Assignment: Homework 4 Report

Name: Horace Fung

Instructor: Professor Simonoff

Course: Regression and Multivariate Data Analysis

Modelling the monthly return of Brazilian arabica coffee beans

This report examines the relationship between the monthly returns on Brazilian arabica coffee beans and several potential predictors. Arabica is one of two main species of coffee, the other being robusta, and Brazil is by far the leading producer of arabica. This report is motivated by the challenges that farmers and farmworkers face due to the volatility of coffee prices, particularly sudden drops in price. In many commodity markets, futures tend to help smoothen volatility and provide both sellers and buyers more certainty. However, coffee bean prices continue to swing and volatility remains a major obstacle for Brazil's coffee industry.

The data is from five sources and spans from January 2009 to December 2016. 2016 was held out as a validation set and accounts for 12.5% of total observations. Coffee prices were collected from the Center for Advanced Studies on Applied Economics at the University of São Paulo. Temperature and rainfall data were collected from The World Bank. Coffee production data was collected from the Brazilian Institute of Geography and Statistics. USD to Real exchange rate data was collected from Federal Reserve Economic Data (FRED). Finally, the Oceanic Niño Index data was collected from Climate Prediction Center NOAA. Coffee prices were adjusted for inflation with CPI data from FRED, using 2016 as the base year.

1. Definition of Variables:

The initial data contains eight variables. The Seasons predictor is encoded into three binary variables. The definitions of all the variables are listed below:

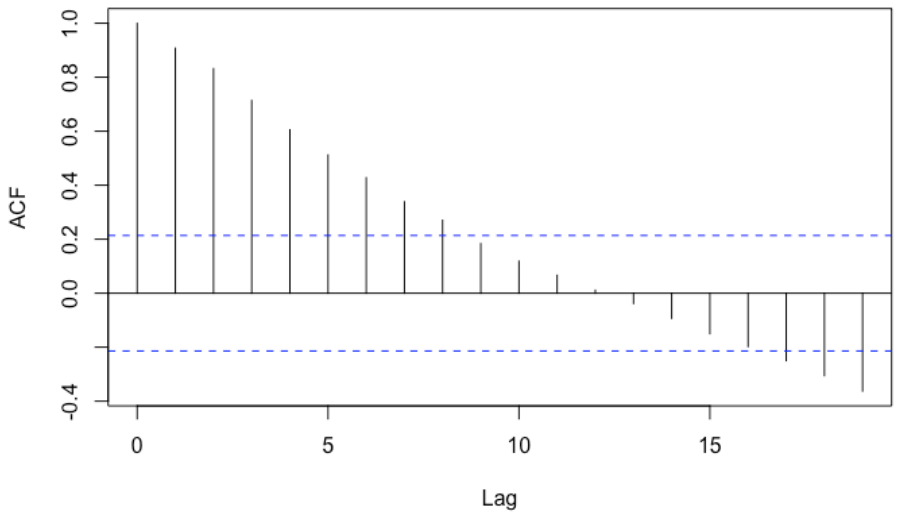
1. *Coffee_Return (Response):* The monthly percentage return on Brazilian arabica coffee beans. The prices used to compute return are denominated in US Dollars.
2. *Temperature:* The average monthly temperature of Brazil measured in Celsius.
3. *Rainfall*: The average monthly rainfall in Brazil measured in millimeters.
4. *Lagged_Temp*: The 12-months lagged monthly average temperature of Brazil measured in Celsius.
5. *Lagged_Rainfall*: The 12-months lagged monthly average rainfall in Brazil measured in millimeters.
6. *FX_Return*: The monthly percentage return on the USD to Real exchange rate.

7. *Production_Change*: The monthly percentage change in quantities of arabica coffee beans produced in Brazil.
8. *Lagged_ONI*: The 12-months lagged monthly Oceanic Niño Index. Numbers above 0.5 is an El Niño event and numbers below -0.5 is a La Niña event.
9. *Seasons*: Three binary variables that indicate winter, spring and summer. In Brazil, winter is from June to August, spring is from September to November and summer is from December to February.

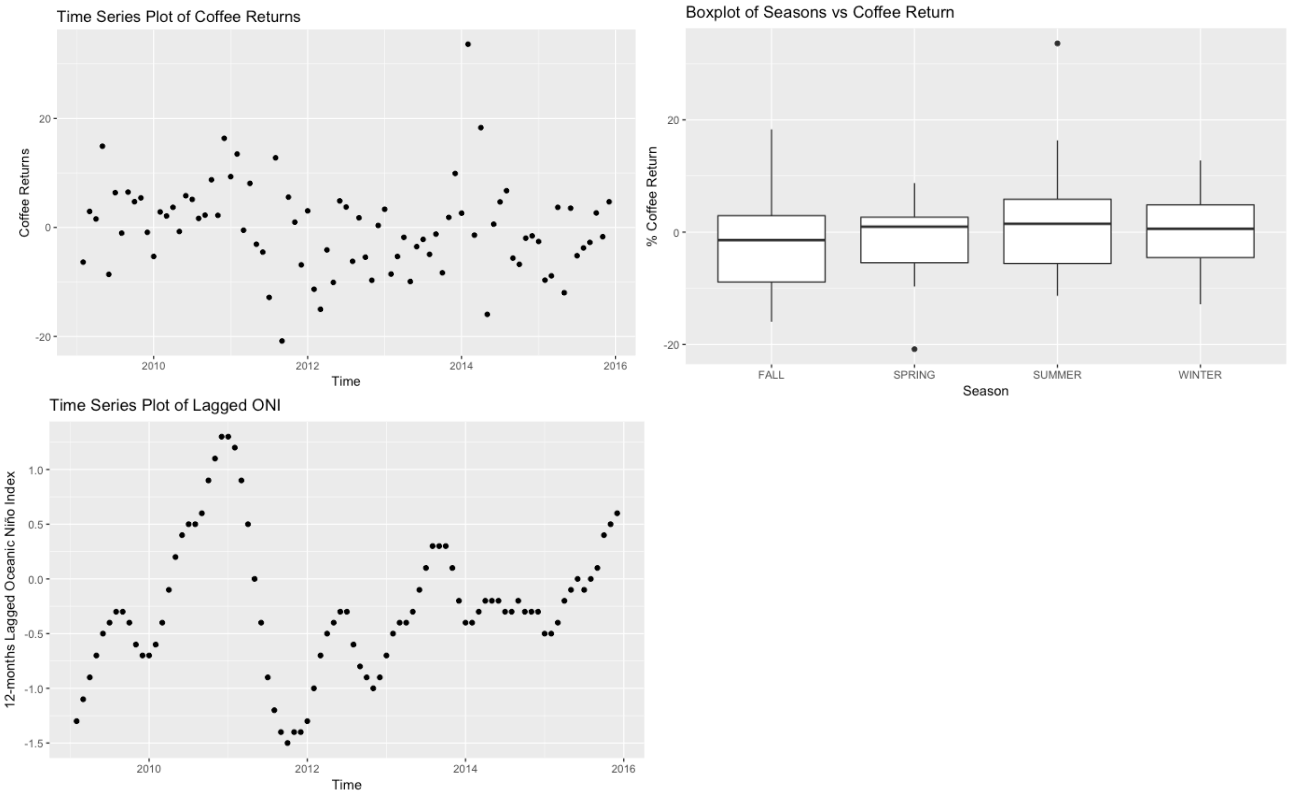
The predictors were selected to capture the relationship of coffee bean returns with climate, exchange rates and production. Climate is of particular interest due to the growing concerns about the impact of climate change on the coffee bean supply chain. The 12-month lagged variables aligns with when the berries of the current batch of coffee were growing. That is when coffee is most sensitive to the environment and defections could hurt yield during harvest. The current month temperature and rainfall predictors may reflect concerns about the quality of future batches, in which case, buyers may attempt to adjust inventory now and potentially influence price and returns.

2. Identifying and addressing autocorrelation:

The initial response variable was Brazilian arabica coffee bean prices. As with most price data, however, the price of coffee beans is likely a non-stationary time series. This means coffee bean prices may follow a random walk where mean and variance changes over time, making price difficult to model. An ACF plot of a regression of coffee prices confirms this:

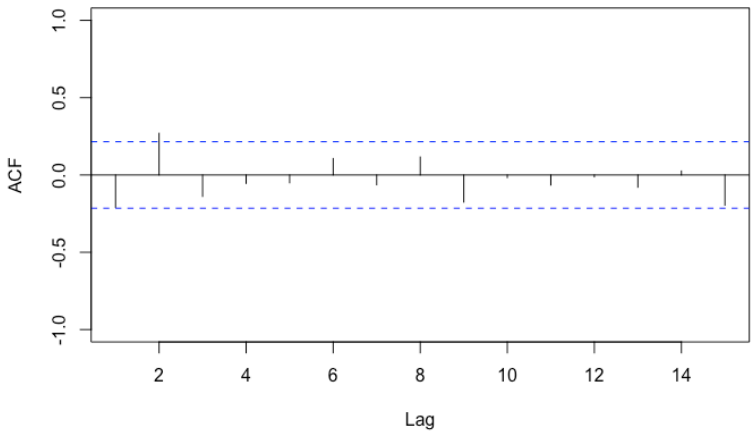


The ACF plot shows clear autocorrelation in the data and the slow decay supports the need to difference the response variable. Taking the log difference, the response variable is approximately transformed into returns, and relevant predictors like production quantity were also transformed into percentage change for better interpretations of the model. The time series plots are shown below:



Looking at the time series, there does not appear to be any broad trends that can be accounted for with a time predictor. In a way, detrending was already performed when the prices were adjusted for inflation. There does not seem to be a strong seasonal effect as the returns do not vary much across seasons. However, there appears to be some cyclical trend every two years in the returns time series. A possible explanation is the El Niño Southern Oscillation. Every few years, regions like Brazil are exposed to a few months of warmer climates (El Niño phase) followed by a few months of cooler climates (La Nina phase). Moreover, most of Brazil's arabica is produced in Minas Gerais, which is located near the coast and is heavily affected by this phenomenon. The Lagged_ONI predictor may help deseaonalize the time series. Lastly, despite the lack of any clear seasonal effects in the boxplot, the analysis will continue to include seasons predictors in the regressions. Seasons could be meaningful after unusual observations are dealt with because

planting seeds during the summer and harvesting during the winter are both critical periods for coffee production. After applying the corrections, the tests for autocorrelation in the full regression model is shown below:

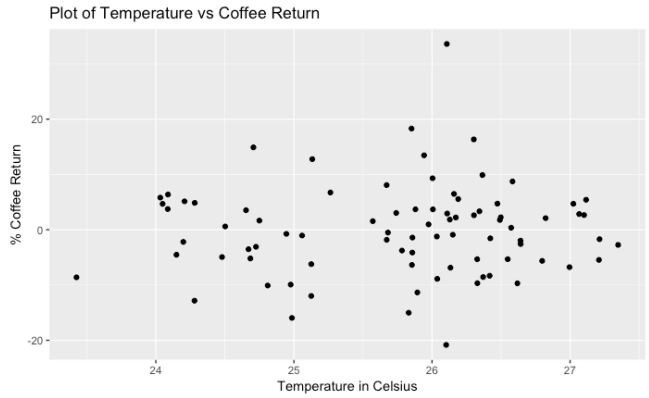


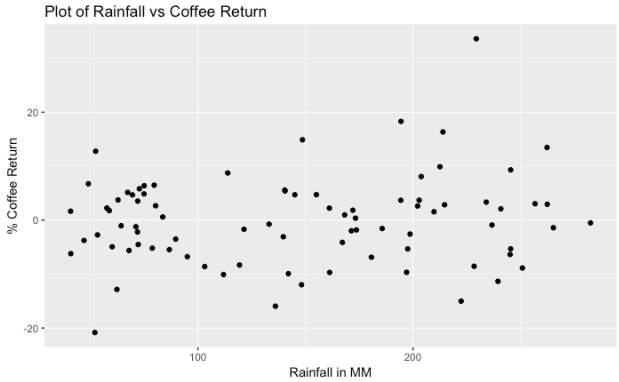

```
Runs Test
data:  std.resf
statistic = 0.88896, runs = 46, n1 = 41, n2 = 41, n = 82, p-value = 0.374
alternative hypothesis: nonrandomness


Durbin-Watson test
data:  diff_fit
DW = 2.391, p-value = 0.9134
alternative hypothesis: true autocorrelation is greater than 0
```

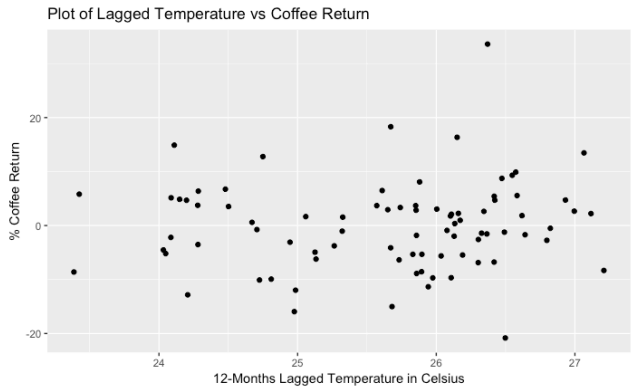
The ACF plot no longer shows strong autocorrelation values or any decaying pattern. However, lag 2 autocorrelation is above the 95% confidence band for zero autocorrelation and some of the autocorrelation values such as lag 1 and lag 9 are quite high. The runs test has p-value = 0.374, which is statistically not significant and the null hypothesis of no autocorrelation in the error is not rejected. The Durbin-Watson test is also highly not significant and the null hypothesis that the errors are normally distributed and uncorrelated is not rejected. In general, it seems like the corrections were largely successful in accounting for autocorrelation. The other diagnostic plots for the initial full model are examined in the next section.

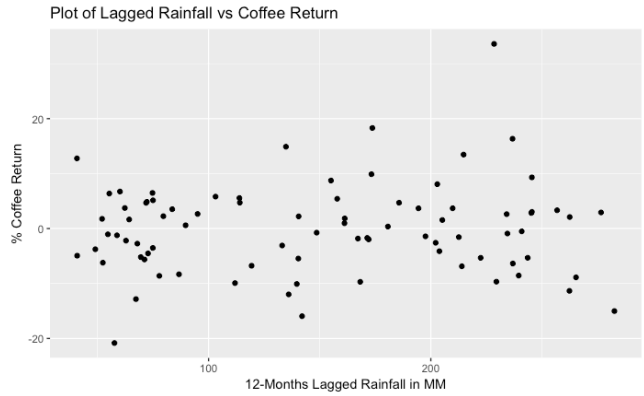
3. Identifying and addressing unusual observations:

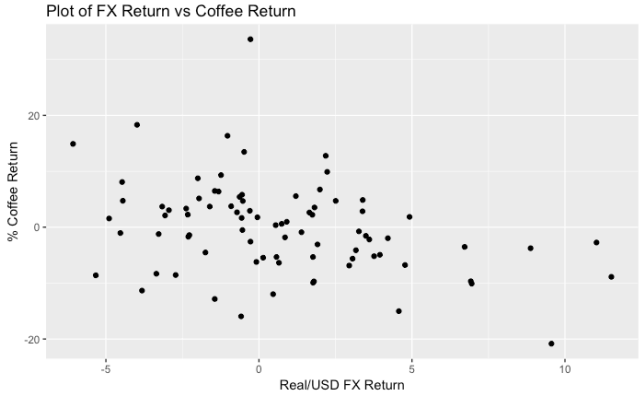
The scatterplots for each predictor against coffee bean returns are shown below:

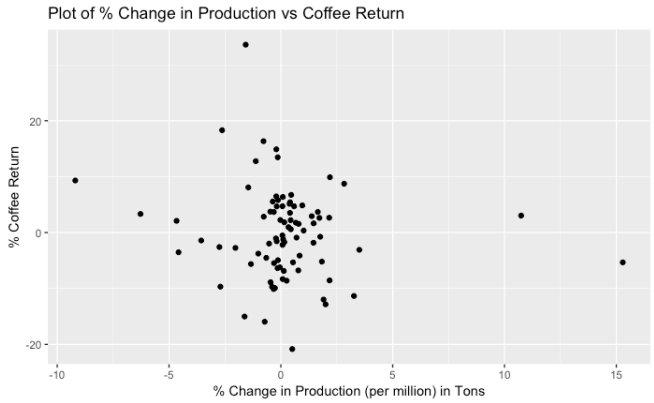


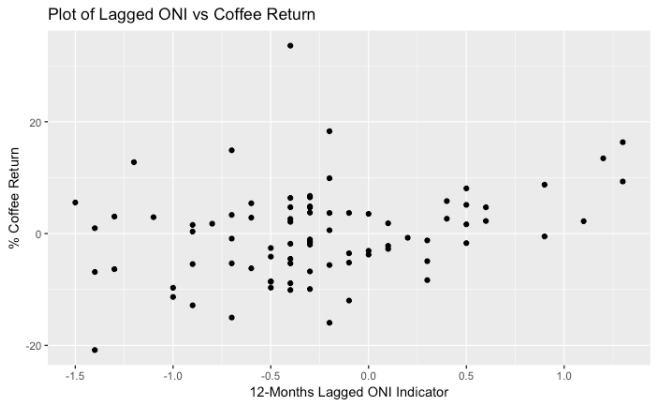






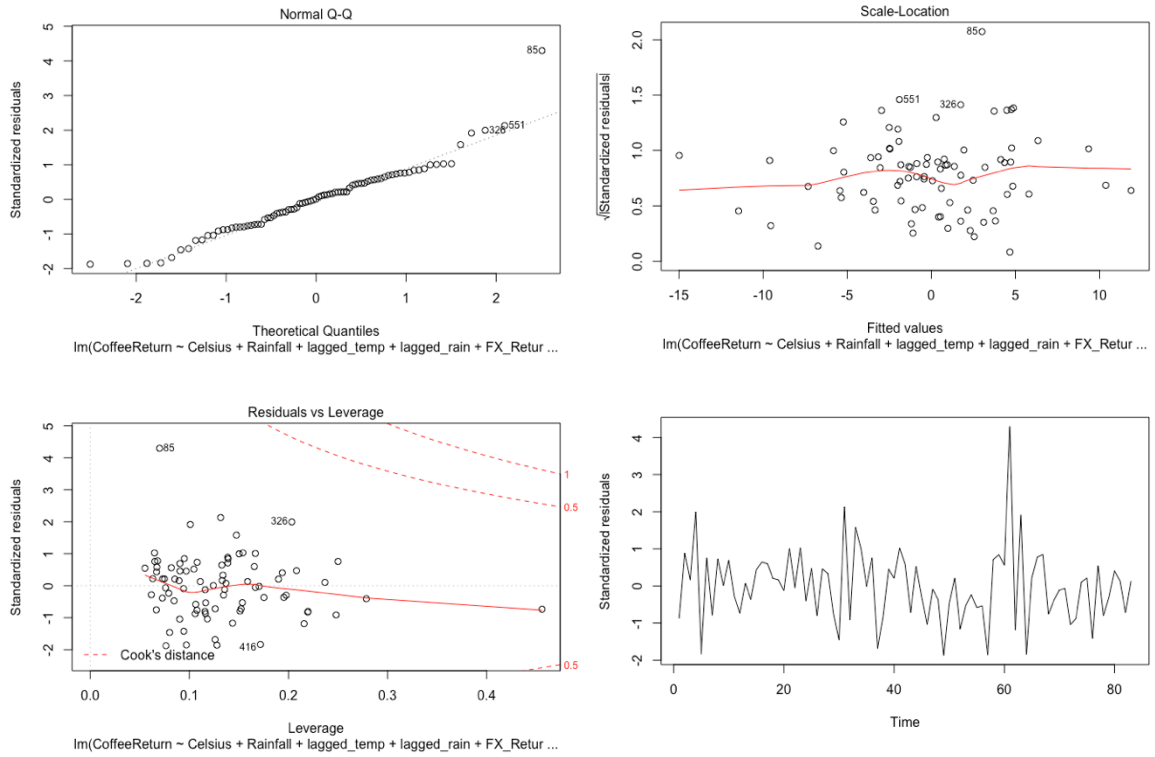






The scatterplots show fairly weak relationships between the predictors and coffee returns. The only two predictors that seems to show a meaningful relationship are FX_Returns and Lagged_ONI. The exchange rate return suggests a negative relationship. That makes sense because an increase in the exchange rate for USD to Real means a weaker Real. From the perspective of a US buyer who purchase Brazilian coffee in Real, a fall in Real would mean their coffee is worth less in US dollars. Hence, we would expect the return on coffee beans in USD to fall. Lagged_ONI seems to show a slight positive relationship, which also makes sense as a large positive ONI value indicates an El Niño event. Hot weathers can severely impact coffee beans and create shortages which triggers price increases and positive returns. There is a clear outlier at around coffee return = 27%. There are also clear leverage points in Production_Change variable. Overall, there is no obvious non-constant variance in any of variables that would warrant log transformations. The output and diagnostics of the full regression is shown below:

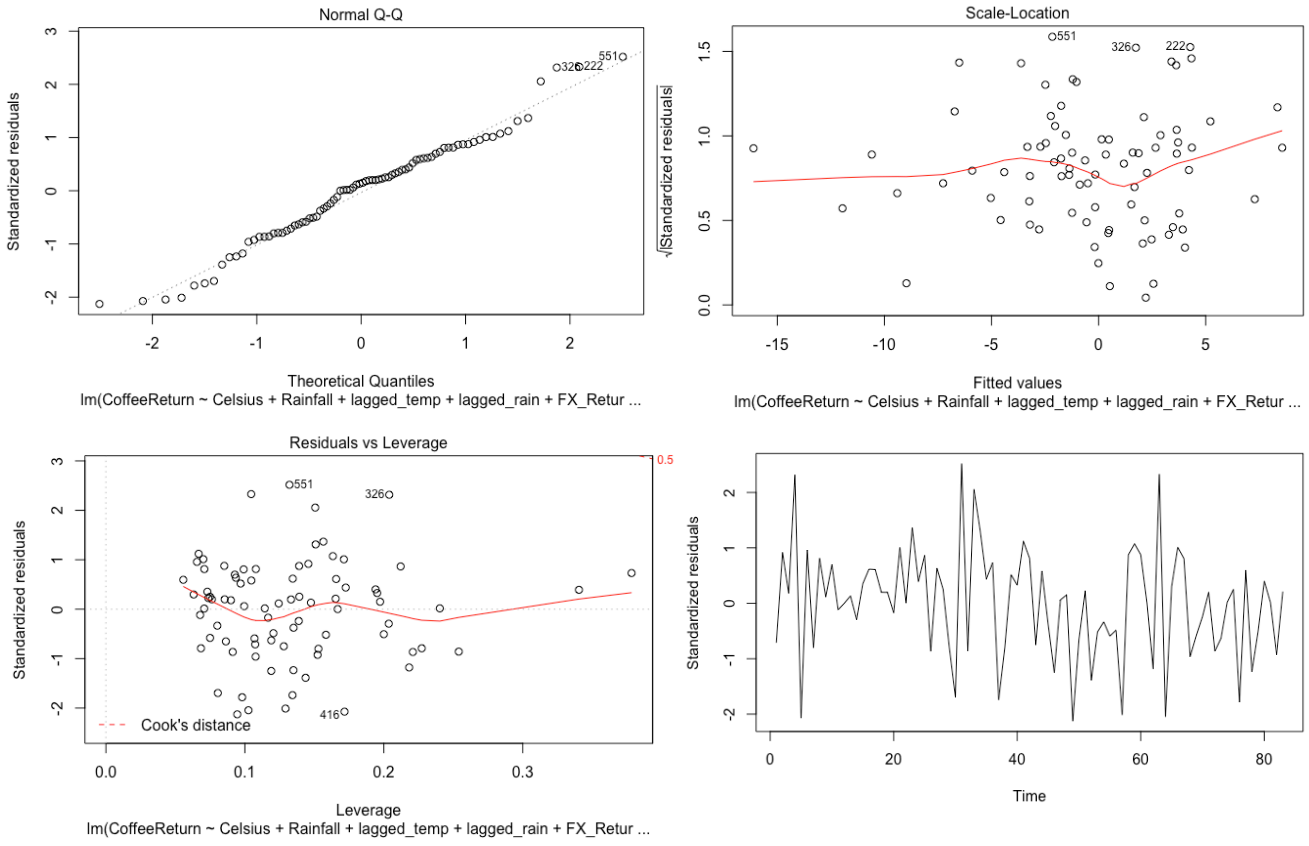
```
lm(formula = CoffeeReturn ~ Celsius + Rainfall + lagged_temp +
    lagged_rain + FX_Return + ProdChange + lagged_oni + Summer +
    Winter + Spring, data = train_diff)

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) -47.14465   52.37994  -0.900  0.37109
Celsius       4.41792    2.53771   1.741  0.08597 .
Rainfall      0.09005    0.04688   1.921  0.05871 .
lagged_temp  -2.69720    2.53786  -1.063  0.29143
lagged_rain  -0.07708    0.04545  -1.696  0.09419 .
FX_Return    -0.73100    0.24086  -3.035  0.00335 **
ProdChange   -0.30780    0.30068  -1.024  0.30941
lagged_oni    3.89309    1.44783   2.689  0.00890 **
Summer        3.76015    2.75845   1.363  0.17709
Winter        5.30611    4.00596   1.325  0.18951
Spring        0.65292    3.91421   0.167  0.86799

Residual standard error: 7.383 on 72 degrees of freedom
Multiple R-squared:  0.2969,    Adjusted R-squared:  0.1993
F-statistic: 3.041 on 10 and 72 DF,  p-value: 0.00289
```



In the normality plot and standardized residuals vs fitted plot, there is one very strong outlier. The observation is February 2014 where coffee return was 33.6%. This was because in January 2014, Brazil was hit with a massive drought that destroyed coffee crops and the damage was not priced in until February. The price shot up due to a fear of shortage and naturally the return spiked. The coffee return for this observation was removed and imputed using linear interpolation. On the standardized residuals vs leverage plot, there is an observation with very high hat value of 0.46, which corresponds to January 2010. Given n = 82 and p = 10, any hat values close to or above 0.34 is suspicious. This observation had unusually high Lagged_ONI value of 1.5 and Production_Change of 15.3% while coffee return was -5.33%. A Jan2010 binary variable was created to mask this leverage point. The full model regression is fitted again:

```
lm(formula = CoffeeReturn ~ Celsius + Rainfall + lagged_temp +
    lagged_rain + FX_Return + ProdChange + lagged_oni + Summer +
    Spring + Winter + Jan2010, data = train_diff2)
```

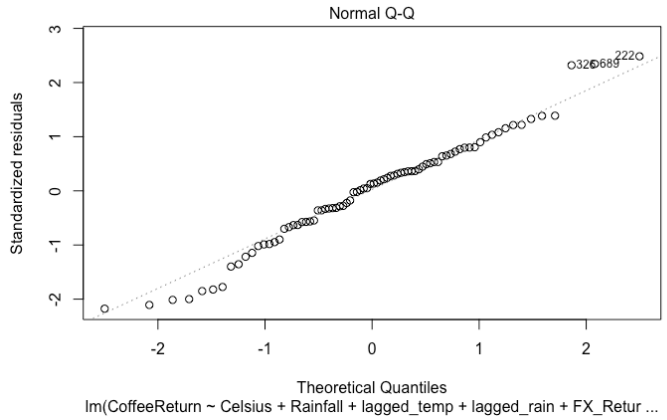
```
Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) -52.88452   45.30849  -1.167 0.247031
Celsius       6.03330    2.28152   2.644 0.010066 *
Rainfall      0.10345    0.04179   2.475 0.015700 *
lagged_temp  -4.10614    2.25187  -1.823 0.072446 .
lagged_rain  -0.08785    0.03986  -2.204 0.030757 *
FX_Return    -0.68258    0.20917  -3.263 0.001695 **
ProdChange    0.03209    0.33318   0.096 0.923550
lagged_oni    4.43925    1.26354   3.513 0.000774 ***
Summer        2.14443    2.38368   0.900 0.371358
Spring        0.49409    3.37623   0.146 0.884064
Winter        5.72724    3.48692   1.642 0.104910
Jan2010      -8.59524    8.63087  -0.996 0.322694

Residual standard error: 6.366 on 71 degrees of freedom
Multiple R-squared:  0.3511,    Adjusted R-squared:  0.2505
F-statistic: 3.492 on 11 and 71 DF,  p-value: 0.0006221
```

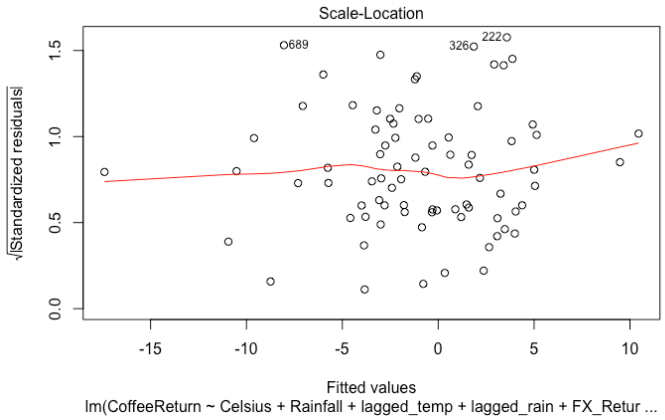


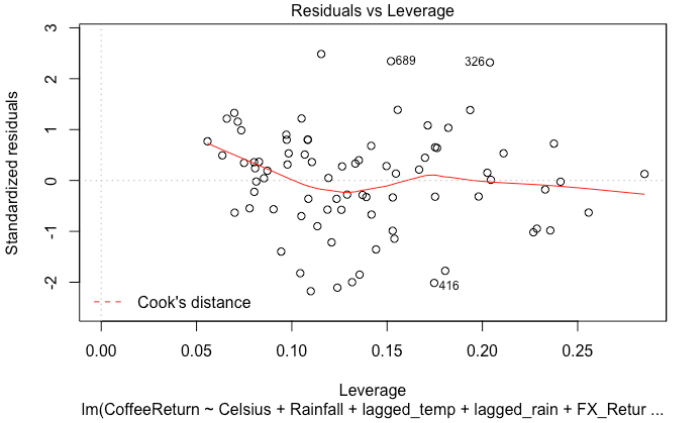
The August 2011 observation is an outlier with standardized residual of 2.52 and a return of 12.77%. There is no clear event that explains why August 2011 generated such high returns, however, the overall sentiment around that period was that growth in Brazil's coffee production was reaching its limit and many were speculating further increases in price. However, this sort of behaviour is very typical in the coffee market and is not a satisfying explanation for why August 2011 is an outlier. The observation is adjusted through linear interpolation, but it is not an outlier and simply reflects the limitations of the model. There were two additional leverage points, January 2011 and January 2010. Given that there are three total leverage points in January, perhaps there is a pattern here. Research, however, shows that both January 2011 and 2010 experienced severe floods and mudslides largely due to poor government infrastructure. Hence, predictors like Temperature and Rainfall could not explain these observations since it was partly poor governance and not weather alone. It seems that any flooding in Brazil since 2011 has not be as devastating. It is reasonable to treat these two months as unusual and unlikely to repeat in the future. They are masked out with binary variables Jan2010 and Jan2011. A final full model is fitted:

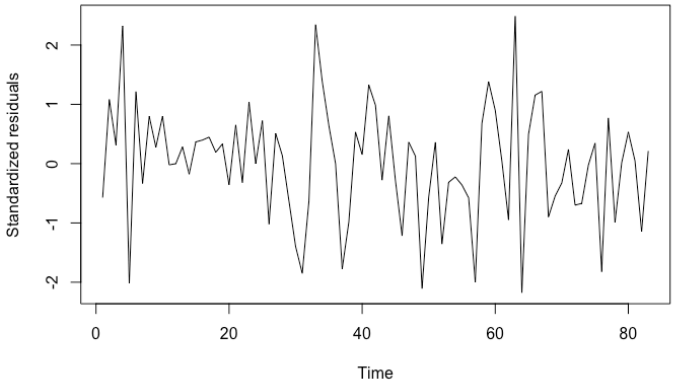
```
lm(formula = CoffeeReturn ~ Celsius + Rainfall + lagged_temp +
    lagged_rain + FX_Return + ProdChange + lagged_oni + Summer +
    Spring + Winter + Jan2010 + Jan2012 + Jan2011, data = train_diff3)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -15.47029   45.42940  -0.341  0.73449
Celsius       4.78222    2.32371   2.058  0.04337 *
Rainfall      0.08878    0.04225   2.101  0.03925 *
lagged_temp  -4.37747    2.24503  -1.950  0.05526 .
lagged_rain  -0.06370    0.03994  -1.595  0.11534
FX_Return    -0.67023    0.20741  -3.232  0.00189 **
ProdChange   -0.02397    0.46532  -0.052  0.95906
lagged_oni    5.66839    1.28627   4.407 3.76e-05 ***
Summer        2.61772    2.42992   1.077  0.28510
Spring        2.63683    3.36963   0.783  0.43658
Winter        3.80693    3.49158   1.090  0.27937
```

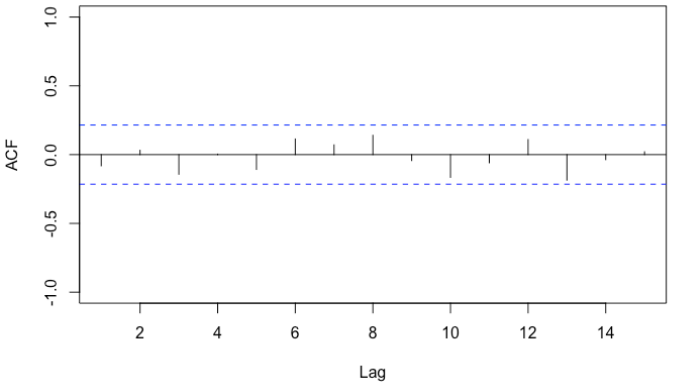
```
Jan2010       -6.83108   10.18782  -0.671  0.50477
Jan2012        5.11541    8.38247   0.610  0.54370
Jan2011       -0.55386    8.13831  -0.068  0.94594

Residual standard error: 6.3 on 69 degrees of freedom
Multiple R-squared:  0.3935,    Adjusted R-squared:  0.2793
F-statistic: 3.444 on 13 and 69 DF,  p-value: 0.0003999
```











```
Runs Test
data:  std.resf
statistic = 0.44448, runs = 44, n1 = 41, n2 = 41, n = 82, p-value = 0.6567
alternative hypothesis: nonrandomness

Durbin-Watson test
data:  diff_fit4
DW = 2.1529, p-value = 0.602
alternative hypothesis: true autocorrelation is greater than 0
```

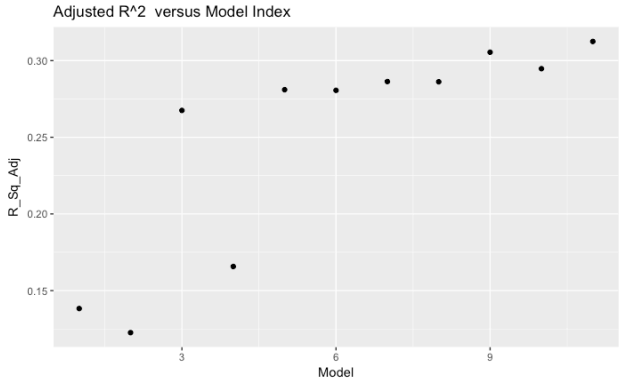
At this point, there are no clear unusual observations for the full model. There is some non-normality in the data which suggests a nonlinear model may fit better. The residuals in the standardized residuals vs fitted plot are evenly spread out which implies constant variance. In the standardized residuals vs fitted plot, there are some observations with high hat values but are within tolerable limits. The standardized residuals vs ordered observations plot seems fairly random; there are high residuals about every 20 months, but the pattern is not obvious. The ACF plot shows low autocorrelation between lagged residuals and both the runs test and Durbin-Watson test are not statistically significant.

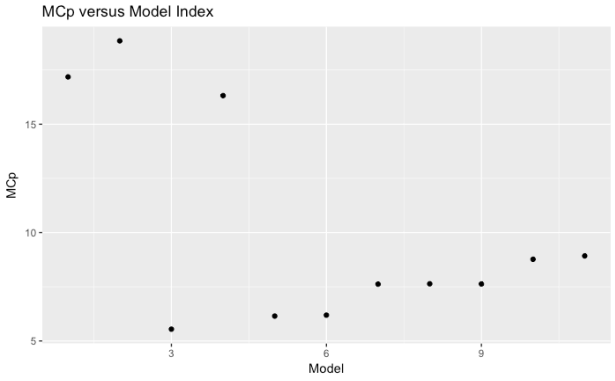
4. Model Selection:

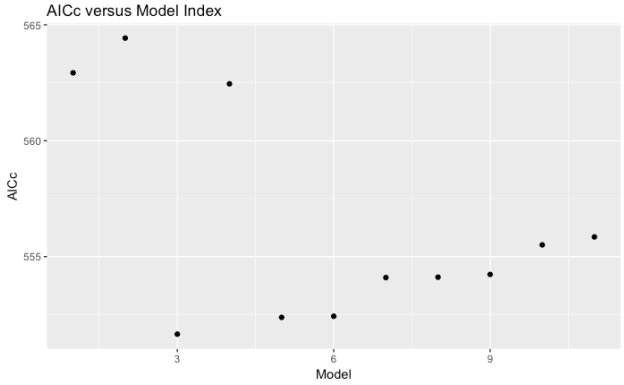
In the full model regression, there are many predictors that were not statistically significant and were removed before model selection. The significant predictors are Temperature, Rainfall, Lagged_Temperature, FX_Return and Lagged_ONI. In addition, Lagged_Rain and Winter were included. Lagged_Rain is very close to statistical significance and may be meaningful in a different model. Winter had high p-values, but the harvesting season in winter tends to be important to coffee farmers and should be further examined. The best subset output is displayed below:

	Jan-10	Jan-12	Jan-11	Temperature	Rainfall	lagged_rain	FX_Return	lagged_oni	Winter
4 (1)	*	*	*					*	
4 (2)	*	*	*				*		
5 (1)	*	*	*				*	*	
5 (2)	*	*	*		*			*	
6 (1)	*	*	*	*			*	*	
6 (2)	*	*	*		*		*	*	
7 (1)	*	*	*	*	*		*	*	
7 (2)	*	*	*	*			*	*	*
8 (1)	*	*	*	*	*		*	*	*
8 (2)	*	*	*	*		*	*	*	
9 (1)	*	*	*	*	*	*	*	*	*

The adjusted R^2, MCp and AICc of each model are compared. All three measures were adjusted for the change in p and n; the best subset function automatically counts the masking indicator variables as predictors and does not subtract those observations from p and n:







For all three measures, the model performance improves rapidly at model three. While the adjusted R^2 continues to improve with additional predictors, MCp and AICc are minimized at model three. Given that the improvements in adjusted R^2 is quite small with more complicated models, the potential best models seem to be model three. Models five and six may also be good candidates. Their regression outputs are displayed below:

Model three, p = 2:

```
lm(formula = CoffeeReturn ~ Jan2010 + Jan2012 + Jan2011 + FX_Return +
    lagged_oni, data = train)

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.8173     0.7843   1.042 0.300666
Jan2010      -1.5918     6.4119  -0.248 0.804602
Jan2012       6.1535     6.5611   0.938 0.351238
Jan2011       1.5275     6.6580   0.229 0.819150
FX_Return    -0.7270     0.1973  -3.686 0.000422 ***
lagged_oni    4.6691     1.1962   3.903 0.000202 ***

Residual standard error: 6.349 on 77 degrees of freedom
Multiple R-squared:  0.3127,    Adjusted R-squared:  0.2681
F-statistic: 7.007 on 5 and 77 DF,  p-value: 1.92e-05

VIF:
FX_Return   lagged_oni
1.025430      1.1393997
```

Model five, p = 3

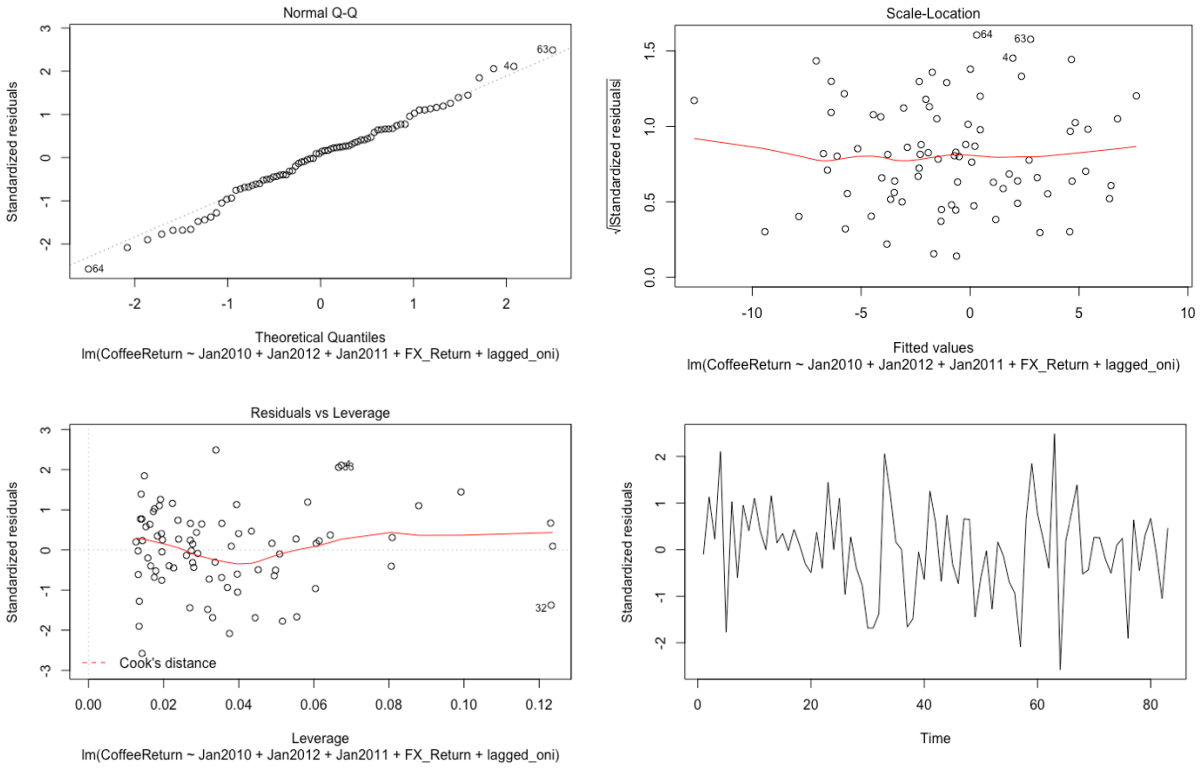
```
lm(formula = CoffeeReturn ~ Jan2010 + Jan2012 + Jan2011 + Celsius +
    FX_Return + lagged_oni, data = train)

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) -22.4346    19.4617  -1.153 0.252624
Jan2010      -2.2885     6.4206  -0.356 0.722503
Jan2012       6.0873     6.5431   0.930 0.355141
Jan2011       1.1697     6.6462   0.176 0.860768
Temperature   0.9048     0.7567   1.196 0.235525
FX_Return    -0.7530     0.1979  -3.805 0.000285 ***
lagged_oni    4.7067     1.1933   3.944 0.000177 ***
```

```
Residual standard error: 6.332 on 76 degrees of freedom
Multiple R-squared:  0.3254,    Adjusted R-squared:  0.2722
F-statistic:  6.11 on 6 and 76 DF,  p-value: 2.927e-05

VIF:
Temperature   FX_Return   lagged_oni
1.023979      1.038001    1.140189
```

Model six, p = 4

```
lm(formula = CoffeeReturn ~ Jan2010 + Jan2012 + Jan2011 + Celsius +
    FX_Return + lagged_oni + Winter, data = train)

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) -40.5483    31.4142  -1.291 0.200750
Jan2010      -2.3573     6.4408  -0.366 0.715397
Jan2012       6.4118     6.5777   0.975 0.332801
Jan2011       1.5303     6.6843   0.229 0.819537
Temperature   1.5896     1.2008   1.324 0.189596
FX_Return    -0.7785     0.2015  -3.864 0.000236 ***
lagged_oni    4.6401     1.2003   3.866 0.000234 ***
Winter        1.8836     2.5595   0.736 0.464061

Residual standard error: 6.351 on 75 degrees of freedom
Multiple R-squared:  0.3302,    Adjusted R-squared:  0.2677
F-statistic: 5.283 on 7 and 75 DF,  p-value: 6.158e-05

VIF:
Temperature   FX_Return   lagged_oni  Winter
2.563104      1.069433    1.146689    2.547989
```

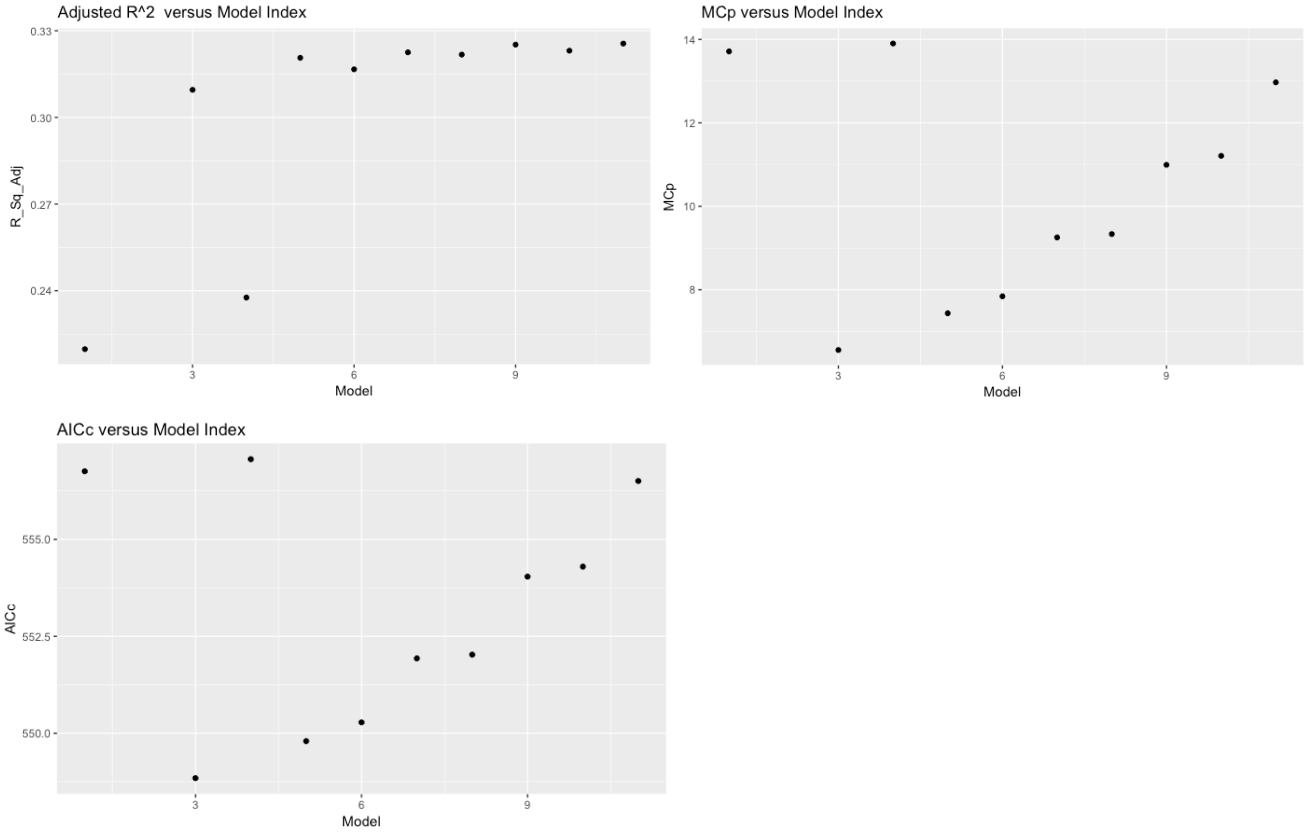
All three models have similar adjusted R^2 of about 27% which is quite low. More importantly, models five and six both have predictors with high p-values that are not statistically significant, which further promotes model three as the best model. All predictors in model three are statistically significant (excluding the binary variables to mask leverage points) and the VIFs are very low which indicates high stability and generalizability. The diagnostics for the model three are shown below:



Overall, the model looks fairly consistent with the assumptions of linear regression. The residuals in the normality plot appears to follow normal distribution and there are no standardized residuals greater than 2.5. On the standardized residuals vs leverage plot, however, there are three leverage points. These observations are September 2011, March 2015 and September 2015. It is not clear which predictors or combination of predictors contributed to their high hat values and whether these observations are truly unusual. However, they have high influence on the regression and should be masked with indicator variables. In a second fit of the model, another leverage point was identified, likely masked by the previous ones. This observation, August 2015, was also removed through an indicator variable. Since unusual

observations were identified and effectively removed, the model selection process was repeated again. The new best subset models and their performance measures are displayed below:

	Jan-10	Jan-12	Jan-11	Sep-11	Mar-15	Sep-15	Aug-15	Temperature	Rainfall	lagged_rain	FX_Return	lagged_oni	Winter
8 (1)	*	*	*	*	*	*	*					*	
8 (2)	*	*	*	*	*	*	*				*		
9 (1)	*	*	*	*	*	*	*				*	*	
9 (2)	*	*	*	*	*	*	*		*			*	
10 (1)	*	*	*	*	*	*	*		*		*	*	
10 (2)	*	*	*	*	*	*	*			*	*	*	
11 (1)	*	*	*	*	*	*	*		*	*	*	*	
11 (2)	*	*	*	*	*	*	*	*	*		*	*	
12 (1)	*	*	*	*	*	*	*	*	*	*	*	*	
12 (2)	*	*	*	*	*	*	*		*	*	*	*	*
13 (1)	*	*	*	*	*	*	*	*	*	*	*	*	*



The performance measures are similar to the previous best subset. Again, model three looks like the best model while model five and six seem to overfit but are worth examining. The outputs of the three regressions are shown below:

Model three, p = 2:

```
lm(formula = CoffeeReturn ~ Jan2010 + Jan2012 + Jan2011 + Sep2011 +
    Mar2015 + Sep2015 + Aug2015 + FX_Return + lagged_oni, data = train3)

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.0856     0.7761   1.399 0.166112
Jan2010      -2.0844     6.2915  -0.331 0.741366
Jan2012       5.3393     6.4422   0.829 0.409925
Jan2011       1.7182     6.5352   0.263 0.793353
Sep2011      -8.7559     6.7654  -1.294 0.199668
Mar2015       0.3637     6.8287   0.053 0.957664
Sep2015       3.9907     6.8093   0.586 0.559635
Aug2015       1.7872     6.6144   0.270 0.787772
FX_Return    -0.7478     0.2427  -3.081 0.002908 **
lagged_oni    4.2962     1.2025   3.573 0.000631 ***

Residual standard error: 6.219 on 73 degrees of freedom
Multiple R-squared:  0.3562,    Adjusted R-squared:  0.2768
F-statistic: 4.488 on 9 and 73 DF,  p-value: 0.0001044

VIF:
FX_Return   lagged_oni
1.618151      1.200197
```

Model five, p = 3:

```
lm(formula = CoffeeReturn ~ Jan2010 + Jan2012 + Jan2011 + Sep2011 +
    Mar2015 + Sep2015 + Aug2015 + Rainfall + FX_Return + lagged_oni,
    data = train3)

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.56005    1.70511  -0.328 0.743521
Jan2010     -3.23142    6.37252  -0.507 0.613644
Jan2012      4.27315    6.50937   0.656 0.513620
Jan2011      0.33645    6.65081   0.051 0.959794
```

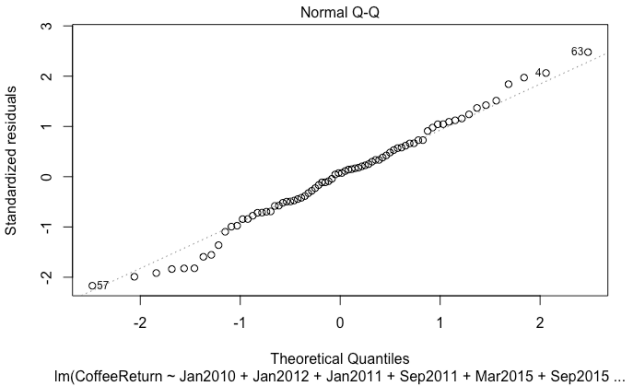
```
Sep2011     -7.74176    6.82182  -1.135 0.260201
Mar2015     -1.14167    6.96059  -0.164 0.870175
Sep2015      4.71989    6.83438   0.691 0.492032
Aug2015      2.65973    6.65539   0.400 0.690608
Rainfall     0.01164    0.01074   1.084 0.282183
FX_Return   -0.72234    0.24355  -2.966 0.004093 **
lagged_oni   4.45168    1.20965   3.680 0.000447 ***

Residual standard error: 6.212 on 72 degrees of freedom
Multiple R-squared:  0.3665,    Adjusted R-squared:  0.2786
F-statistic: 4.166 on 10 and 72 DF,  p-value: 0.0001458

VIF:
Rainfall   FX_Return   lagged_oni
1.233057   1.633344     1.217332
```

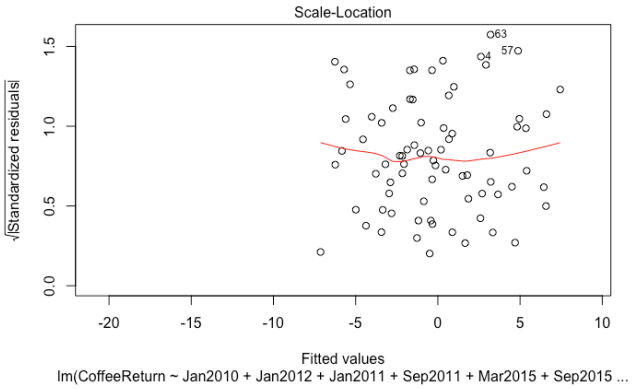
Model six, p = 4:

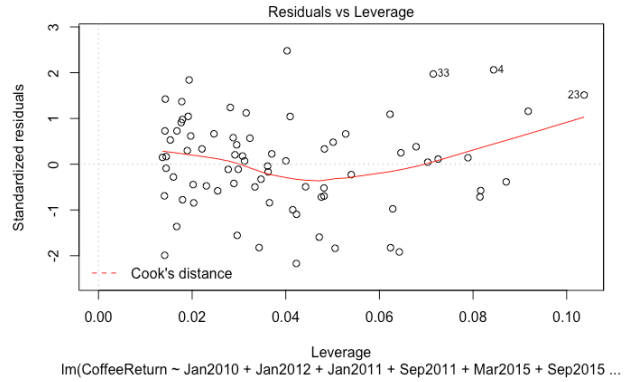
```
lm(formula = CoffeeReturn ~ Jan2010 + Jan2012 + Jan2011 + Sep2011 +
    Mar2015 + Sep2015 + Aug2015 + Rainfall + lagged_rain + FX_Return +
    lagged_oni, data = train3)

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.48316    1.72326  -0.280 0.780004
Jan2010     -3.32387    6.41156  -0.518 0.605779
Jan2012      4.14722    6.55190   0.633 0.528780
Jan2011      0.41149    6.69018   0.062 0.951129
Sep2011     -8.00568    6.88538  -1.163 0.248842
Mar2015     -1.23418    7.00265  -0.176 0.860603
Sep2015      4.59742    6.87813   0.668 0.506039
Aug2015      2.41026    6.71589   0.359 0.720743
Rainfall     0.02508    0.03195   0.785 0.435045
lagged_rain -0.01376    0.03078  -0.447 0.656130
FX_Return   -0.69810    0.25084  -2.783 0.006895 **
lagged_oni   4.41709    1.21889   3.624 0.000542 ***
```

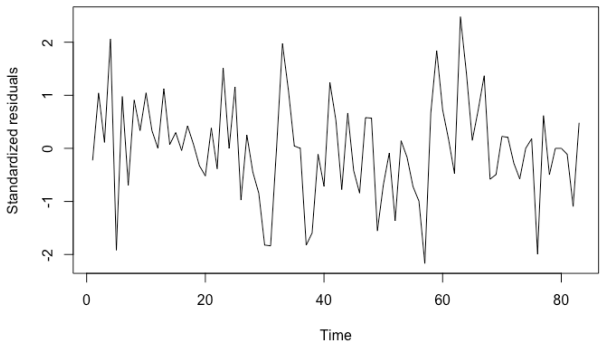
```
Residual standard error: 6.246 on 71 degrees of freedom
Multiple R-squared:  0.3683,    Adjusted R-squared:  0.2705
F-statistic: 3.763 on 11 and 71 DF,  p-value: 0.0002913
VIF:
Temperature    FX_Return    lagged_oni    Winter
2.563104       1.069433     1.146698    2.547989
```

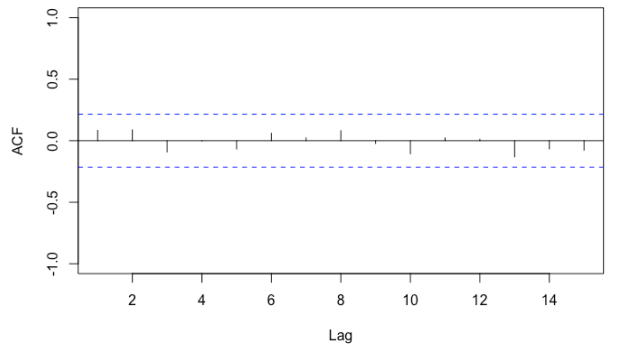
It is very clear that model three is the best model; it has practically the same adjusted R^2 to the other two and all of its predictors are highly significant. Moreover, its simplicity means it will likely generalize better for out of sample predictions. The diagnostic plots are shown below:

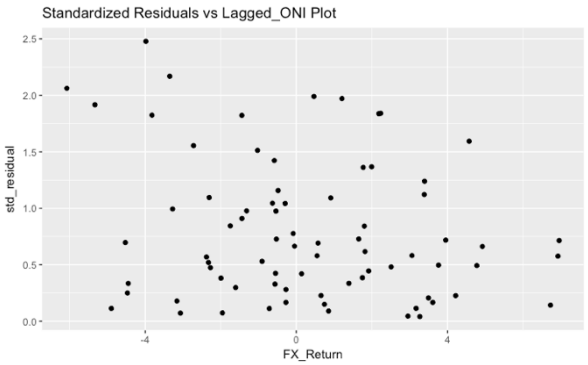


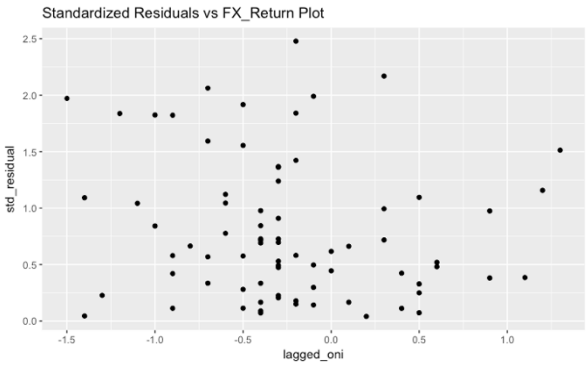












The best model looks consistent with the assumptions of linear regression. There is some nonnormality but the true relationship that explains coffee returns is likely nonlinear. There are no unusual observations and the ordered residuals do not display any clear patterns. Furthermore, the ACF plot suggests autocorrelation has been mostly corrected, and both the runs test and Durbin-Watson test are not statistically significant. Lastly, the plots of standardized residuals vs predictors both appear randomly distributed. For Lagged_ONI, there seems to be a slight concentration of smaller residuals at higher Lagged_ONI values.

5. Interpreting the best model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9$$

$$y = 1.09 - 0.75 x_1 + 4.30 x_2 - 2.08 x_3 + 1.72 x_5 + 5.34 x_4 - 8.76 x_6 + 0.36 x_7 + 3.99 x_8 + 1.79 x_9$$

x_1: Percentage return of USD to Real currency exchange
x_2: Oceanic Niño Index
x_3: Indicator variable for January 2010
x_4: Indicator variable for January 2011
x_5: Indicator variable for January 2012
x_6: Indicator variable for September 2011
x_7: Indicator variable for March 2015
x_8: Indicator variable for September 2015
x_9: Indicator variable for August 2015

The intercept β_0 represents the return on coffee beans for a month where the return on USD to Real exchange rate is zero, the ONI measure is zero and is not any of the indicator months. Since the exchange rate usually changes between months, there is no practical meaning to the intercept. β_1 = -0.75 means that a one percentage point change in the return on USD to Real currency exchange is associated with a -0.75 percentage point change in the return on Brazilian arabica coffee beans. β_2 = 4.30 means that a one point change in the ONI measure is associated with a 4.30 percentage point change in the return on Brazilian arabica coffee beans Overall, the model is a poor predictor of coffee bean returns. The adjusted R^2 is only 27.68% and the prediction intervals are wide. Validated against the held out 2016 data, the predictions and 95% prediction intervals are:

	fit	lwr	upr	CoffeeReturn
1	4.91543	-7.7359	17.56675	-2.756756
2	8.68841	-4.359	21.73582	11.257125
3	6.81484	-5.98375	19.61343	-3.690342
4	5.00707	-7.72261	17.73675	-3.933633
5	7.85094	-5.06799	20.76986	20.551712
6	9.41459	-3.6793	22.50848	-1.862258
7	8.68552	-4.4302	21.80124	-1.17909
8	7.34964	-5.99956	20.69883	2.252903
9	11.2363	-2.37361	24.84623	8.13883
10	6.98067	-7.0443	21.00563	-5.940641
11	10.5839	-3.31547	24.48328	-5.658214

Moreover, this wide interval needs to be further adjusted for post selection inference:

$$\tilde{s} = s\sqrt{\frac{n-p-1}{n-p*-1}} = 6.219\sqrt{\frac{75-2-1}{75-6-1}} = 6.399$$

Hence, the 95% prediction interval is approximately ± (2)(6.399) = ± 12.798. After the adjustment, the model predicts with 95% probability that the return for the first value in the validation set lies within (-7.88257, 17.71343), which is only about 1.2% larger than the unadjusted prediction interval. The biggest problem with the wide prediction interval is that the lower bound is negative and the upper bound is positive; that means the model cannot even provide information on the direction of coffee returns and prices.

6. Conclusion:

The best model does not explain much of Brazilian arabica coffee bean returns and is a very poor predictor. It did successfully account for the majority of autocorrelation; hence, the biggest problem is that meaningful predictors were not selected for the analysis. Possible predictors could be monthly unemployment rate, since coffee production is a high labour-intensive process. However, the methodology for surveying unemployment in Brazil changed in 2012 and the climate data used in this analysis only extends to 2016. Hence, including unemployment would have a produced a very small dataset. Furthermore, many of the leverage points that were masked out contained valuable information about coffee production in Brazil. In particular, several of them corresponded to poor infrastructure and governance during natural crises. So perhaps, government risk needs to be measured and included into the analysis. Another key factor that was ignored was speculation in coffee futures market. Speculation, particularly algorithmic trading, is suspected to be a major contributor to volatility in coffee prices. That relationship, however, is probably circular in that spot prices also influence future prices and is difficult to isolate the effect of speculation alone. Finally, the model did show statistical significance with Lagged_ONI and a positive coefficient, which supports concerns over climate change and the possibility of warmer weathers driving up prices.

Data:

1) "CEPEA/ESALQ Arabica Coffee Price Index." *Center for Advanced Studies on Applied Economics*, https://www.cepea.esalq.usp.br/en/indicator/coffee.aspx.
2) "Climate Change Knowledge Portal." *The World Bank Group*, https://climateknowledgeportal.worldbank.org/download-data.
3) *Brazilian Institute of Geography and Statistics*, https://sidra.ibge.gov.br/tabela/6588.
4) "Brazil / USD Foreign Exchange Rate." *Federal Reserve Economic Data*, https://fred.stlouisfed.org/series/DEXBZUS.
4) "Cold & Warm Episodes by Season." *Climate Prediction Center NOAA*, https://origin.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ONI_v5.php.
5) "Consumer Price Index for All Urban Consumers: All Items." *Federal Reserve Economic Data*, https://fred.stlouisfed.org/series/CPIAUCSL.