Project B Report

STAT 331, Fall 2017

Rosie Zou (20588049)
Simon Guo (20600133)
Azoacha Forcheh (20558994)

Summary

The objective of this report is to model the Apple Stock return rate, defined as $ret.AAPL_i = \frac{AAPL_i - AAPL_{i-1}}{AAPL_{i-1}}$, based on key metrics such as volatility, $S\&P$ 500 Index, and etc.

To model the return rate, our group has surveyed a wide range of models including an autoregressive model, linear models from automated model selection, as well as transformed linear models.

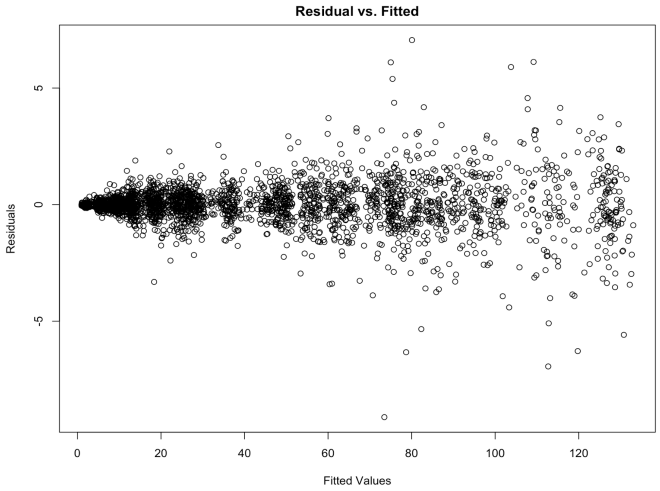
From the autoregressive model, we noticed the heteroscedascity caused by autocorrelation, which led us to determine that directly modelling the stock price is not the most ideal way to model the return rate. We also found discrepancy between the adjusted R-Squared value and the residual plots in the backfitted model, which will be further discussed later in this report. Through our investigation, our group has concluded that there is no truly ideal linear model for the Apple Stock return rate. Nevertheless, we were able to identify the inadequate models and gain insights on why those models did not work.

Model Selection

The first model to be considered is the autoregressive model, which directly models the stock price of day i using the price from day $i-1$ and has the following expression:

$$AAPL_i = \beta_0 + \beta_1 AAPL_{i-1}$$

The residual plot from the fitted model has a clearly heteroscedastic property, since the residuals take on a fan shape.



We tried to resolve the heteroscedascity issue by implementing a Weighted Least Squares autoregressive model but the results remained the same (see Appendix for full code and plot comparison). After some

research, it turns out that this is an irreducible property of the autoregressive model because the data is serially correlated. We realized that we needed something better than directly modelling the price. Hence we decided to directly model the return which, to re-iterate, is defined as $ret.AAPL_i = \frac{AAPL_i - AAPL_{i-1}}{AAPL_{i-1}}$.

Correlation between the coefficients was also expected from the data. Since it is a time-series financial data set, we knew and verified that time is correlated to variables such as the Apple Stock Price and the $S\&P$ 500 Index. The detailed matrix can be found in the Appendix section. We decided that this wasn't too much of an issue since it is an inevitable part of any time-series data.

Moving on from the autoregressive model, we then considered automated model selecting using forward selection, backward elemination, and stepwise selection, from which we obtained the following models (full code in Appendix):

Forward selection: $Return = \beta_0 + \beta_1 VIX + \beta_2 AAPL2 + \beta_3 AAPL + \beta_4 SPGSCITR$

Backward elimination: $Return = \beta_0 + \beta_1 AAPL + \beta_2 VIX + \beta_3 SPGSCITR + \beta_4 AAPL2$

Stepwise selection: $Return = \beta_0 + \beta_1 VIX + \beta_2 AAPL2 + \beta_3 AAPL + \beta_4 SPGSCITR$

Interestingly, all selection methods returned the same model. Before we proceeded any further, we consulted with Prof. Zeng. After receiving helpful advice from the professor, we realized that since *Return* is calculated as a function of $AAPL$ and $AAPL2$, the two variables cannot be included in the same model. We then adjusted the full model by excluding the variable $AAPL$, hence only keeping the previous-day price, and obtained the results below:

Forward selection: $Return = \beta_0 + \beta_1 VIX + \beta_2 AAPL2 + \beta_3 SPGSCITR + \beta_4 EEM$

Backward elimination: $Return = \beta_0 + \beta_1 VIX + \beta_2 AAPL2 + \beta_3 SPGSCITR + \beta_4 EEM$

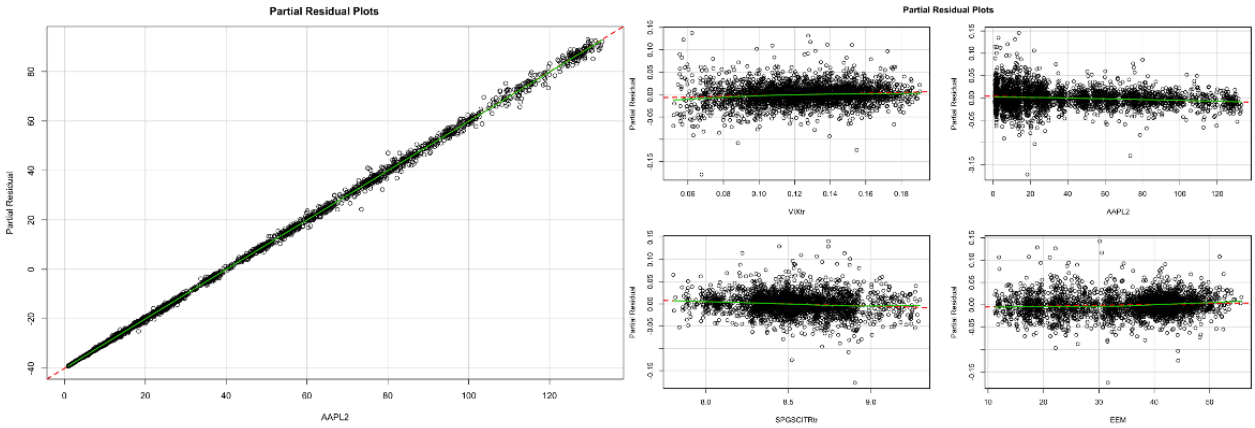
Stepwise selection: $Return = \beta_0 + \beta_1 VIX + \beta_2 AAPL2 + \beta_3 SPGSCITR + \beta_4 EEM$

We decided to further analyze this model in the model diagnostics section.

Additionally, we considered how the distribution of the independent variables could affect the model fit. Since both VIX and $SPGSCITR$ had heavily skewed histograms, we decided to log-transform the variables to make the distributions resemble a normal curve and obtained the following model: **Transformed model:** $Return = \beta_0 + \beta_1 (log(VIX)^{-2}) + \beta_2 AAPL2 + \beta_3 (log(SPGSCITR)) + \beta_4 EEM$. This will be the second model to be discussed in the model diagnostics section.

Model Diagnostics

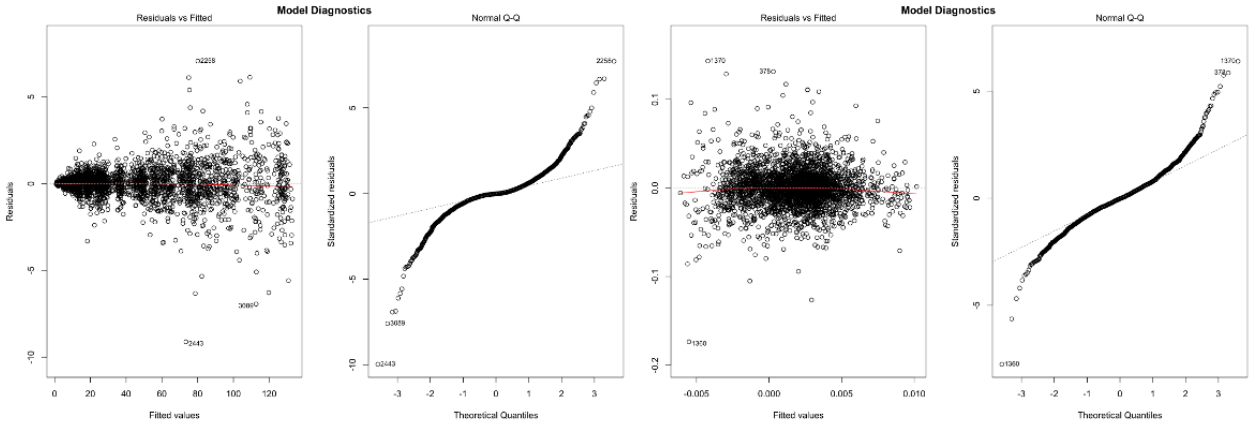
We first look into some diagnostic plots of the two models for the sake of residual analysis and outlier detection.



(a) Autoregressive Model (b) Automated Model

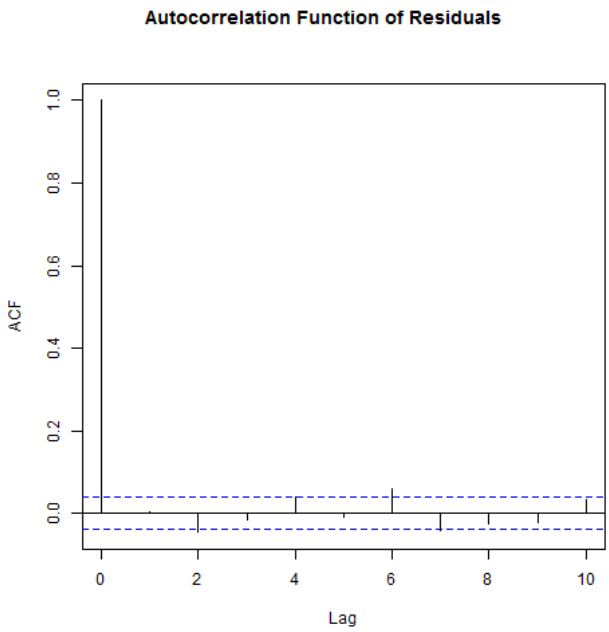
There is a strong linear trend in the partial residual plot of the autoregressive model, and no deviation from the straight line. We see this trend with the covariates of the automated model as well, with only the 'EEM' covariate and the log-transformed 'VIX' displaying a small amount of deviation from the straight line. Hence, the assumption of linearity is satisfied for both models.

The residuals of the autoregressive model are heteroscedastic as the plot of the residuals versus the fitted values has a fan-shaped pattern. In addition, the Q-Q plot of the residuals is heavy-tailed and exhibits curvature. In comparison, the plot of the residuals versus the fitted values of the automated model has more randomness and lacks the fan-shaped pattern indicative of heteroscedasticity; the Q-Q plot, while still being heavy-tailed, follows the linear trend more closely.

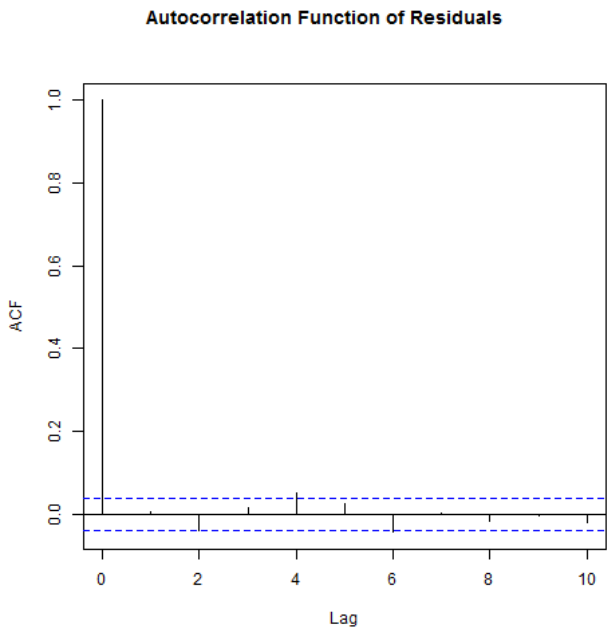


(a) Autoregressive Model (b) Automated Model

Hence, the automated model produces errors that satisfy the assumption of constant error variance and that have non-normality that might be reasonably ignored, while the plots of the autoregressive model indicate that there are non-constant variances of the error and non-normality that we may be able to fix.



(a) Autoregressive Model



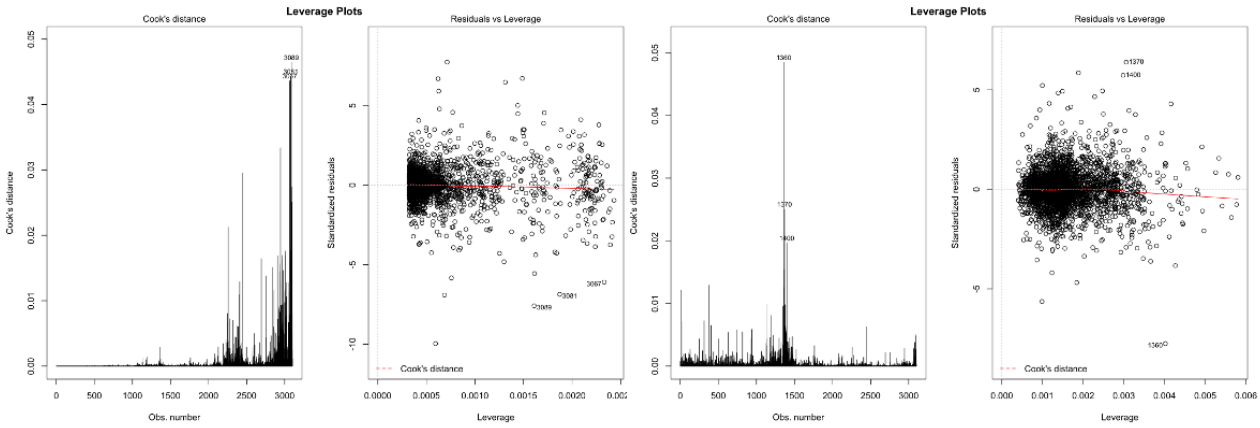
(b) Automated Model

We then looked into the last assumption of linear regression models: independence of the errors. As the stock data was time-series data and contained serial correlation, a scatterplot matrix was ruled out as a useful tool for detecting correlation. We instead considered ACF plots of the residuals, in combination with the results of the Durbin-Watson test. As can be seen in the ACF plots above, both models have explained away the autocorrelation in the residuals very well, as almost all of the sample correlations are within the limits. The results of the Durbin-Watson test on the models (shown in the table below) confirms this: for both models, the p-value of the test is high enough to accept the null hypothesis that there is no autocorrelation between the residuals.

	Test Statistic	p-value
Autoregressive	1.989882	0.7398676
Automated	1.992961	0.8450008

* alternative: true autocorrelation is not 0

The next concern we addressed was potential outliers and high-leverage cases. In both Cook's Distance plots below, there are no concerning observations that have a distance greater than 1, but most observations have a Cook's Distance greater than $\frac{p}{n}$, which is 0.001612 and 0.000403 for the automated and the autoregressive model, respectively. However, we see in both Residuals versus Leverage plots that there are no cases that are influential to the regression results. All points are well enough within the Cook's distance lines that said lines do not appear on either plot.



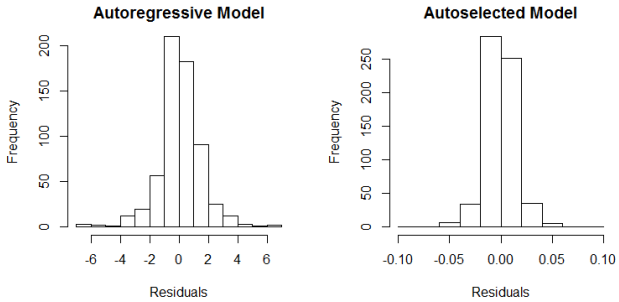
(a) Autoregressive Model

(b) Automated Model

Once we had performed residual analysis and outlier detection, we performed cross validation for the two models to determine if they were overfitting. The k-fold cross validation method was used - with $k = 10$ - and the $RMSE$ over the 10 folds was calculated. For the sake of comparison, since the models have different dependent variables and hence different scales, the $RMSE$ had to be normalized by dividing it by the range of the observed values. We denote this calculation as the $NRMSE$.

The autoregressive model has an $NRMSE$ of 0.0071775, a value very close to 0, indicating that this model is a very good fit. While still being a well-fitted model, the automated model will generally produce less accurate predictions than the former model as it has a higher $NRMSE$ of 0.0706321.

We explored this claim by using the two models to predict responses for a test set of data, obtained by splitting the original dataset via an 80:20 split. The plots below are of the distributions of the error of prediction obtained from this test.



Prediction Error Histograms

The autoregressive model produces far more inaccurate results when predicting, with the value of the residual being as high as 6 in some cases. The autoselected model performs extremely well, producing a higher number of accurate predictions and a much smaller range of error.

Based on our aforementioned results, the final model we decided to choose was the automated model, i.e. $Return = \beta_0 + \beta_1(log(VIX)^{-2}) + \beta_2 AAPL2 + \beta_3(log(SPGSCITR)) + \beta_4 EEM$. While both models are well-fitted with most of the linear regression assumptions satisified and no significant outliers, the residuals of the automated model are homoscedastic and are closer to normality than those of the autoregressive model. In addition, though the autoregressive model is less overfitted than the automated model (as measured by the $NRMSE$), it is not by a significant difference and the automated model tends to produces more accurate predictions.

The paramters estimates and confidence interval of the final model are displayed in the table below.

	Estimate	Intervals
(Intercept)	0.0729816	[0.0311, 0.1149]
VIXtr	0.0829201	[0.047, 0.1189]
AAPL2	-0.0000878	[-1e-04, 0]
SPGSCITRtr	-0.0098632	[-0.0151, -0.0046]
EEM	0.0001745	[0, 3e-04]

We conclude that neither model is a good enough fit. However, while the automated model has a low Adjusted R^2, it predicts very accurately and is a strong model otherwise. Given more time and resources, we could potentially develop a better model using the knowledge we now have about the relationship between the covariates and the response.

For example, an artificial neural net could be implemented instead, especially as this is a popular method for predicting stock returns and there is more research that can be considered during implementation. Alternatively, we could build on the autoregressive model to create an ARIMA model, which a good choice of model for time series data such as this.

Discussion

We answer the questions posed to us in the Instructions file in this section.

1. *What are the most important factors affecting Apple Stock returns?*

   According to the results from the R code below, we can check the p-values for the significance of coefficients with the corresponding covariates. For the transformed model, VIX term has the lowest p-value as shown in the summary, which means it would make largest change to the response if it drops out of the model; while $AAPL2$ term, which is the pervious day's price, and $SPGSCITR$ term's coefficients ealso have relatively small p-values. For our selected model before transformation, VIX has the lowest p-value as well, and again $AAPL2$ and $SPGSCITR$ terms' coefficients have considerably small p-values too. So based on our models, VIX would be the most important factor, and $SPGSCITR$ and $AAPL2$ are considerably significant as well.

2. *Does including information for several days ago have much of an impact on tomorrow's return, if we account for today's information?*

   As shown in the R code in the Appendix, after attempting to instead of using previous day's price to calculate return, we take the average of the previous 5 days' stock prices and still fitting the same transformed model. Comparing to original transformed model, the performance of the model seems to be better in temrs of coefficient of determination and the AIC values, while there is a small increase in adjusted R^2 and a small decrease for AIC. So it seems that from our attempt, including information for several days ago would give a better model fitting the tomorrow's return, while we are still accounting for today's information.

3. *Does the model predict accurately?*

   Based on our results on our k-fold cross validation with the comparsion of training and testing sets of data, the model performs generally well for prediction accuracy. Although in the summary of the model, the coefficient of determination is not ideally high, as we have discussed in class and the fact that R^2 can be affected by other factors (such as number of variables that are involved), so it would be acceptable to say that the model predicts the future value with reasonably accurate results.

4. *Are there any coefficients with high p-values retained in the final model? If so, why?*

According our models, fortunately we don't have any coefficients with high p-values retained in the final model. All of the coefficients have descent p-values that are lower than 0.05.

5. *Are any of the regression assumptions of the final model violated? If so, which ones?*

   This question has been mentioned along with the Model Diagnostics part. All of the regression assumptions seems to be satisfied by our final model, instead of there is a little bit bias on the independence of residuals.

6. *What are the possible deficiencies of the final model? How the model can be improved?*

   As discussed in previous analysis part, and also by looking at the residual plots and the summary of aborted models, the data has a tendency to be fitting a non-linear model. For example, there is a curved trend for the residual against fitted values plot, even after the chosen transformation. And notice that from the R results we can see that the adjusted and unadjusted coefficients of determination is very small and not close to 1 at all, which is not ideal as well. To improve the fittness of the model, it might make more sense to fit a non-linear model instead. Alternatively, there is the option of adding more covariates to the model, but that would increase the possiblity of overfitting the model.

7. *Are there any outlying observations that might be appropriate to remove?*

   By judging from the results of residual plots and leverage plots, it can be seen that there seems to have several outliers concentrated around the year 2008, which has the index 1370. The possible reason for the outlying results could be the finacial crisis during that time that would cause the drop in variates' values, but AAPL stock price didn't be affected that much, possible explanation could be the release of new series of iPhone which sort of stablize the change in stock price for Apple. Under such circumstance, there might exist possible reasons for the unexpected change, here it would be more reasonable to ignore the outlying results, as it is not really "unexpected" or it is abnormal. Therefore, it would not be appropriate to remove these values.

Appendix

Below is a collection of R code used in this report.

Pre-process the data

```
market_origin <- read.csv("market_index_clean.csv")
first <- market_origin[1,]
market <- market_origin[2:3104,]
market$AAPL2 <- market_origin[,"AAPL"][1:3103]
market$Return <- (market$AAPL - market$AAPL2)/market$AAPL2

market$Date <- market_origin$Date[2:3104]
date <- market$Date
first <- date[1]
day <- c()
for(i in 1:length(date)) {
  diff = round(difftime(date[i], first))
  day = c(day, diff)
}
market$Date <- day

## Splitting the data into a training and test set (80/20)
n <- length(market[,1])
split <- round(0.8*n)
market_train <- market[1:split,]
market_test <- market[(split+1):n,]
```

Full Model

```
full_model <- lm(Return~., data = market)
summary(full_model)
```

Null Model

```
null_model <- lm(Return~1, data = market)
summary(null_model)
```

Autoregressive Model

```
autoregressive_model <- lm(AAPL~AAPL2, data = market_train)
plot(fitted(autoregressive_model),residuals(autoregressive_model),
     main = "Residual vs. Fitted", xlab = "Fitted Values", ylab = "Residuals")
```

WLS Autoregressive Model

```
wts <-1/(fitted(lm(abs(residuals(autoregressive_model))~fitted(autoregressive_model)))^2)
wls <- lm(market_train$AAPL~market$AAPL2, weights = wts)
par(mfrow=c(1,2))
plot(fitted(autoregressive_model), autoregressive_model$residuals,
     main="unweighted", xlab = "fitted", ylab = "residuals")
plot(fitted(wls), wls$residuals, main="weighted", xlab = "fitted", ylab = "residuals")
```

There is no apparent difference between the weighted least squares and the un-weighted autoregressive models.

Correlation Matrix

```
cor(market_train)
```

Forward Selection

```
step(null_model, scope = list(upper=full_model), direction="forward")
summary(lm(formula = Return ~ VIX + AAPL2 + AAPL + SPGSCITR, data = market_train))
```

Backward Elimination

```
step(full_model, scope = list(lower=null_model), direction="backward")
summary(lm(formula = Return ~ AAPL + VIX + SPGSCITR + AAPL2, data = market_train))
```

Stepwise Selection

```
step(null_model, scope = list(upper=full_model), direction="both")
```

Forward Selection without AAPL

```
full_model <- lm(Return~SPX+VIX+SPGSCITR+BNDGLB+EEM+AAPL2, data = market)
step(null_model, scope = list(upper=full_model), direction="forward")
```

Backward Elimination without AAPL

```
step(full_model, scope = list(lower=null_model), direction="backward")
```

Stepwise Selection without AAPL

```
step(null_model, scope = list(upper=full_model), direction="both")
```

Log-transformed model from stepwise selection results

```
## note that the VIX needed to be raised to the -2, after the log transformation
##  in order to obtain a distribution close to the Gaussian curve
market.tr <- market
market.tr$VIXtr <- log(market$VIX)^(-2)
market.tr$SPGSCITRtr <- log(market$SPGSCITR)
market.tr_train <- market.tr[1:split,]
market.tr_test <- market.tr[(split+1):n,]

transformed_model <- lm(Return~VIXtr+AAPL2+SPGSCITRtr+EEM, data=market.tr_train)
summary(transformed_model)
```

Function for plotting diagnostic tools

```
diagnostic_plots <- function(model) {
  p <- length(model$coef) - 1 # the number of covariates in the model
  nrows <- p%/%2
  savepar <- par(mfrow=c(1,1))
  if (nrows > 0) {
    savepar <- par(mfrow=c(nrows, 2))
  }

  # Partial Residual v.s. Covariates Plot
  savepar
  crPlots(model, ylab="Partial Residual", smooth=T,
          main='')
  title("Partial Residual Plots",
        outer=T, line=-1)

  # Q-Q Plot and Residual v.s. Fitted Plot
  par(mfrow=c(1,2))
  plot(model, which=c(1,2))
  title(paste("Model Diagnostics"),
        outer=T, line=-1)

  ## Leverage Plots
  par(mfrow=c(1,2))
  plot(model, which=c(4,5))
  title(paste("Leverage Plots"),
        outer=T, line=-1)
}

diagnostic_plots(autoregressive_model)
diagnostic_plots(transformed_model)
```

ACF Plots for the Residuals

```
## Correlation calculations and plot
## Using ACF Plots instead of scatterplot matrix
png('arm_acf_residuals.png')
```

```
acf(resid(auto_regressive_model), main='', lag.max=10)
title("Autocorrelation Function of Residuals",
      outer=T, line=-1)
dev.off()

## Using ACF Plots instead of scatterplot matrix
png('fm_acf_residuals.png')
acf(resid(final_model), main='', lag.max=10)
title("Autocorrelation Function of Residuals",
      outer=T, line=-1)
dev.off()
```

Durbin-Watson Test for autocorrelation

```
library(magrittr)
library(knitr)
library(kableExtra)

fm_dwtest <- dwtest(transformed_model, alternative=c("two.sided"),
                    data=market.tr_train)
arm_dwtest <- dwtest(auto_regressive_model,
                     alternative=c("two.sided"), data=market_train)

dwtest_dt <- data.frame(
  c(fm_dwtest$statistic, arm_dwtest$statistic),
  c(fm_dwtest$p.value, arm_dwtest$p.value)
)
rownames(dwtest_dt) = c('Autoregressive', 'Automated')
colnames(dwtest_dt) = c('Test Statistic', 'p-value')

kable(dwtest_dt, "latex") %>%
  kable_styling(full_width = F) %>%
  add_footnote(paste("alternative:",fm_dwtest$alternative),
               notation = "symbol")
```

Cross-validation and calculation of the Normalized RMSE

```
library("lattice")
library("DAAG")
arm_cv = cv.lm(data=market_train, form.lm=auto_regressive_model,
               m=10, printit=F)
fm_cv = cv.lm(data=market.tr, form.lm=final_model,
              m=10, printit=F)

nrmse <- function(mse, ylim) {
  yrange = ylim[2] - ylim[1]
  sqrt(mse)/yrange
}

arm_nrmse = nrmse(attr(arm_cv, "ms"), range(market_train$AAPL))
fm_nrmse = nrmse(attr(fm_cv, "ms"), range(market.tr$Return))
```

Predicting the response on the test dataset

```
arm_preds = predict(autoregressive_model, market_test)
fm_preds = predict(transformed_model, market.tr_test)
arm_test_resids = market_test[,'AAPL'] - arm_preds
fm_test_resids = market.tr_test[,'Return'] - fm_preds

## Plots of the residuals from prediction
par(mfrow=c(1,2))
hist(arm_test_resids, xlab='Residuals', cex=13,
     main="Autoregressive Model")
hist(fm_test_resids, xlab='Residuals',
     main="Autoselected Model")
```

Parameter Estimates and CI of Final Model

```
ci_df <- data.frame(round(confint(transformed_model), 4))
ci_df['Estimate'] <- unname(transformed_model$coef)
intervals = paste(ci_df[,1], ci_df[,2], sep=', ')
ci_df['Intervals'] = paste0('[', intervals , ']')
ci_df[,1:2] = NULL
kable(ci_df, "latex") %>%
  kable_styling(full_width = F)
```

The impact of including more data from previous days

```
## suppose we include previous 5 days data information

## attempt to using the average value of previous 5 days AAPL
AAPL5 <- c()
for (i in 1:(nrow(market_origin) - 4)){
  set <- c(market_origin$AAPL[i], market_origin$AAPL[i+1], market_origin$AAPL[i+2], market_origin$AAPL[
  AAPL5 <- c(AAPL5, mean(set))
}
new_AAPL = market$AAPL[4:length(market$AAPL)]
new_Return <- (new_AAPL - AAPL5)/new_AAPL
new_VIXtr = VIXtr[4:length(VIXtr)]
new_SPGSCITRtr = SPGSCITRtr[4:length(SPGSCITRtr)]
new_EEM = market$EEM[4:length(market$EEM)]
new_transformed_model <- lm(new_Return~new_VIXtr+new_SPGSCITRtr+new_EEM+AAPL5)
summary(new_transformed_model)
AIC(new_transformed_model)
AIC(transformed_model)
```