# APPLE STOCK PROJECTION REPORT

Submitted By:

Mark Erenberg

Valerie Tsang

Mark Nishikawa

Wednesday, December 6th, 2017

## PART I: SUMMARY

The objective of this report is to determine a linear regression model that could model the return of Apple Inc. stock (NASDAQ: AAPL) on a given day using current and past market information. To identify the best possible models, we analyzed the skewness of the explanatory variables, then checked the model assumptions and the mean squared prediction error, and used automated model selection to select the correct number of covariates. After narrowing down two candidate models, we then evaluated the influence of outliers, the raw and studentized residual plots, and the models' predictions of the response based on publicly available market information.

Through these analyses, we uncovered many insights about how best to model the response. We found that models without historic variables were unable to explain much of the variance in the response, and were a poor fit for the data. We also found that adding a historic price variable allowed the models to capture more of the total variance in the response, and have a smaller prediction error. Through model diagnostics, we ultimately discovered that the best fitted model to predict the response, given the context of the data, was a model that included the price, historic price, volatility index, and commodity index variables.

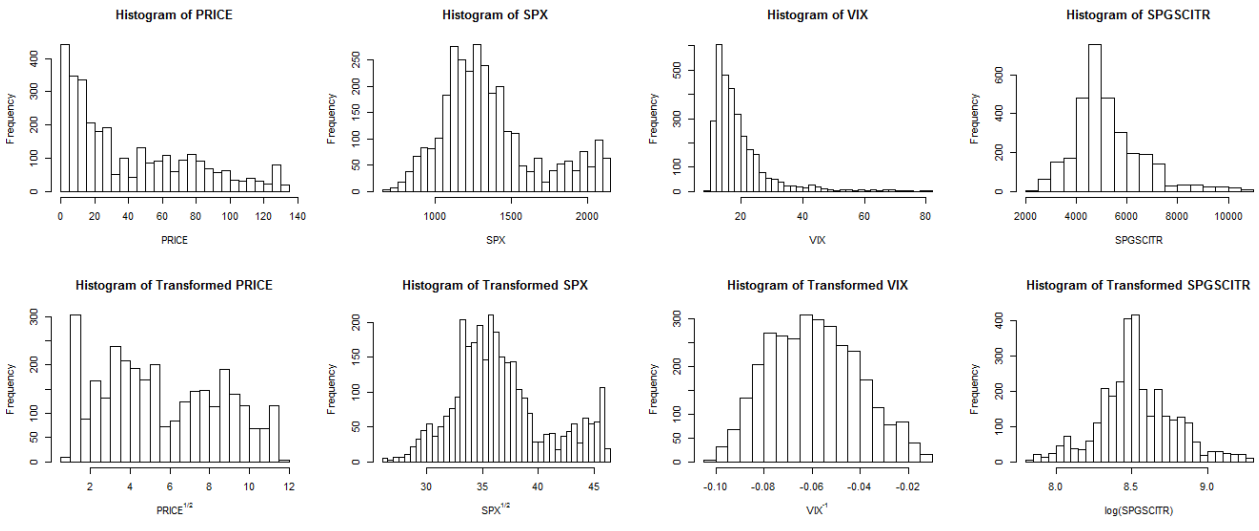
## PART II: MODEL SELECTION

For our data set, each observation of data corresponds to information that was recorded on a respective weekday dating from April 14$^{th}$, 2003, up until September 14$^{th}$, 2015. The response variable is the return of Apple stock on a given day, which will be hereby be denoted as RETURN. The explanatory variables include this day's date, Apple stock quote, S&P 500 Index, CBOE Volatility Index, S&P Goldman Sachs Commodity Index, Dow Jones Barclays Capital Bond Index, and Morgan Stanley Emerging Markets Index, which will hereby be denoted as DATE, PRICE, SPX, VIX, SPGSCITR, BNDLGB, and EEM respectively.

In selecting a model, an *initial model* was fitted that included these seven explanatory variables: DAYS, PRICE, SPX, VIX, SPGSCITR, BNDGLB, and EEM. Note that DATE was converted to days passed since the first observation, denoted by DAYS. Initial diagnostics of the histograms of each covariate showed that the PRICE, SPX, VIX and SPGSCITR covariates showed some skewness. Figure 1 displays the skewness levels for these covariates. The following power transformations were applied in order to even out these histograms and make them more evenly distributed across the range (i.e. decrease skewness):

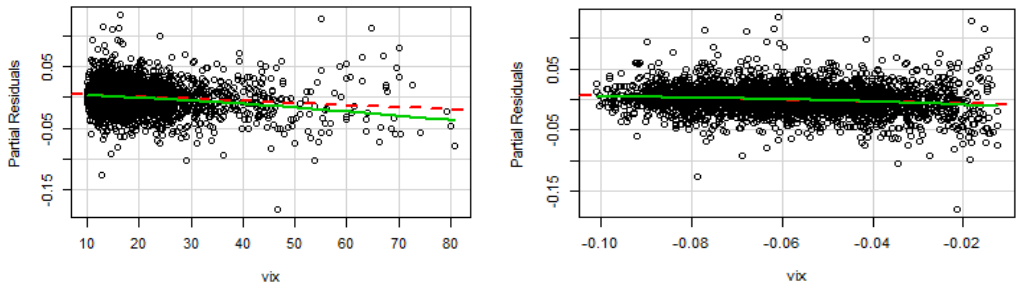
$$\boldsymbol{PRICE = PRICE^{1/2},\ SPX = SPX^{1/2},\ VIX = VIX^{-1},\ SPGSCITR = log(SPGSCITR)}$$

*Figure 1: Transformations effect on histograms of covariates*



In addition to decreasing the skewness, these transformations also spread out the data points in the corresponding residual plots, and helped reveal any systematic patterns in the residuals. Also, the partial residual plot for VIX in the *initial model* with untransformed variables deviated from the line of best fit, as is depicted in red in Figure 2 below. After transforming VIX, the same model did not show any deviation in the VIX partial residual plot. This transformation's improvement in the distribution of residuals, in addition to its reduction of the deviations in the VIX covariate's partial residuals plot, are graphically illustrated in Figure 2. The rest of the partial residual plots did not show any deviation from the line of best fit, indicating that the covariates had a linear relationship with the response in our model. The residuals vs fitted values plot showed no systematic patterns or curves, which would have been indication of an inadequate fit. Moreover, the residual plots and Q-Q plots showed no indication against homoscedasticity or normality of errors. By observing the autocorrelation function, otherwise known as the ACF, and the Durbin Watson Test statistic for this model, there was no evidence of autocorrelation in the errors, as the ACF stayed within the range for all but two lags, which extended barely beyond the top band. Further, we had no evidence against $H_0$: $corr(\varepsilon_t, \varepsilon_{t-1}) = 0$ from the DW-Test statistic of 1.9844.

*Figure 2: Transformations effect on partial residual plot of VIX*




The *initial model* had an $R^2$ value of 0.01125 and a mean squared prediction error value, otherwise known as an MSPE value, of 0.0005035645. The $R^2$ value was very low, meaning the model barely captured any of the total variance, and did a poor job fitting the data. With such a poor fitting initial model, any model selected through automated model selection would only have produced a worse fitting model. So rather than proceeding with this model, we added more covariates to contribute more information about the response.

It is important to note that we did not address the issue of multicollinearity within our data set. The reason for this was because the data given was observed sequentially in time, and the explanatory variables were therefore expected to be highly correlated to each other. As such, we did not consider the Variance Inflation Factor for our covariates, since this would have incorrectly skewed the fitted models, and provided inaccurate prediction values.

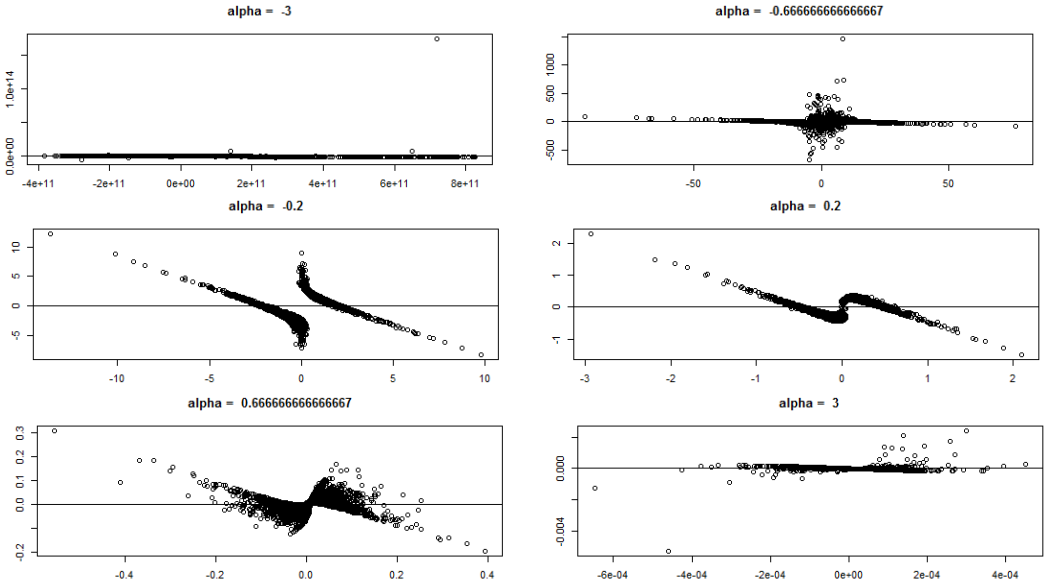
Next, it occurred to us that an autoregressive model that incorporates the information from a prior period would have more information about the response and would better fit the data. Thus we fit a *return autoregressive model* incorporating the prior day's returns to have covariates: DAYS, PRICE, SPX, VIX, SPGSCITR, BNDGLB, EEM, LAG1RETURN where the $i^{th}$ observation of LAG1RETURN refers to the $(i-1)^{th}$ return observation. This model had an $R^2$ value of 0.01294 and a MSPE value of 0.0005035122; thus, it was slightly better, but was still a poor prediction model for the data. The t-statistic for the LAG1RETURN variable was extremely high at 0.835, which hinted that this variable may have been insignificant. By looking at the partial residual plot for LAG1RETURN, we further saw that the residuals were all in one large

cluster in a small interval around zero. This confirmed that adding LAG1RETURN as a covariate did not contribute any significant information to the model, so more covariates were needed to provide more information about the response.

Considering the response variable is the stock's return, the two variables that it depends on are the given day's price and the previous day's price. Thus, we expected that a model that incorporates these two variables would explain more of the response's variance and provide a closer fit to the data. As such, a *simple price autoregressive model* was fit next with two covariates: PRICE and LAG1PRICE, where the $i^{th}$ observation of LAG1PRICE refers to the $(i-1)^{th}$ price observation. Immediately, we saw that this model fit the data much closer with an $R^2$ value of 0.7526 and an MSPE value of 0.0001261915. This makes sense, since a model that incorporates a stock's historic price would be able to explain more of the variance in the stocks' return. Additionally, we saw the partial residual plots for this model were perfectly linear, fitting exactly to the line of best fit, as can be seen in Figure 3. This was an indication that there was a strong linear relationship between the covariates and the response. The residuals/studentized residuals vs fitted values plots also had an "X" shaped distribution, where the residuals approached zero as the fitted values approached zero. That is, the variance of the residuals was proportional to the fitted values, and was not constant. Moreover, the Q-Q plot showed slightly heavy tailed residuals, meaning that the error in this model may not have been normally distributed. The Durbin Watson Test statistic of 1.9572 showed no evidence against $H_0$: $corr(\varepsilon_t, \varepsilon_{t-1}) = 0$, and the ACF was within the band for almost all lags. Thus, we had no evidence of autocorrelation in the errors.

The next step was to look at another *price autoregressive model* that included all the explanatory variables. This model had covariates: DAYS, PRICE, SPX, VIX, SPGSCITR, BNDGLB, EEM and LAG1PRICE. This model had no linearity issues in the partial residual plots, but had the same issue as the simple price autoregressive model – an "X" shape in the residuals vs fitted values plot, and non-constant variance that was proportional to the fitted value. The Q-Q plot was similarly heavy tailed. Note that no applicable power transformations on the response were able to eliminate systematic patterns in the residual plots, as is shown in Figure 3.

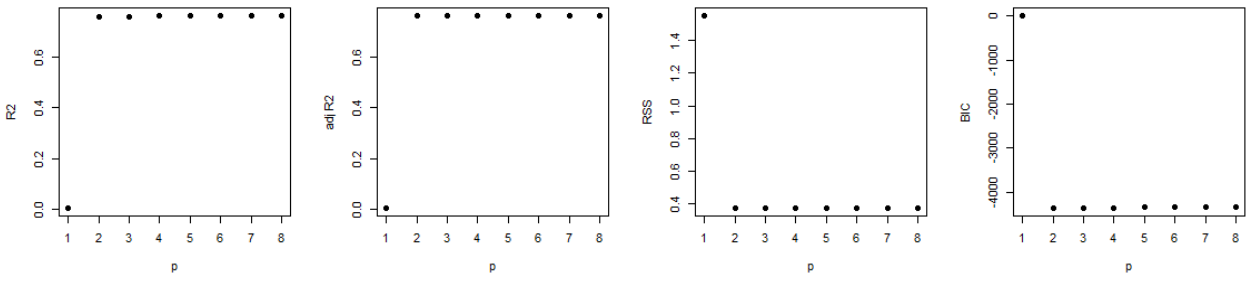
*Figure 3: Residuals vs Fitted Values plots for models with different powers of RETURN*



Although the third-degree power transform straightens the residual plot and fixes the Q-Q plot, the transformed data is all essentially zero, which leads to a worse fitted model. The weighted least squares regression model was also created by regressing the absolute residual against the fitted values as weights, but this model also could not resolve the heteroscedasticity. Again, the AFC and Durbin Watson Test showed no indication of autocorrelation in the model's errors.

Using this *price autoregressive model* as our decided model, we used the three automated model selection methods to narrow down the relevant covariates. The *stepwise regression model* and *backward elimination models* were identical, with covariates: PRICE, VIX, SPGSCITR, LAG1PRICE, whereas the *forward selection model* had six covariates: PRICE, SPX, VIX, SPGSCITR, EEM, and LAG1PRICE. Since the model with one historic price performed fit so well, we also investigated how the model behaved as we introduce more historic price variables. The next autoregressive model, the *full autoregressive model*, included the initial seven explanatory variables, with twenty-one additional explanatory variables: one for each of the twenty-one prior days' price. Note that this made the first twenty-one data observations incomplete, and they were thus excluded. The All Subsets Regression Method was used with the *full autoregressive model* to identify the best covariates to include for each value of p, where p is the number of covariates in the model. Using this method, we saw that VIX did the best job at modelling return, however once more than one covariate was considered, the combination of PRICE and LAG1PRICE fit better. Additionally, Figure 4 shows how the $R^2$, adjusted $R^2$, RSS and BIC changed as p increased, and how it appeared as though the model quality did not improve after the model with PRICE and LAG1PRICE as covariates, at p=2. That is, the *simple price autoregressive model* with covariates PRICE and LAG1PRICE did an equally good job at capturing the total variance as the best model with eight covariates.

*Figure 4: $R^2$, Adj. $R^2$, RSS and BIC plots for p = 1 to 8*




At this point we had come up with several models that we could compare using the MSPE and adjusted $R^2$ values. MSPE values were a primary focus to compare the models, as it measures the accuracy in prediction models. To supplement MSPE, adjusted $R^2$ values were also a pivotal focus, as they measure the proportion of total variance that is explained by the model, or the model's overall fit. Furthermore, adjusted $R^2$ values allowed us to compare the models that had different numbers of covariates. Figure 5 illustrates a summary of these values for each model.

*Figure 5: Summary Table of Constructed Models*

| | p | MSPE | Adj. R2 | AIC | BIC |
|---|---|---|---|---|---|
| Initial Model | 7 | 0.0005035645 | 0.01073928 | -14739.62 | -14685.26 |
| Return AutoReg. Model | 8 | 0.0005035122 | 0.01038664 | -14733.19 | -14672.79 |
| Simple Price AutoReg. Model | 2 | 0.0001261915 | 0.75246085 | -19037.76 | -19013.60 |
| Full AutoReg. Model | 28 | 0.0001206101 | 0.75931697 | -19002.39 | -18821.39 |
| Price AutoReg. Model | 8 | 0.0001259014 | 0.75255069 | -19032.90 | -18972.50 |
| Forward Selection Model | 6 | 0.0001259439 | 0.75262714 | -19035.85 | -18987.53 |
| Stepwise Regression Model | 4 | 0.0001259576 | 0.75276005 | -19039.52 | -19003.28 |
| Backward Elimination Model | 4 | 0.0001259576 | 0.75276005 | -19039.52 | -19003.28 |

Since our goal was to find the best prediction model, we used MSPE values to compare models and choose the best two. The s*imple price autoregressive model* had only two covariates, but still had an MSPE value 0.000126 and an adjusted $R^2$ value of 0.75, and thus was chosen as the first model to move forward with. The second model chosen was the b*ackward elimination model*, from the original model with the seven given covariates and LAG1PRICE. It had only four covariates, and had an MSPE value of 0.0001259576 and an adjusted $R^2$ values of 0.75. The equations of the selected models are shown below:

*Simple Price Autoregressive Model:*

***RETURN = 0.00255 + 0.33173(PRICE) – 0.33206(LAG1PRICE)***

*Backward Elimination Model:*

***RETURN = 0.01702 + 0.33123(PRICE) – 0.33161(LAG1PRICE) + 0.01944(VIX) – 0.00179(SPGSCITR)***

These two models will be referred to as *Model 1* and *Model 2* respectively.

## PART III: MODEL DIAGNOSTICS

The first measure we used to compare the two models was the presence of outliers, and any notable cases of leverage or influence. We began by calculating the leverage cases for each model. *Model 1* was deemed to have 191 high leverage cases, which is approximately 6% of the total response values, with a maximum leverage value of 0.0286447. This tells us that *Model 1* is sensitive to changes in 6% of the response values, which indicates that this model has a fair amount of rigidity with responses to changes in the response. To examine the influential cases for *Model 1*, we examined the Cook's statistics. As we can see from Figure 6, most of the Cook's statistics are significantly close to zero, indicating that the observed data values are generally non-influential on the overall fit of the model. The exception appears to be the three statistics occurring at observations 1359, 2442, and 3088. To assess the influence of these observations, we fitted a second model using the same covariates as *Model 1*, but without these observations in the data set. The results were that the model's adjusted $R^2$ value lowered to 0.7492 from 0.7525, and the MSPE value also lowered to 0.0001237097 from 0.0001261915. This tells us that removing these outliers made the model account for 0.44% less of the total variance in the data, but increased the prediction accuracy of the model by 1.97%.

*Figure 6: Cook's Statistic Plot for Model 1*

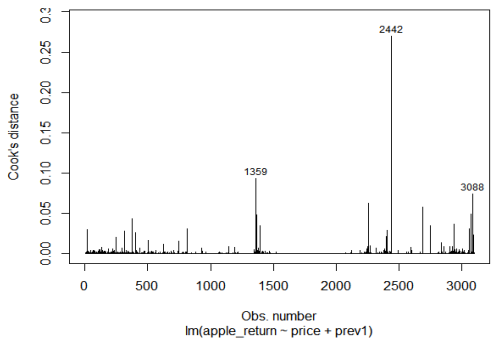
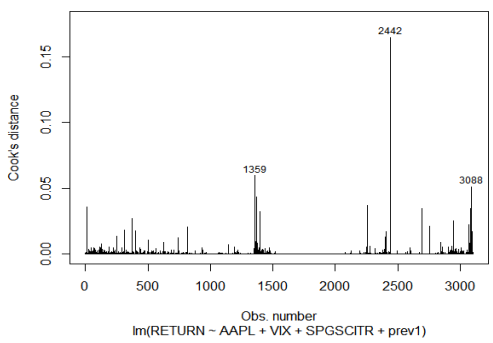



*Figure 7: Cook's Statistic Plot for Model 2*



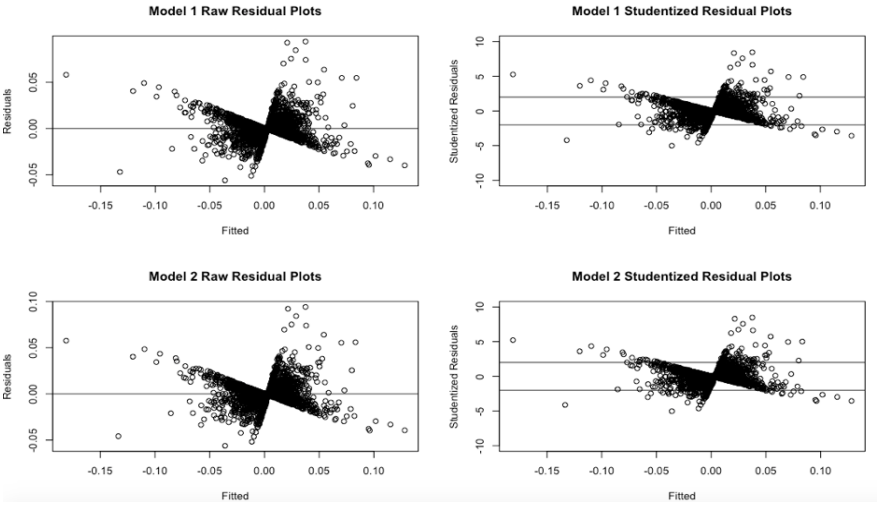

Similarly, we evaluated leverage and influence in *Model 2*. *Model 2* had 205 high leverage cases, which is approximately 7% of the total response values, and a maximum leverage value of 0.02955348. Thus, *Model 2* is sensitive to changes in 7% of the response values, and is

equivalently as rigid as *Model 1*. Figure 7 illustrates the plot of the Cook's statistics for *Model 2*, and shows that most of the Cook's statistics in this model are also close to zero, indicating that the data is non-influential in this model as well. The outliers for this model occurred at the same observations as in *Model 1*, except that the influence of these outliers was smaller in comparison. To assess the influence of these observations, we again fitted a model using the same covariates as *Model 2*, but without the outliers removed. *Model 2* 's adjusted $R^2$ value lowered to 0.7493 from 0.7525, and the MSPE value also lowered to 0.000123484 from 0.0001259576, indicating that the outliers made the model account for 0.43% less of the total variance in the data, but increased the prediction accuracy of the model by 1.96%.

One other measure we used to compare predictive accuracy of each model was cross-validation. Using 10-fold cross validation, the overall mean square of *Model 1* and *Model 2* was 0.000127 for both cases, as the sum over all 310 folds. Even when trying different variations of folds, the two models consistently had the same mean square values. This may be attributed to the low magnitude of the response variable.

The partial residual plots for PRICE and LAG1PRICE are perfectly linear in both models. Moreover, the VIX and SPGSCITR partial residual plots for *Model 2* also had no deviation from the line of best fit. Thus, there was no evidence against the linearity assumption in our models. Another measure that was used to assess the two models was a comparison of raw and studentized residuals. Both the models produced nearly identical raw and studentized residuals plots that showed that the variance of residuals in both models were proportional to the fitted values and thus was not constant. We attempted to resolve this breach of the homoscedasticity using power transformations in addition to Weighted Least Squares (WLS) regression. Using these methods, we fit new models, but were unable to stabilize the variance. Looking at the return formula, we see that it is not a linear function and thus the true relationships between our covariates and response is not linear, this would explain our inability to stabilize the variance. Figure 7 shows a comparison of both the raw residuals and studentized residuals for both model one and model two:

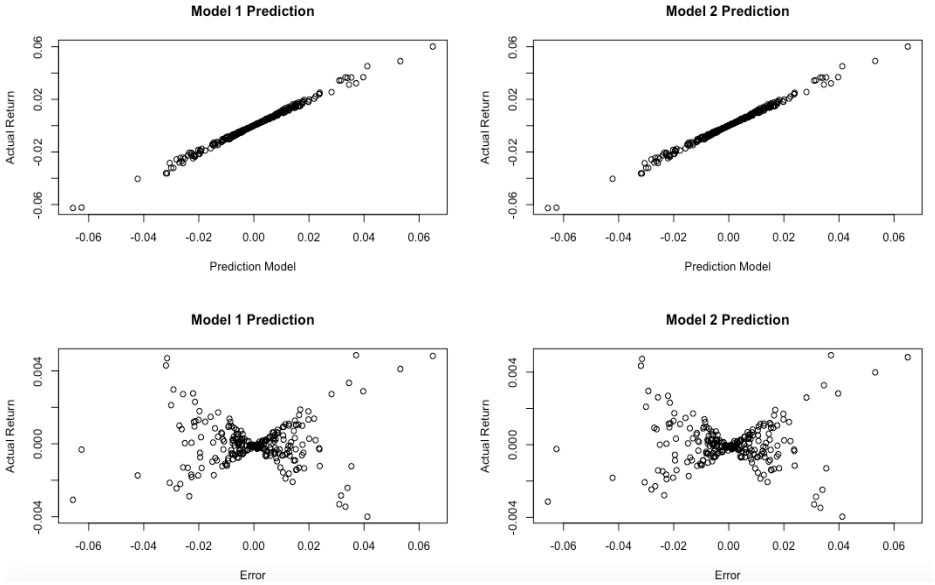
*Figure 7: Raw Residual and Studentized Residual Comparison Plots*




The final assessment used to compare the two models was to generate predicted return values using publicly available data, and to assess the accuracy of these predictions. The predicted values for both models were approximately equal, and hence have similar prediction and error plots. Figure 8 depicts the predicted return for both models against the actual return, and the error of the predicted returns against the actual return. From Figure 8, we are able to determine

that the predicted values for both models are highly correlated to the actual return, which indicates that the predicted values are fairly reliable. Looking at the error plots, we see that a majority of the prediction errors lie around zero. In reality, this is a logical outcome due to the fact that if a model could accurately predict the stock market, there would be no need for active investing.

*Figure 8: Model Prediction Plots and Prediction Error Plots*




From these varying forms of model diagnostics, we can thus deduce certain properties about the two models. Namely, that both models are fairly resistant to changes in the observed values, while the influential outliers in *Model 1* shifted the prediction accuracy of the model by 0.1% more than the shift in prediction accuracy from the influential outliers in *Model 2*. The comparison of studentized residual versus fitted values displayed that both models had an equal violation of homoscedasticity, but that the other model assumptions were not violated. Also, the predictions of the two models also appeared identical as is illustrated in Figure 8, and as is reaffirmed by the results of our 10-fold cross validation. Thus, since there is such a strong similarity between the predictive accuracy of the two models, we chose the *backward elimination model* as our final model, since it had a higher adjusted $R^2$ value and a smaller MSPE value. While it did have more covariates than the s*imple price autoregressive model*, this implied that it was better able to account for external market factors outside of the Apple stock price, and was thus a better indication of the influence of the financial indices on the stock's return. Thus, the formula for our final fitted model is:

***RETURN = 0.01702 + 0.33123(PRICE) – 0.33161(LAG1PRICE) + 0.01944(VIX) – 0.00179(SPGSCITR)***

## PART IV: DISCUSSION

The p-values for our final model are summarized below:

*Figure 8: Coefficients and p-values of final model:*

| FINAL MODEL | | |
|---|---|---|
| COVARIATE | COEF | P-VALUE |
| PRICE | 0.33123 | $2.0 \times 10^{-16}$ |
| LAG1PRICE | −0.33161 | $2.0 \times 10^{-16}$ |
| VIX | 0.01944 | 0.0818 |
| SPGSCITR | −0.00179 | 0.0482 |

For our final model, PRICE and LAG1PRICE both had p-values of essentially zero, however VIX and SPGSCITIR had higher p-values of 0.0818 and 0.0482 respectively. At the usual 0.05 significance level, the VIX covariate would not be significant, however at the 0.15 significance level for backwards elimination, VIX was left in the model. Given the high p-values of VIX and SPGSCITR relative to those of PRICE and LAG1PRICE, we might consider VIX and SPGSCITR to be relatively unnecessary in our model. However, we decided to keep them in our model on account that they decreased the overall MSPE of the model, and are a more revealing indication of the influence of external market factors on the stock's return. Also, since our final model had PRICE and LAG1PRICE as the highest magnitude coefficients, this was an indication that the stock price and historic stock price were the two most important factors affecting the response. Also, since this model had a much higher adjusted $R^2$ value than the models that did not include historic stock price as a covariate, we can conclude that including previous information on price had a significant impact on the predicted return.

It is also worth noting the influence of outliers on our final model. We found that the model was sensitive to changes in only 7% of the response values through 205 determined cases of high leverage. However, a more telling measure of the outliers' influence was the Cook's statistics, the plot for which revealed that there were three outliers amongst the data set. Since their Cook's statistics were still relatively low, ranging from 0.05 to 0.20, this meant that the data had relatively low influence on the fitted values. Even after we removed these three outliers from the data set and refit the model, we found that the outliers made the model account for only 0.43% more of the total variance in the data, and decreased the prediction accuracy by only 1.96%.

For our final model, the partial residual plots revealed no deviation from the line of best fit, and showed no evidence against the linearity assumption. The error of this model was not constant, and we were unable to stabilize this variance with transformations and WLS regression. However, the heteroscedasticity was explained by the true relationship seen in the return formula that clearly shows the relationship between RETURN, PRICE and LAG1PRICE as non-linear. The Q-Q plot assessing normality of error showed slightly heavier tales, but no compelling evidence against the normality assumption. As well, the ACF function and Durbin Watson Test showed no indication of autocorrelation in the errors. Therefore, with the exception of the homoscedasticity assumption, all other model assumptions for this model were satisfied.

Lastly, since the prediction values calculated using publicly listed market information had a high correlation to the actual response values, and had predominantly zero prediction errors, we can conclude that this model does the predict the response with a strong level of accuracy. This is also confirmed by the small mean square value obtained from 10-fold cross validation, and the model's relatively small overall MSPE value.