W271 Lab 3

Brandon Shurick, Alejandro J. Rojas, Olivier Zimmer

August 13 2016

1 *Forecast the Web Search Activity for Flight Demand*

The file, google*_correlate_flight.csv*, contains the relative web search activity for the phrase *flight prices* over time.

Your data science team believes that search activity for this phrase is positively correlated with consumer demand for flying (and possibly prices).

Your task is to forecast the relative demand of this phrase for the year 2016. For the purposes of this assignment, ignore the units of the data as they are not relevant here.

Remember to explain and justify each of your modelling decisions. Also, comment on your forecast. Do you notice anything interesting? Do you notice anything worth worrying about?

1.1 Visual inspection of Time Series: *Flight Prices* Web Search

Exploring time series we conclude:

- Period includes 625 weeks from 2004 to 2016
- No clear yearly trend but data looks seasonal with peaks in the beginning of the year declining at a constant rate as the year unfolds.
- Histogram shows that data is skewed and does not follow a normal distribution
- Autocorrelation function shows a strong but declining correlation function as lags go from 1 to 12 weeks

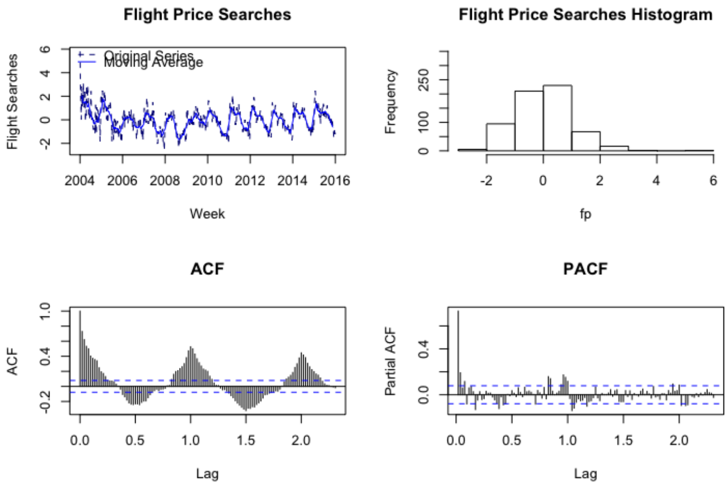


Figure 1: Exploratory visuals of Time Series

- Partial autocorrelations seems to suggest an autocorrelation of order 2, meaning that autocorrelation to the prior two weeks seems to determine the behavior of the time series

1.2 Autocorreation model: *No convergence*

Running the autocorrelation function, we find that model does not converge:

- Confirms model is not stationary
- Leads us to look at how to remove seasonality effect to make it stationary

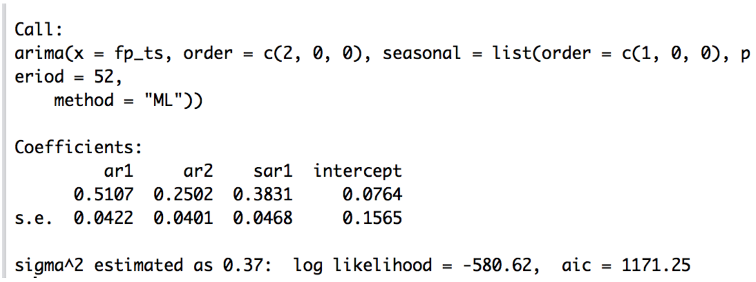
1.3 Autocorreation model with seasonality: *Initial guess*

Given that:

- From ACF and PACF we suspect an AR(2) could work but need to take into account seasonality
- Visually we can see that seasonality runs on a 52 week period, and we initially guess and seasonal AR(1)

We run the following model:

and plot its residuals results:



```
Call:
arima(x = fp_ts, order = c(2, 0, 0), seasonal = list(order = c(1, 0, 0), p
eriod = 52,
    method = "ML"))

Coefficients:
         ar1     ar2    sar1  intercept
      0.5107  0.2502  0.3831     0.0764
s.e.  0.0422  0.0401  0.0468     0.1565

sigma^2 estimated as 0.37:  log likelihood = -580.62,  aic = 1171.25
```

Figure 2: (2,0,0) (1,0,0) Initial model

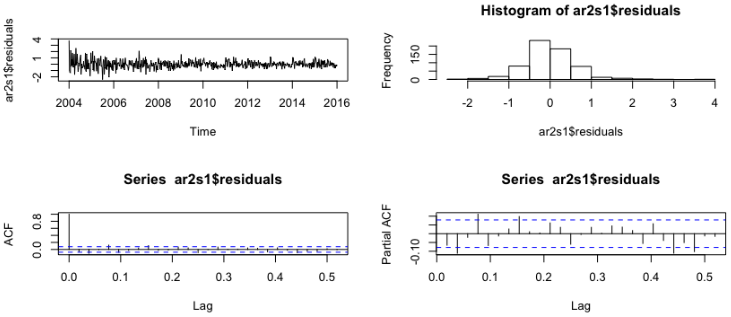


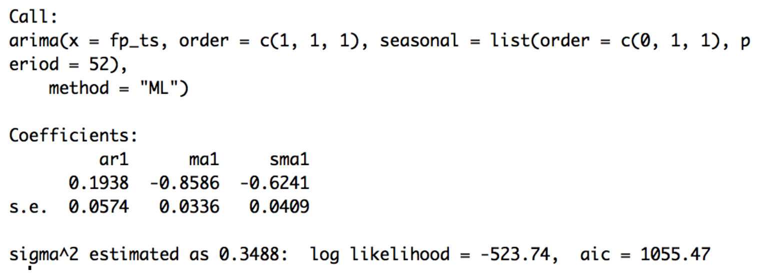
Figure 3: (2,0,0) (1,0,0) Initial model Residuals Plots

From that model we learn that:

- As suspected coefficients for AR1, AR2 and SAR1 are all significant showing that we're making some progress in understanding the mechanics behind the time series
- However, the intercept is not meaning that there are still some trends in the overall mean of the time series that we will need to remove. This suggest that we still need to make the series stationary probably by integrating it
- Looking at the residuals show higher volatility at the early years of the series so we know that variance is not constant across the whole period of the time series
- Eyeballing the residuals graph also seem to suggest that our initial model still breaks the zero conditional mean rule
- This leads us to look at a differentiated model so that underlying trend can be removed

1.4 Arima model: *Making time series stationary*

To remove trend, we created a new model with integration I(1) and we also include Moving Average MA(1) and AR(1). We include a seasonality effect also with integration I(1) and MA(1). Here's the model results:



```
Call:
arima(x = fp_ts, order = c(1, 1, 1), seasonal = list(order = c(0, 1, 1), p
eriod = 52),
    method = "ML")

Coefficients:
         ar1      ma1     sma1
      0.1938  -0.8586  -0.6241
s.e.  0.0574   0.0336   0.0409

sigma^2 estimated as 0.3488:  log likelihood = -523.74,  aic = 1055.47
```

Figure 4: (1,1,1) (0,1,1) Arima model

and its residuals plots:

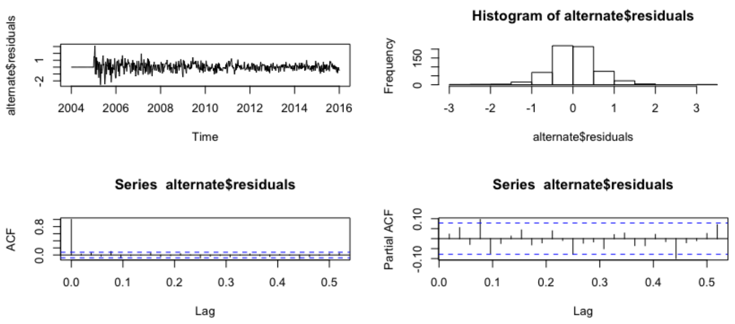


Figure 5: (1,1,1) (0,1,1) Arima model Residuals Plots

This Arima model looks better than the one discussed above:

- The time series plot now look closer to white noise. There are still some problems during the initial period up to 2008 where we still find higher variance and not compliance to the zero conditional mean.
- However, the histogram of the residuals now look more normal

- All coefficients show high significant values
- ACF and PACF look similar to what you would expect from white noise
- The model's AIC is lower that the previous model AIC

1.5 2016 Forecast:

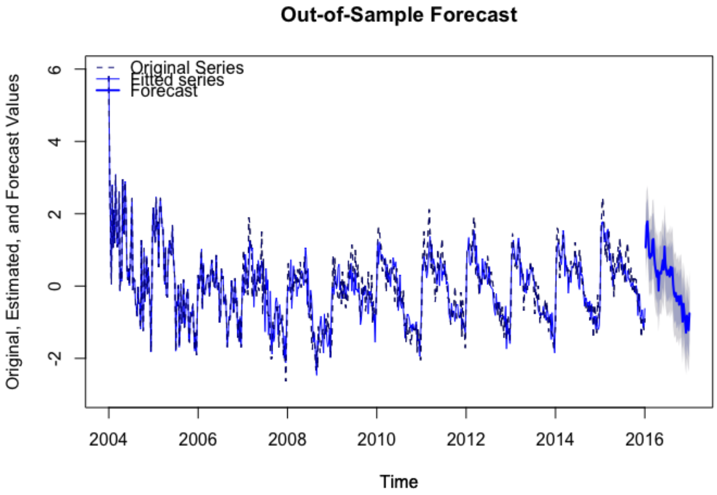


Figure 6: (1,1,1) (0,1,1) Arima Model Out-of-sample Forecast

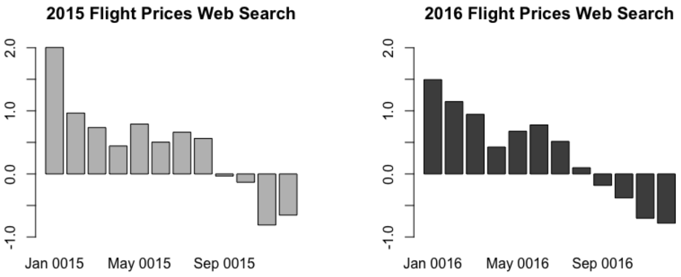


Figure 7: Actual vs Predicted Monthly Averages for Flight Prices Web Search

Comparing monthly averages of 2015 with 2016, we foresee a significant drop in demand in January of 2016 though we expect better performance for the next two months. The model is also predicting worse market conditions for the end of 2016 compared to 2015. It seems in general that market conditions are getting worse so that we may have to be ready to be more aggressive on pricing.

2 *Forecast Inflation-Adjusted Gas Price*

During 2013 amid high gas prices, the Associated Press (AP) published an article about the U.S. inflation adjusted price of gasoline and U.S. oil production. The article claims that there is "evidence of no statistical correlation" between oil production and gas prices. The data was not made publicly available, but comparable data was created using data from the Energy Information Administration. The workspace and data frame *gasOil.Rdata* contains the U.S. oil production (in millions of barrels of oil) and the inflation-adjusted average gas prices (in dollars) over the date range the article indicates.

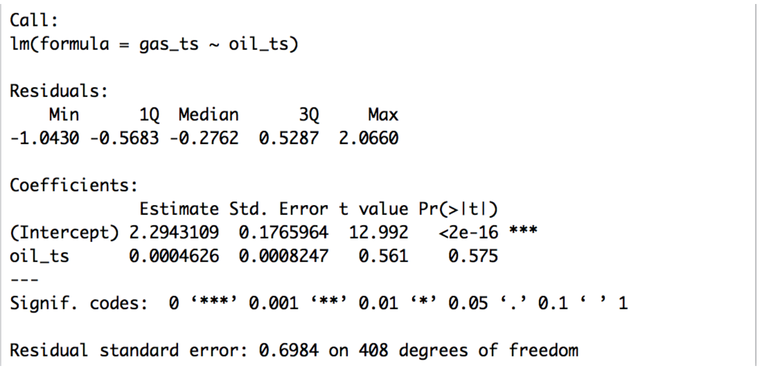
In support of their conclusion, the AP reported a single p-value. You have two tasks for this exercise, and both tasks need the use of the data set *gasOil.Rdata*.

Your first task is to recreate the analysis that the AP likely used to reach their conclusion. Thoroughly discuss all of the errors the AP made in their analysis and conclusion.

Your second task is to create a more statistically-sound model that can be used to predict/forecast inflation adjusted gas prices. Use your model to forecast the inflation-adjusted gas prices from 2012 to 2016.

2.1 Recreate AP Analysis

We run a simple regression using Gas prices as a dependent variable and Oil production as an independent variable:



```
Call:
lm(formula = gas_ts ~ oil_ts)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.0430 -0.5683 -0.2762  0.5287  2.0660 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2.2943109  0.1765964  12.992   <2e-16 ***
oil_ts      0.0004626  0.0008247   0.561    0.575    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6984 on 408 degrees of freedom
```

Figure 8: Probable regression made by AP Analysis. Linear Model 1

- As AP reported a single p-value, it was likely from a regression fitting Oil Production to Gas Price.
- Our regression results in a coefficient for Oil Production of 0.0005 (±0.0016)
- The Oil Production coefficient has a p-value of 0.5752
- This is likely the simple analysis that AP conducted to conclude that there is "evidence of no statistical correlation" between Oil Production and Gas Prices.
- Failing to reject the null hypothesis should not be considered as evidence that the null hypothesis is true.

The issue with this analysis is that the linear regression model breaks many of the Gauss-Markov Assumptions so any statistical inference is not valid. In fact let's take a look at the residuals plot to see what we can find:

Looking at the residual plot we find that:

- Residuals are very far from normal
- They do not comply to the zero conditional mean
- We find correlations all the way to lag 25
- This inspection confirms that linear regression model is flawed and cannot be used to make any statistical inference

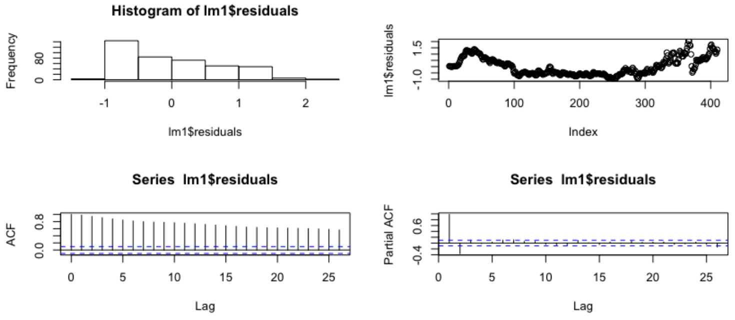


Figure 9: Residuals Ispection of Linear Model 1

2.2 Visual Inspection: *Gas Prices and Oil Production*

Let's take a quick look at both underlying time series: Gas Prices and Oil Production.

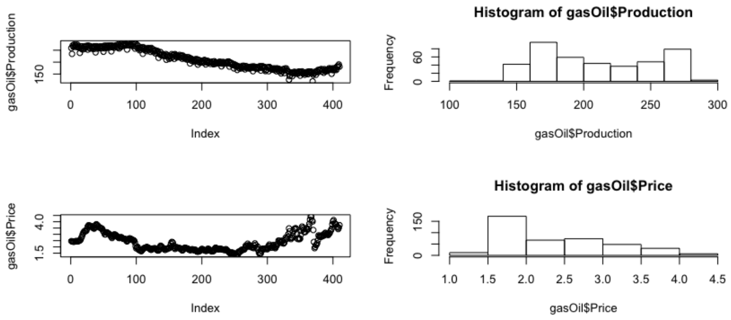


Figure 10: Time Series: Gas prices and Oil production

We observe the following:

- Monthly data from 1978 to 2012: 410 months
- There might be some seasonality
- Oil production trends downwards
- Gas prices more volatile. Double u-shape
- None of them are stationary

Taking a closer look at Gas Prices we observe the following:

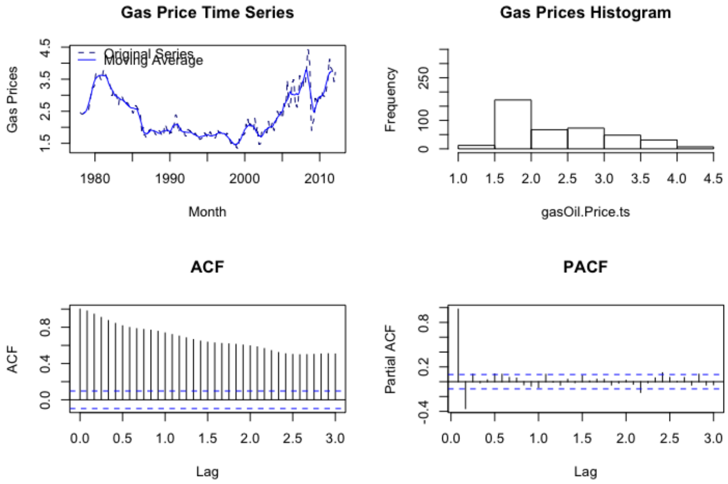


Figure 11: Exploratory visuals of Time Series

- Period includes 410 months from 1978 to 2012
- No clear trend overall
- Seems to be some seasonality in gas prices
- The ACF tails off very slowly
- The PACF drops-off abruptly after lag 2

2.3 Linear Model 2: *Regressing stationary time series*

To construct a model that is valid we first need to make both time series stationary. A simple way to do that is to look at the integrated I(1) model for each time series.

From these two charts we observe the following

- Underlying time series now look stationary
- But we still find correlations and possible seasonality trends
- Though data is not perfect we try now running a linear regression to see what we get.

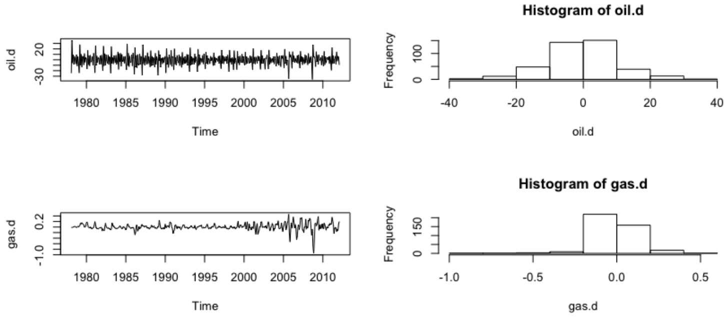


Figure 12: First difference: Integrated I(1) models for each time series

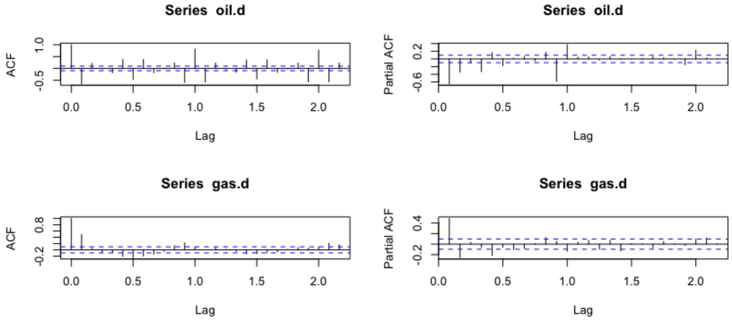


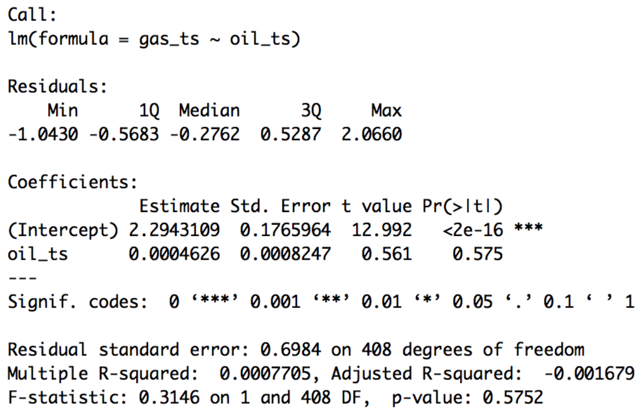
Figure 13: First difference: ACF and PACF of Integrated I(1) models for each time series

Here we run a linear regression on the integrated I(1) model of each time series:

We now observe that the linear regression coefficients are significant for a p-value of 0.1:

- P-value of oil.d coefficient is 0.08
- F statistic of model is 3.043, a figure 10 times higher than our initial model
- Model looks better but further examination of the residuals is required

We now take a look at the residuals plot for the Linear Model 2. We suspect that we may find problems given what we found at the underlying time series:



```
Call:
lm(formula = gas_ts ~ oil_ts)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.0430 -0.5683 -0.2762  0.5287  2.0660 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2.2943109  0.1765964  12.992   <2e-16 ***
oil_ts      0.0004626  0.0008247   0.561    0.575    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6984 on 408 degrees of freedom
Multiple R-squared:  0.0007705, Adjusted R-squared:  -0.001679 
F-statistic: 0.3146 on 1 and 408 DF,  p-value: 0.5752
```

Figure 14: Linear Model 2: Regression of Integrated I(1) models for each time series

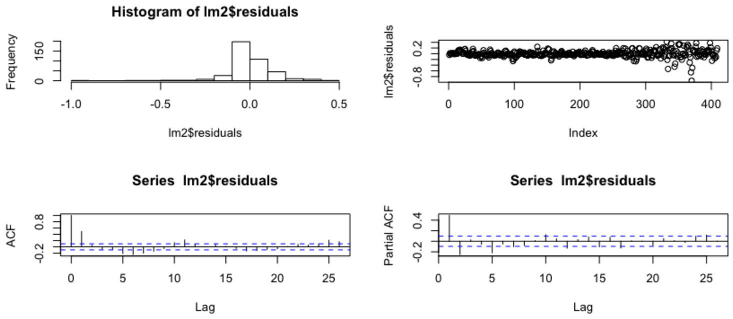


Figure 15: Linear Model 2: Residuals inspection

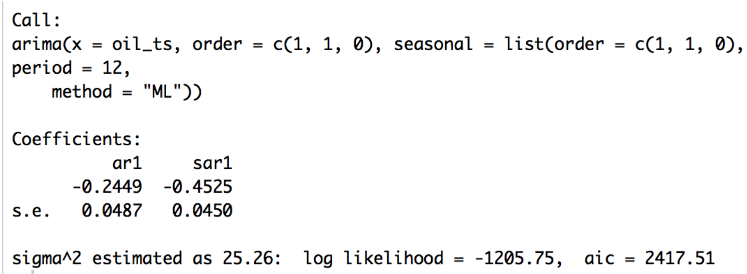
We still find problems with our residual plots which would make invalid any statistical inference. In particular:

- We found autocorrelation of order 1 in the ACF graph of the residuals
- We found that some outliers make our residuals not normal
- We found that the zero conditional mean i probably broken as we approach the end of period
- the PACF and ACF graphs seem to suggest some trend and seasonality effects that need to still be removed

2.4 Arima models 1: *Remove Trend and Seasonality Effects on Oil Production*

To move forward we need to remove sesonality effects. We can do this using ARIMA models.

For oil production we find that an Arima model (1,1,0)(1,1,0) provides a good approximation to the observed time series:



```
Call:
arima(x = oil_ts, order = c(1, 1, 0), seasonal = list(order = c(1, 1, 0),
period = 12,
    method = "ML"))

Coefficients:
          ar1     sar1
      -0.2449  -0.4525
s.e.   0.0487   0.0450

sigma^2 estimated as 25.26:  log likelihood = -1205.75,  aic = 2417.51
```

Figure 16: Oil Production Arima: (1,1,0)(1,1,0)

Inspecting the residuals, we observe the following:

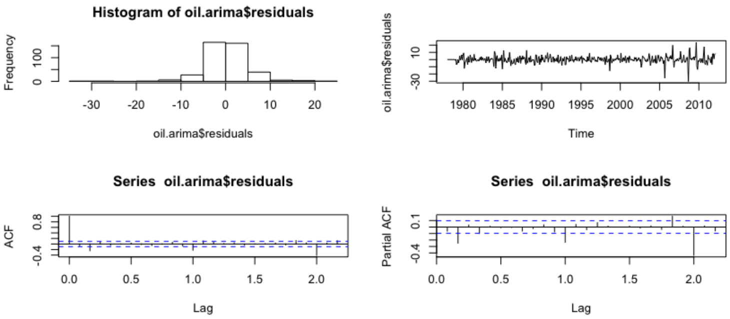


Figure 17: Oil Production Arima: Residual Plots (1,1,0)(1,1,0)

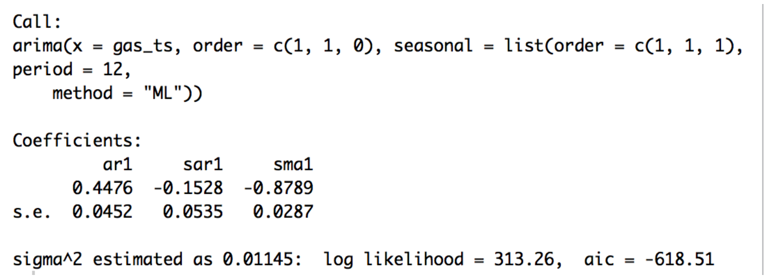
- The model coefficients show high significant values
- We still find some values outside the confidence intervals in the PACF graph

- We examine other possible models and we find that this model has the lowest AIC and given its simplicity we decide to adopt it

2.5 Arima models 2: *Remove Trend and Seasonality Effects on Gas Prices*

Just like we did for oil production, we now look to construct an Arima model to fit the Gas prices time series.

The model with the lowest AIC that we found is the following Gas Prices Arima (0,1,0)(1,1,1). However we decided to go for (1,1,0)(1,1,1) because even though its AIC was a bit higher, its residual plots look closer to white noise :



```
Call:
arima(x = gas_ts, order = c(1, 1, 0), seasonal = list(order = c(1, 1, 1),
period = 12,
    method = "ML"))

Coefficients:
         ar1     sar1     sma1
      0.4476  -0.1528  -0.8789
s.e.  0.0452   0.0535   0.0287

sigma^2 estimated as 0.01145:  log likelihood = 313.26,  aic = -618.51
```

Figure 18: Gas Prices Arima: (1,1,0)(1,1,1)

Inspecting the residuals, we observe the following:

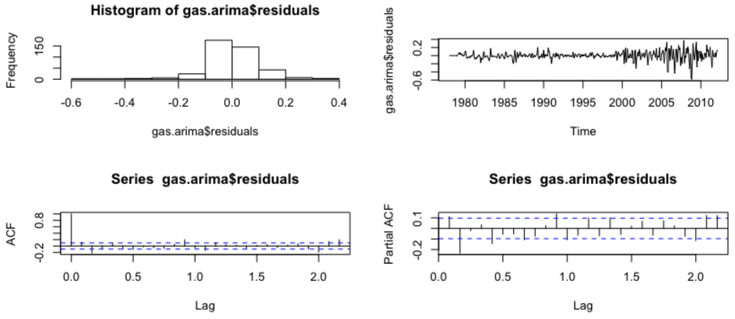


Figure 19: Gas Prices Arima: Residual Plots (1,1,0)(1,1,1)

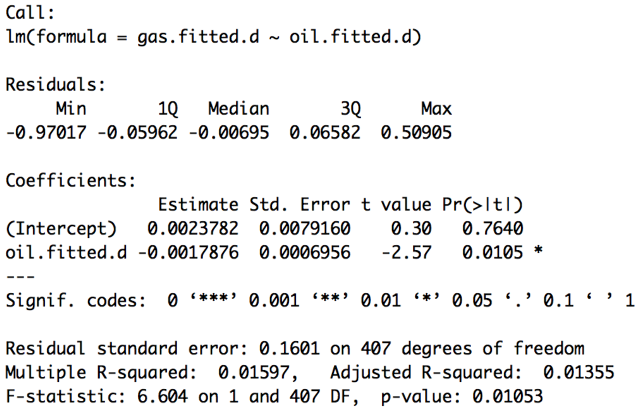
Reviewing this model, we notice the following points:

- This model also exhibits coefficients that are highly significant
- ACF and PACFs graphs are for the most part consistent with what you woudl expect for white noise
- We still find higher volatility towards the end of the period probably meaning that variance is not constant.
- We decided to adopt this model because its residuals look closer to white noise

2.6 Linear Model 3: *Regressing modeled time series*

Once we have determined Arima models that resemble each of the time series, we can now try using those models to see if we find a significant relationship between Gas Prices and Oil Production.

We first fit the data using each model for each time series. We then look at the difference I(1) of each modeled time series. We regressed the modeled difference on Gas Prices on the modeled difference on Oil Production.



```
Call:
lm(formula = gas.fitted.d ~ oil.fitted.d)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.97017 -0.05962 -0.00695  0.06582  0.50905 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)  
(Intercept)   0.0023782  0.0079160    0.30   0.7640  
oil.fitted.d -0.0017876  0.0006956   -2.57   0.0105 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 0.1601 on 407 degrees of freedom
Multiple R-squared:  0.01597,	Adjusted R-squared:  0.01355 
F-statistic: 6.604 on 1 and 407 DF,  p-value: 0.01053
```

Figure 20: Linear Model 3: Regression of Integrated I(1) of time series modeled using Arima

and we also take a look at the residuals:

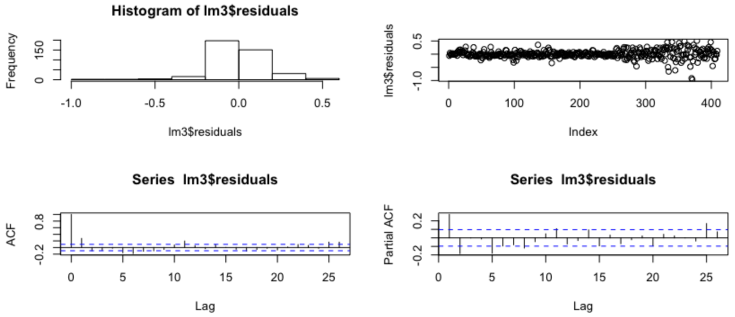


Figure 21: Linear Model 3: Residual Plots

This model leads us to conclude:

- The oil production coefficient now exhibits a highly significant p-value of 0.0105
- for each change in oil production, the change is Gas prices is affected by -0.00179
- We now have evidence that higher changes in oil production leads to lower changes in Gas Prices
- Residuals plot look a lot better than the residual plots of the prior two linear models
- However, we still find that variance is not constant though the zero conditional mean seems to be in line.
- We still find some outliers in our residual histograms that prevents it from being normally distributed
- ACF and PACF look very close to what you would expect to find in white noise
- With this model we can refute AP conclusions

2.7 Forecast 2012-2016: *Predicting Gas Prices*

We use the Arima model to predict Gas prices over the period 2012-2016. Notice that for such a long period, predictions cover a wide range, given the volatility of that market.

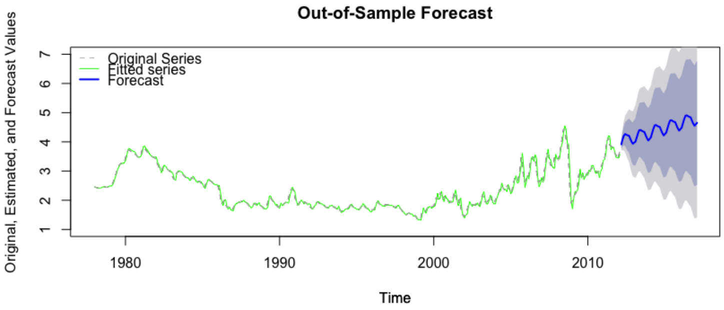


Figure 22: 2012-2016 Gas Prices: Forecast based on Fitted Model

3 *Forecast two series*

The file ex3series.csv contains two monthly time series. You task is to (1) build a forecasting model using techniques covered in lecture 8 - 13, which includes the async. lectures, live sessions, and the assigned readings, and (2) produce a 6-step ahead (monthly) forecast.

We start again with a visual analysis of the two time series. We notice the following:

- Series 1 is decreasing until 2009 and then the structure appears to change and begins to have increasing trend
- Series 2 follows a similar pattern but starts to increase in 2010 instead of 2009
- Series 2 has clearly visible seasonality, but series 1 does not
- ACF for series 1 declines slowly
- ACF for series 1 declines and then increases, showing strong seasonality
- Series 1 PACF drops off sharply after lag 1
- Series 2 PACF drops off after lag 2

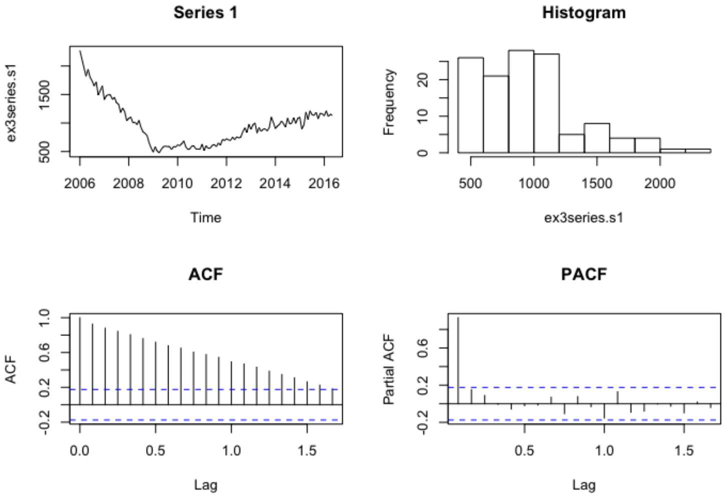


Figure 23: Exploratory Visuals of Series 1

Next, we test for unit roots on each series.

- Augmented Dickey-Fuller test for series 1 fails to reject the null hypothesis that there are unit roots, with p-value equal to 0.19
- Phillips-Perron unit root test for series 1 fails to reject the null hypothesis that there are unit roots, with p-value equal to 0.78
- Augmented Dickey-Fuller test for series 2 fails to reject the null hypothesis that there are unit roots, with p-value equal to 0.46
- Phillips-Perron unit root test for series 2 fails to reject the null hypothesis that there are unit roots, with p-value equal to 0.09
- we conclude that both series are likely to contain unit roots, and are thus likely not stationary

Since our tests for unit roots on both series have failed, we will test for cointegration. If the test for cointegration rejects the null hypothesis, we will conclude that the cointegrated series is stationary and proceed with creating a VAR model.

- The Phillips-Ouliaris Cointegration test between series 1 and series 2 produce a p-value of 0.052, which we will accept as evidence that the two series are cointegrated.
- We will proceed with creating a VAR model for the two series

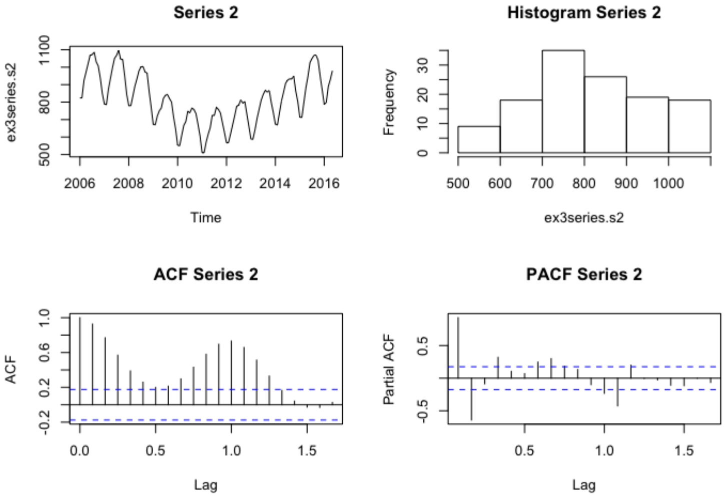
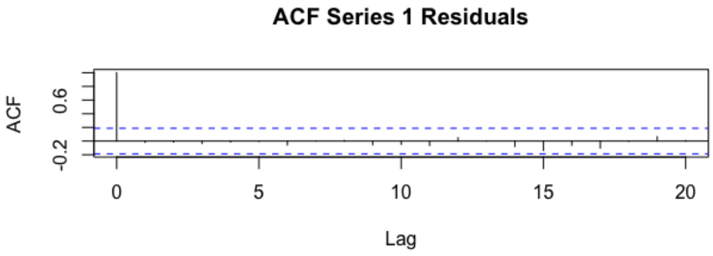


Figure 24: Exploratory Visuals of Series 2

We use the 'ar' function in R to fit a VAR model to the two time series and we discover that the optimal lag order for a VAR model is 15. We produce an ACF chart for the residuals of each series. Since the residuals show no autocorrelation after lag 1, we have validated that the residuals are approximately bivariate white noise.

Next, we review the plotted results of our out-of-sample forecast from our VAR model. The forecast for each series correctly follows trend and seasonality.



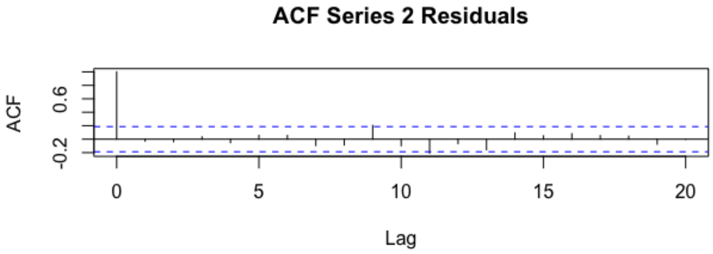
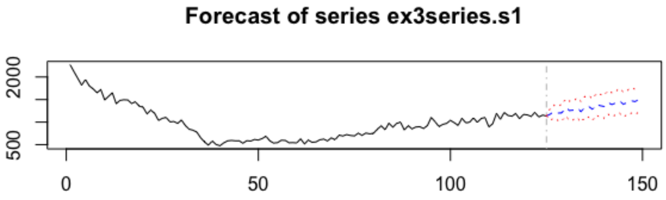


Figure 25: VAR Residuals for Series 1 and 2



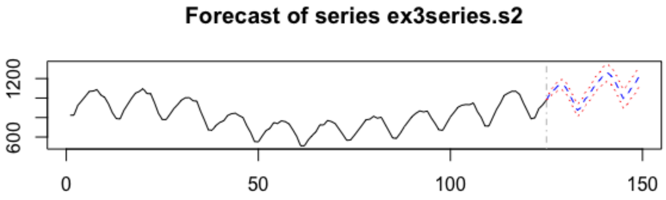


Figure 26: VAR forecast of Series 1 and 2