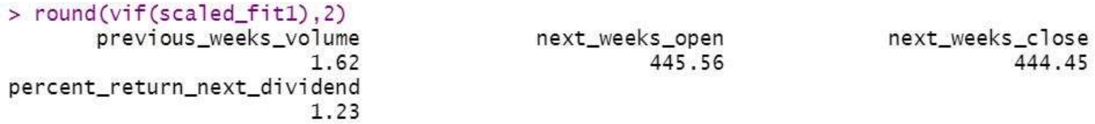
Now we compute determinant of X_{cs}’X_{cs} (centred and scaled design matrix) and observe that it is very close to 0. So, we suspect presence of multicollinearity. Also, correlation matrix gives high correlation among various predictor variables.

Now we compute Variance Inflation Factor for our scaled fit model and use 10 as our cut off.

$$VIF_j = \left(1 - R_{x_j}^2\right)^{-1};$$

where R_(x_j) is the coefficient of determination obtained when X_j is linearly regressed on the remaining p-1 predictor variables.

VIFs are as follows:



```
> round(vif(scaled_fit1),2)
       previous_weeks_volume              next_weeks_open             next_weeks_close
                        1.62                       445.56                       444.45
percent_return_next_dividend
                        1.23
```

From the results we obtain that the regression coefficients corresponding to predictor variables ‘next weeks open’ and ‘next weeks close’ are poorly estimated because of multi-collinearity.

We now compute Condition Indices with cut off as 50.

$$C_j = \frac{\lambda_1(R)}{\lambda_j(R)}, j = 1, 2, \dots, p$$

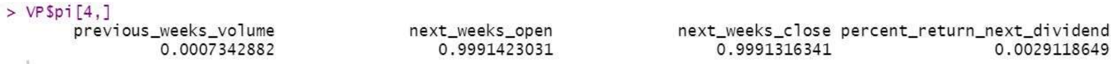
The condition indices are (1, 1.99, 5.58, -38992.74}.

We observe that last condition index exceeds our pre-set cut off of 50. Therefore, we suspect that last principal component of the predictor variables may be responsible for multicollinearity.

Measures Based on Variance Decomposition

$$p_{jk} = \frac{\frac{v_{jk,R}^2}{\lambda_k(R)}}{\sum_{k=1}^{p} \frac{v_{jk,R}^2}{\lambda_k(R)}}$$

p_{jk} indicates the proportion of the contribution of kth principal component on the variance of $\widehat{\beta_{J,cs}}$



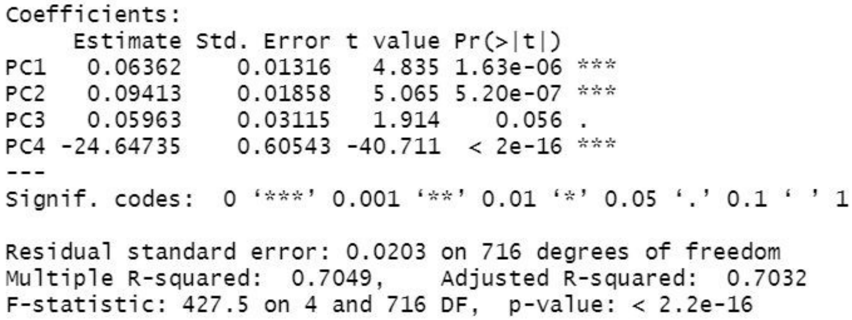
```
> VP$pi[4,]
      previous_weeks_volume               next_weeks_open              next_weeks_close percent_return_next_dividend
               0.0007342882                  0.9991423031                  0.9991316341                 0.0029118649
```

On computing these measures, we make the following observations:

Regression coefficient corresponding to

- 'next weeks open' is suffering from the 1st principal component. The measure is 0.9914.
- 'next weeks close' is suffering because of the 1st principal component. The measure is 0.9913

Principal Component Analysis:

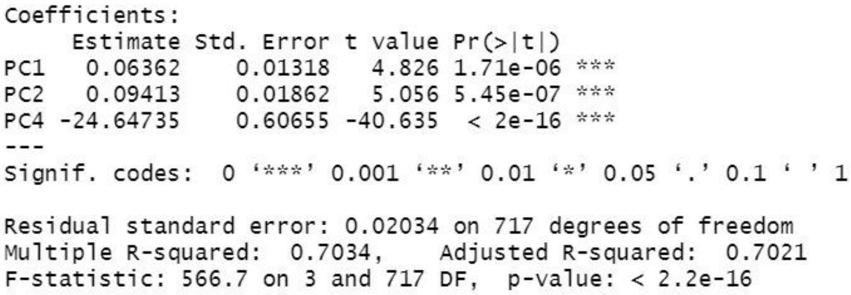


```
Coefficients:
     Estimate Std. Error t value Pr(>|t|)    
PC1   0.06362    0.01316   4.835 1.63e-06 ***
PC2   0.09413    0.01858   5.065 5.20e-07 ***
PC3   0.05963    0.03115   1.914    0.056 .  
PC4 -24.64735    0.60543 -40.711  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0203 on 716 degrees of freedom
Multiple R-squared:  0.7049,	Adjusted R-squared:  0.7032 
F-statistic: 427.5 on 4 and 716 DF,  p-value: < 2.2e-16
```

We fit PCA model to the given data to work around the multi-collinearity detected above.

On fitting the model, we observe that only PC1, PC2, and PC4 are significant.

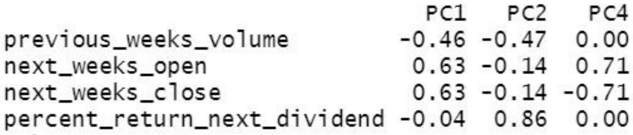
Hence, we again apply PCA but only with the selected Principal Components.



```
Coefficients:
     Estimate Std. Error t value Pr(>|t|)    
PC1   0.06362    0.01318   4.826 1.71e-06 ***
PC2   0.09413    0.01862   5.056 5.45e-07 ***
PC4 -24.64735    0.60655 -40.635  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02034 on 717 degrees of freedom
Multiple R-squared:  0.7034,	Adjusted R-squared:  0.7021 
F-statistic: 566.7 on 3 and 717 DF,  p-value: < 2.2e-16
```

From above, we achieve an approximately same adjusted R^2 of 0.7021 compared to 0.7032 (previously when all PCs were used). In this case, $R^2 = 0.7034$, i.e., 70.34% of the total variation of the response variable is explained by the PCA fitted multiple linear regression model.

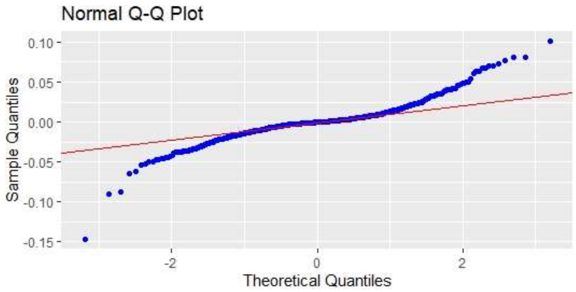
The linear combination of each principal component is as shown:



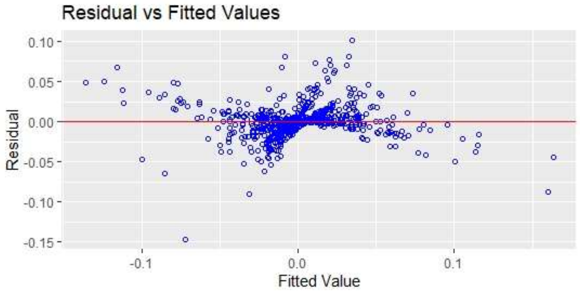
```
                              PC1   PC2   PC4
previous_weeks_volume        -0.46 -0.47  0.00
next_weeks_open               0.63 -0.14  0.71
next_weeks_close              0.63 -0.14 -0.71
percent_return_next_dividend -0.04  0.86  0.00
```

Verification of assumptions on the random errors:

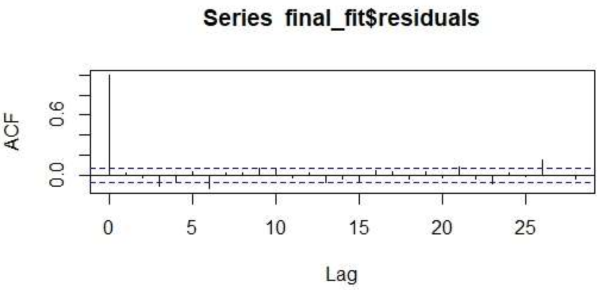
1. Normality assumption
   We perform normality tests to test the null hypothesis H_0: Error terms are normally distributed, against, H_1: Error terms are not normally distributed and we have enough evidence to reject the null hypothesis and hence normality assumption is violated.



2. Homoscedasticity assumption
   The best way to detect homoscedasticity is to construct residual vs fitted value plot and look for patterns. As clustering is clearly observed in the residual plot, homoscedasticity assumption doesn't hold.

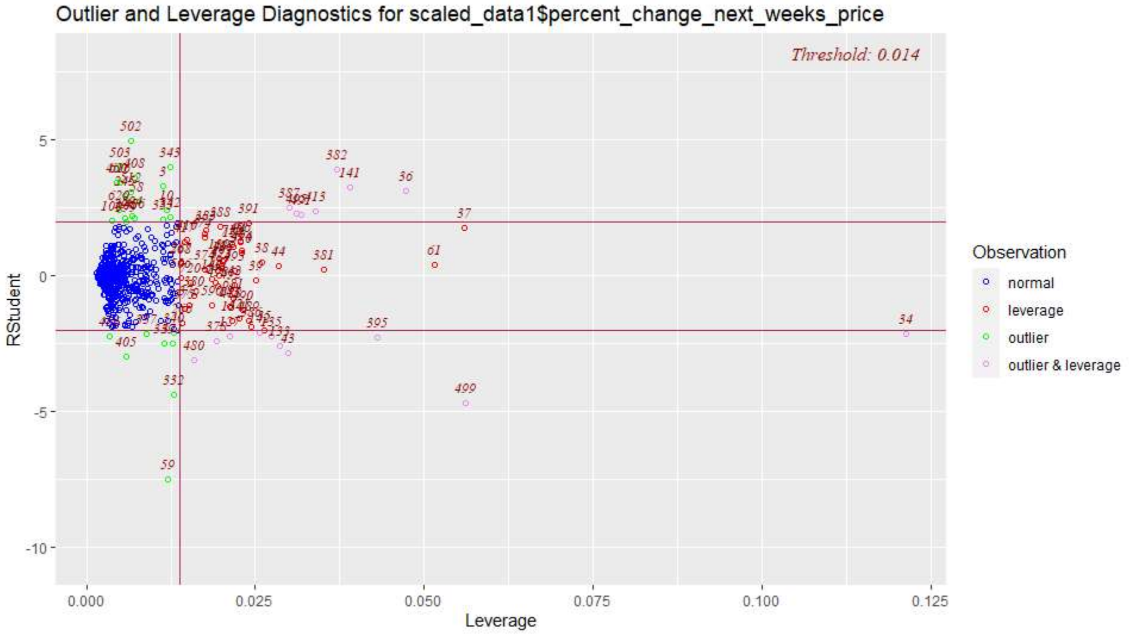


3. Assumption of Random errors being correlated

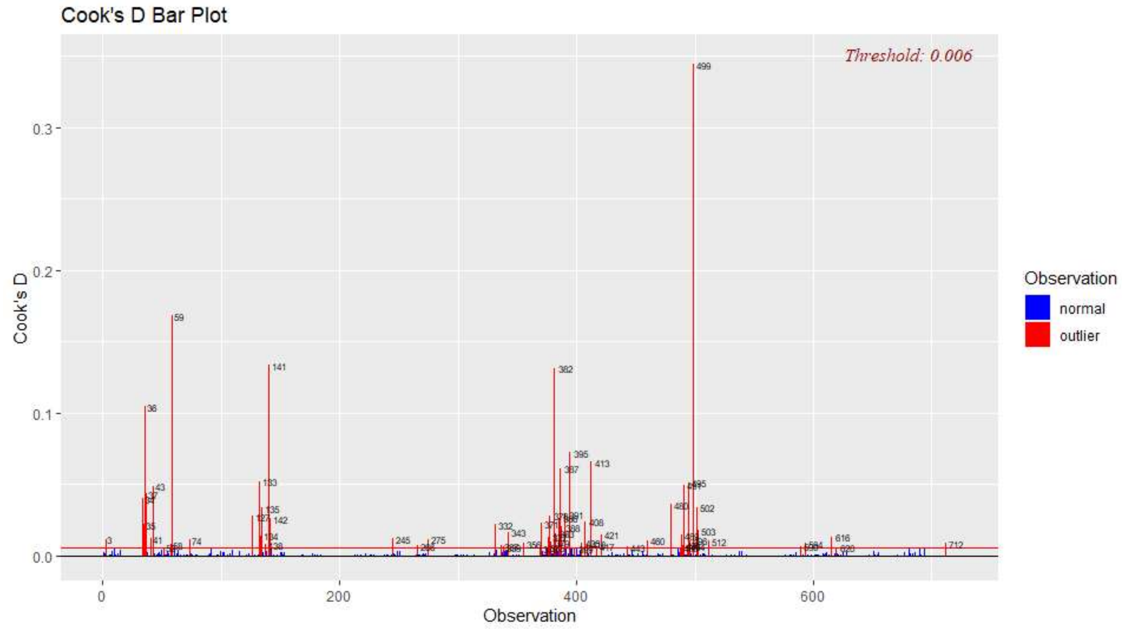


Outlier detection

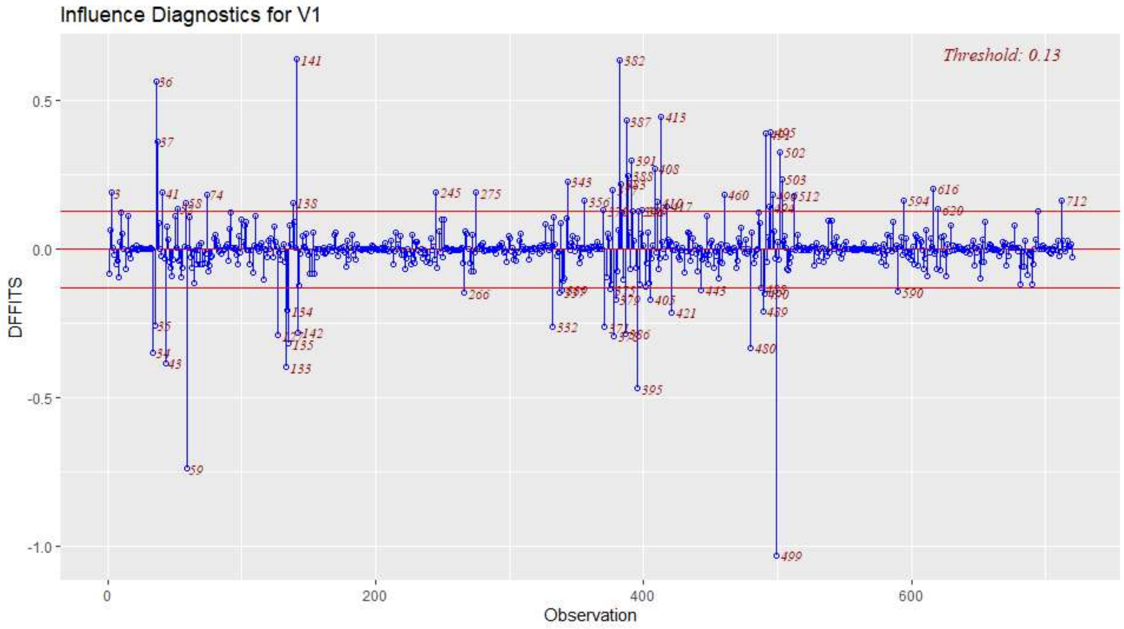
For outlier detection we plot the leverage measure for all the observations, use Cook's distance statistics, DFFITS and DFBETAS on our final model.



By the leverage measure, we obtain 35, 37, 38, 39, 41, 42, 44, 61, 74, 134, 136, 137, 138, 139, 140, 143, 266, 267, 268, 370, 373, 374, 375, 376, 377, 380, 381, 383, 386, 388, 391, 410, 417, 421, 487, 488, 489, 490, 492, 493, 494, 496, 497, 498, 506, 590, 644, 681, 712, 720 observations as leverage points and 3, 10, 58, 59, 100, 245, 332, 333, 337, 339, 340, 342, 343, 371, 396, 399, 405, 408, 456, 460, 486, 502, 503, 512, 594, 616, 620 observations as outliers.

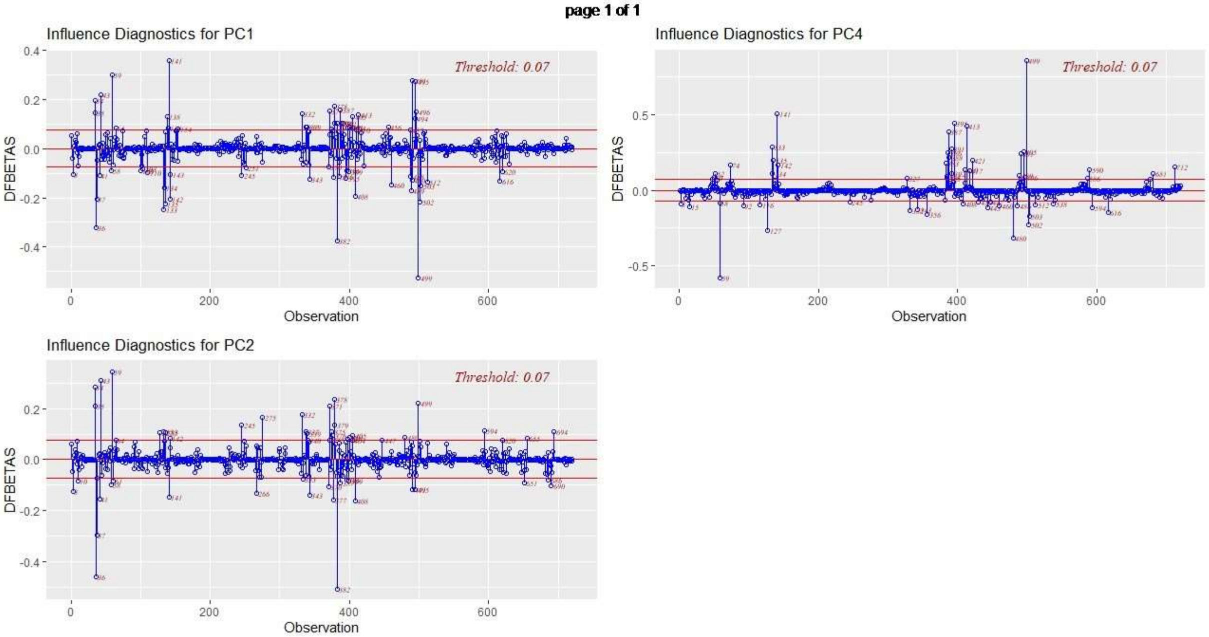


By Cook’s distance statistic we obtain 3, 34, 35, 36, 37, 41, 43, 52, 58, 59, 74, 127, 133, 134, 135, 138, 141, 142, 245, 266, 275, 332, 337, 339, 343, 356, 370, 371, 375, 377, 378, 379, 382, 383, 386, 387, 388, 391, 395, 399, 405, 408, 410, 413, 417, 421, 443, 460, 480, 488, 489, 490, 491, 494, 495, 496, 499, 502, 503, 512, 590, 594, 616, 620, 712 observations as outliers.



By DFFITS we obtain 3, 34, 35, 36, 37, 41, 43, 52, 58, 59, 74, 127, 133, 134, 135, 138, 141, 142, 245, 266, 275, 332, 337, 339, 343, 356, 370, 371, 375, 377, 378, 379, 382, 383, 386, 387, 388, 391, 395, 396, 399, 405, 408, 410, 413, 417, 421, 443, 460, 480, 488, 489, 490, 491, 494, 495, 496, 499, 502, 503, 512, 590, 594, 616, 620, 712 observations as outliers.

By DFBETAS we obtain the following:


page 1 of 1
Influence Diagnostics for PC1
Threshold: 0.07
DFBETAS
Observation
Influence Diagnostics for PC4
Threshold: 0.07
DFBETAS
Observation
Influence Diagnostics for PC2
Threshold: 0.07
DFBETAS
Observation