MATH 642 - Data Project - yl714

Yuchen Li

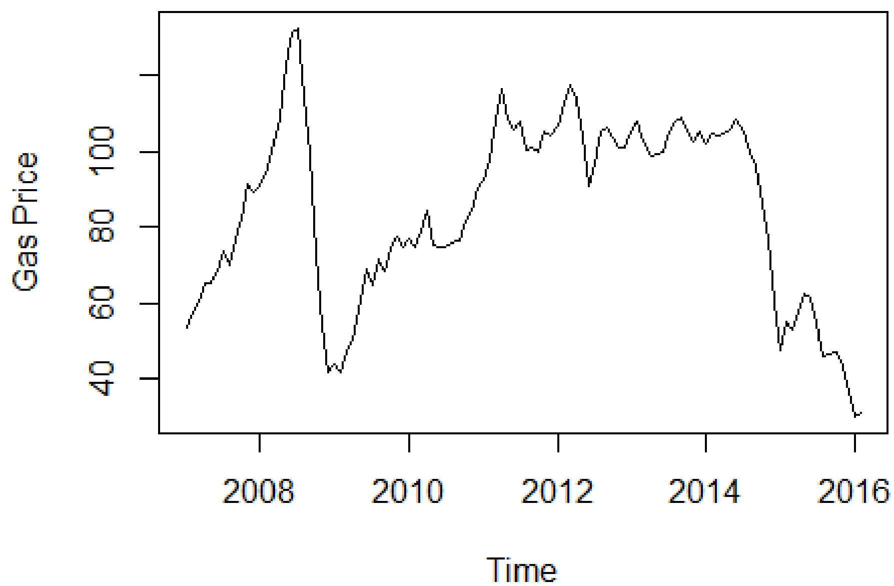
February 17, 2016

Is It Time to Long Oil Stocks? Insights from Historical Data

Project Background:

It can not be denied that crude oil is the most important commodities, hence its price movement is a crucial financial determinant.

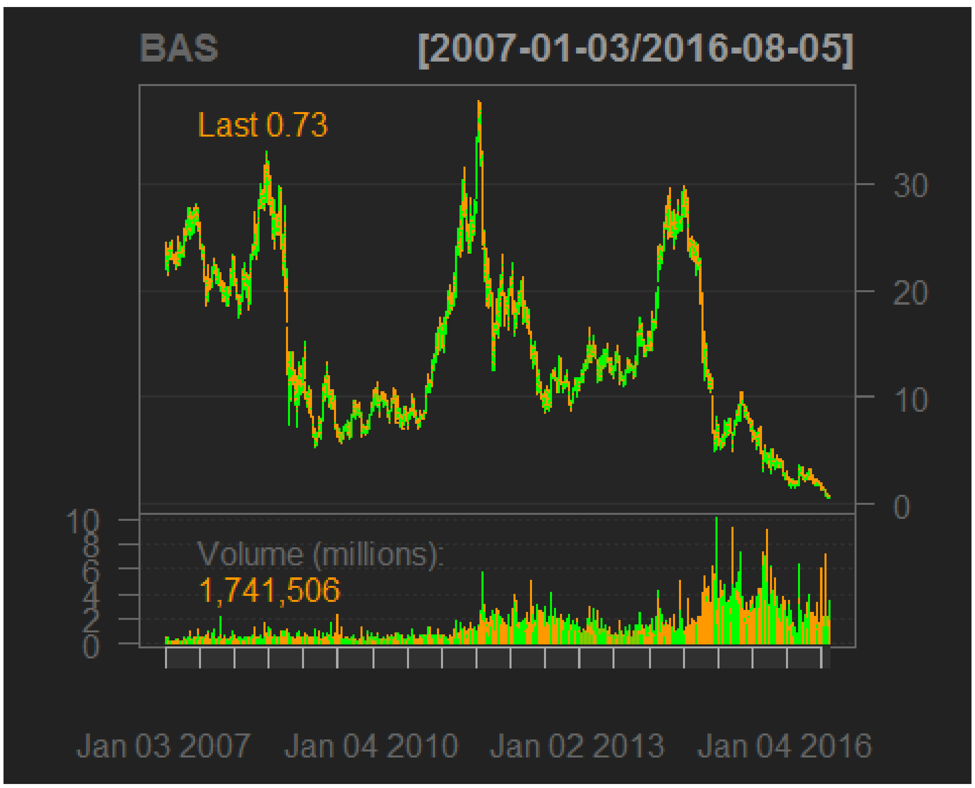
The following graph shows the oil price from January 2007 till Present.



BAS is one of the stocks I am currently investing in, and the following graph is the price of BAS from January 2007 till Present, which bears great resemblance to oil price.

```
## Loading required package: xts

## Loading required package: zoo

##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: TTR

## Version 0.4-0 included new data defaults. See ?getSymbols.

##     As of 0.4-0, 'getSymbols' uses env=parent.frame() and
##  auto.assign=TRUE by default.
##
##  This  behavior  will be  phased out in 0.5-0  when the call  will
##  default to use auto.assign=FALSE. getOption("getSymbols.env") and
##  getOptions("getSymbols.auto.assign") are now checked for alternate
defaults
##
##  This message is shown once per session and may be disabled by setting
##  options("getSymbols.warning4.0"=FALSE). See ?getSymbols for more details.

## Warning in download.file(paste(google.URL, "q=", Symbols.name,
## "&startdate=", : downloaded length 95943 != reported length 200

## [1] "BAS"
```



Project Goals and Customers:

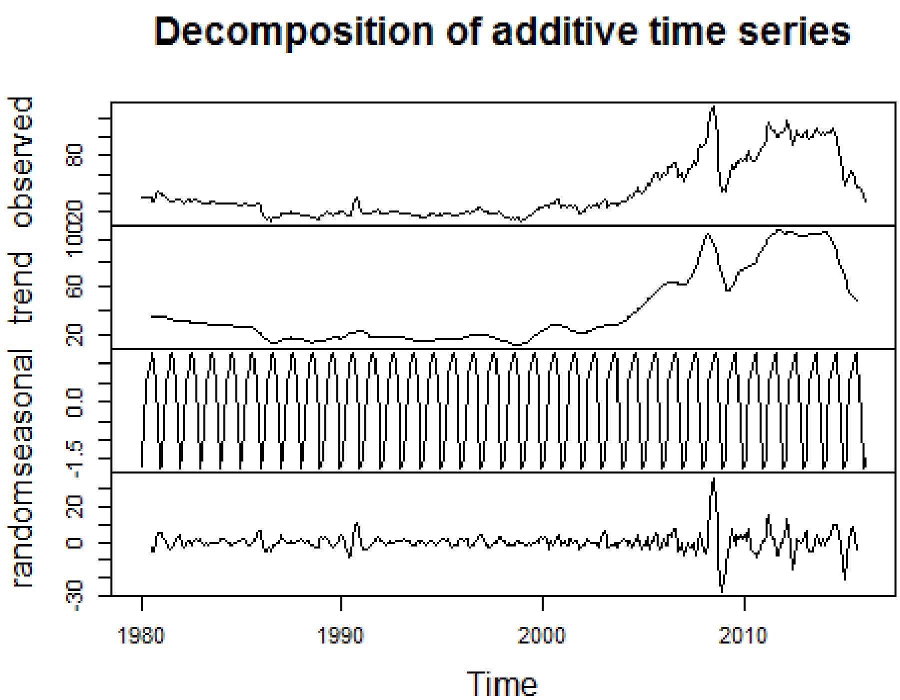
To provide short-term forecasts of oil prices based on historical prices from January 1980 to February 2015, and, to give recommendations(i.e. buy, hold or sell) for oil stock investors.

Holt-Winters Exponential Smoothing

If you have a time series that can be described using an additive model with increasing or decreasing trend and seasonality, you can use Holt-Winters exponential smoothing to make short-term forecasts.

Holt-Winters exponential smoothing estimates the level, slope and seasonal component at the current time point. Smoothing is controlled by three parameters: alpha, beta, and gamma, for the estimates of the level, slope b of the trend component, and the seasonal component, respectively, at the current time point. The parameters alpha, beta and gamma all have values between 0 and 1, and values that are close to 0 mean that relatively little weight is placed on the most recent observations when making forecasts of future values.

First, let's check if oil price can be described using an additive model with trend and seasonality.



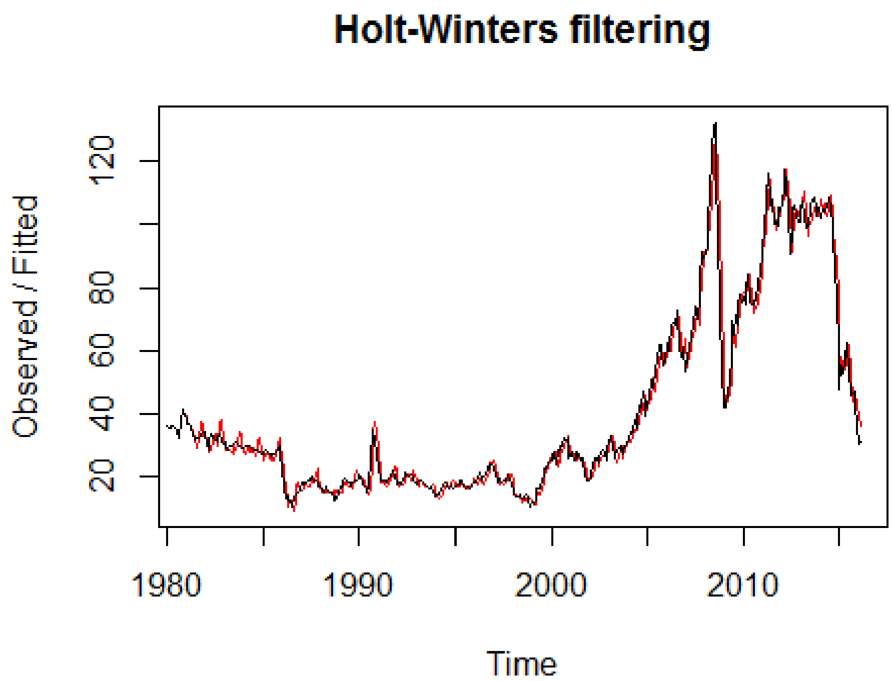
The plot shows that historical oil price can be described as an additive model with increasing trend, seasonality and white noise.

To make forecasts, we can fit a predictive model using the HoltWinters() functions.

```
## Holt-Winters exponential smoothing with trend and additive seasonal
component.
##
## Call:
## HoltWinters(x = oil.price.ts)
##
## Smoothing parameters:
##  alpha: 0.8857118
##  beta : 0
##  gamma: 1
##
## Coefficients:
##           [,1]
## a   28.8332558
## b   -0.1429283
## s1   3.9133791
## s2   5.3669166
## s3   4.0070705
## s4   3.0209644
## s5   2.8298391
## s6   0.8308616
## s7  -0.5656688
## s8  -3.0431981
## s9  -3.9251532
## s10 -4.7228559
## s11 -3.1337960
## s12  2.2167442

## [1] 7776.565
```

The estimated value of alpha, beta and gamma are 0.885, 0, and 1, respectively. The value of alpha (0.885) is relatively high, indicating that the estimate of the level at the current time point is based more upon recent observations than observations in the more distant past. The value of beta is 0.00, indicating that the estimate of the slope b of the trend component is not updated over the time series, and instead is set equal to its initial value. This makes good intuitive sense, as the level changes quite a bit over the time series, but the slope b of the trend component remains roughly the same. In contrast, the value of gamma (1) is high, indicating that the estimate of the seasonal component at the current time point is just based upon very recent observations.

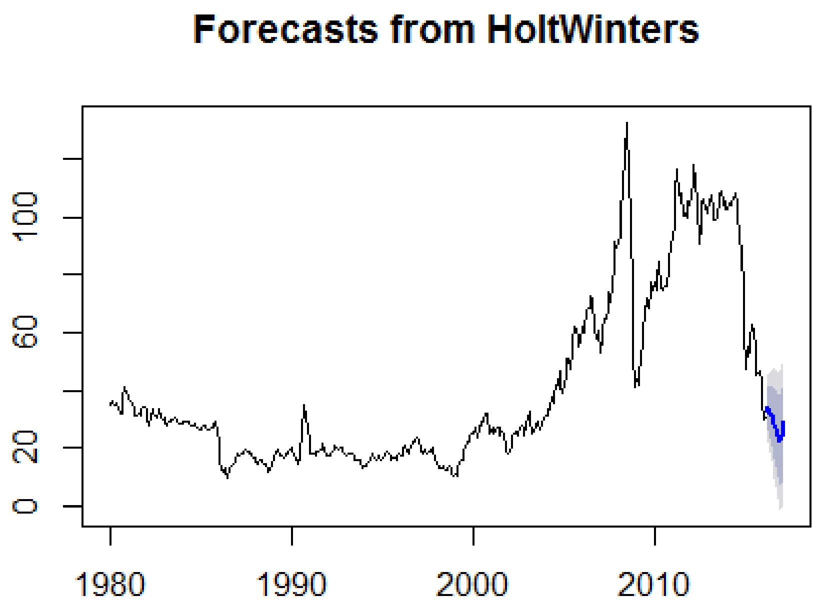
As for Holt Winter's exponential smoothing, we can plot the original time series as a black line, with the forecasted values as a red line on top of that:



(a) Forecasting using Holt-winters exponential smoothing

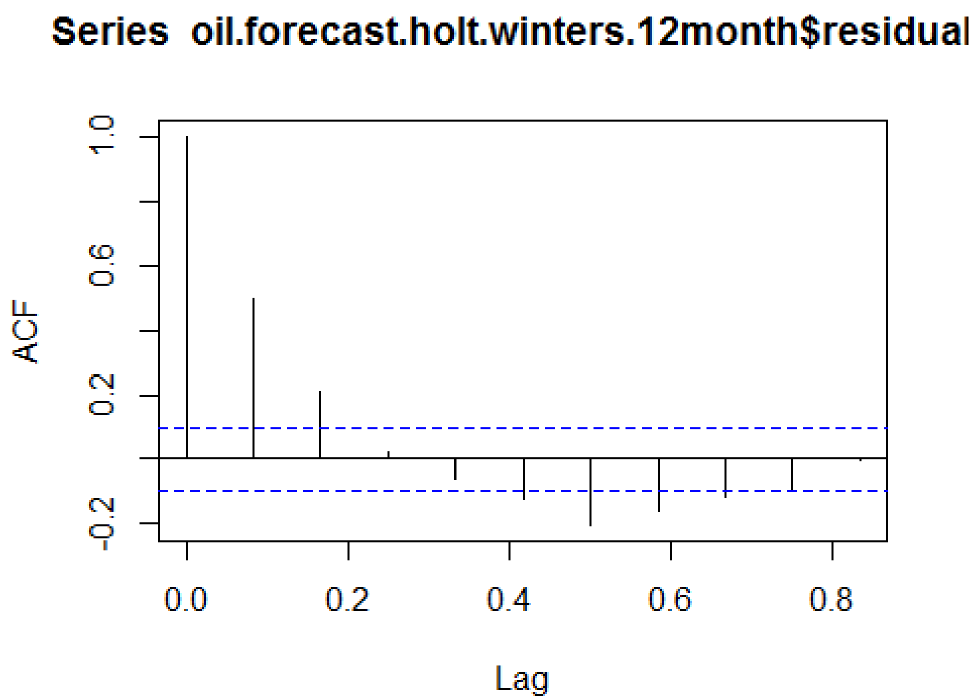
To make forecasts for future times not included in the original time series, we use the "forecast.HoltWinters()" function in the "forecast" package. For example, the original data for the oil prices is from January 1980 to February 2015 If we wanted to make forecasts for March 2015 till present(February 2016) (12 more months), and plot the forecasts, we would type:

```
## Loading required package: timeDate

## This is forecast 7.1

##          Point Forecast     Lo 80    Hi 80      Lo 95    Hi 95
## Mar 2016       32.60371 27.098737 38.10868 24.1845827 41.02283
## Apr 2016       33.91432 26.560523 41.26811 22.6676603 45.16097
## May 2016       32.41154 23.588172 41.23491 18.9173622 45.90572
## Jun 2016       31.28251 21.201563 41.36345 15.8650337 46.69998
## Jul 2016       30.94845 19.750283 42.14662 13.8223302 48.07458
## Aug 2016       28.80655 16.592924 41.02017 10.1274226 47.48567
## Sep 2016       27.26709 14.116189 40.41799  7.1545226 47.37965
## Oct 2016       24.64663 10.620949 38.67231  3.1962015 46.09706
## Nov 2016       23.62175  8.772730 38.47077  0.9121343 46.33136
## Dec 2016       22.68112  7.052076 38.31016 -1.2214386 46.58367
## Jan 2017       24.12725  7.755306 40.49919 -0.9114770 49.16597
## Feb 2017       29.33486 12.252294 46.41743  3.2093290 55.46039
```



The forecasts are shown as a blue line, and the blue and grey shaded areas show 80% and 95% prediction intervals, respectively.

We can investigate whether the predictive model can be improved upon by checking whether the in-sample forecast errors show non-zero autocorrelations at lags 1-10, by making a correlogram and carrying out the Ljung-Box test:



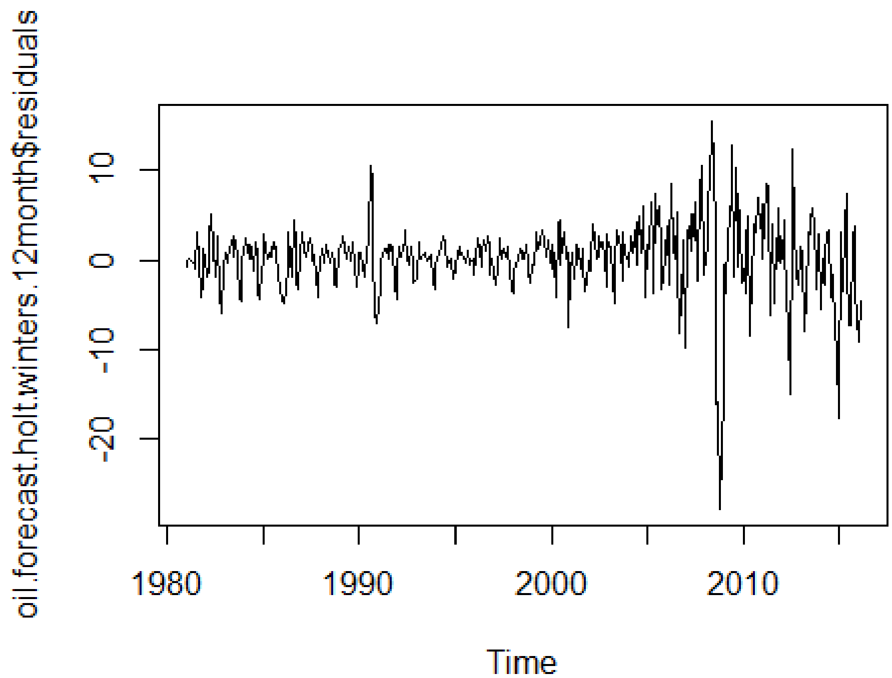

```
##
##  Box-Ljung test
##
## data:  oil.forecast.holt.winters.12month$residuals
## X-squared = 170.78, df = 10, p-value < 2.2e-16
```

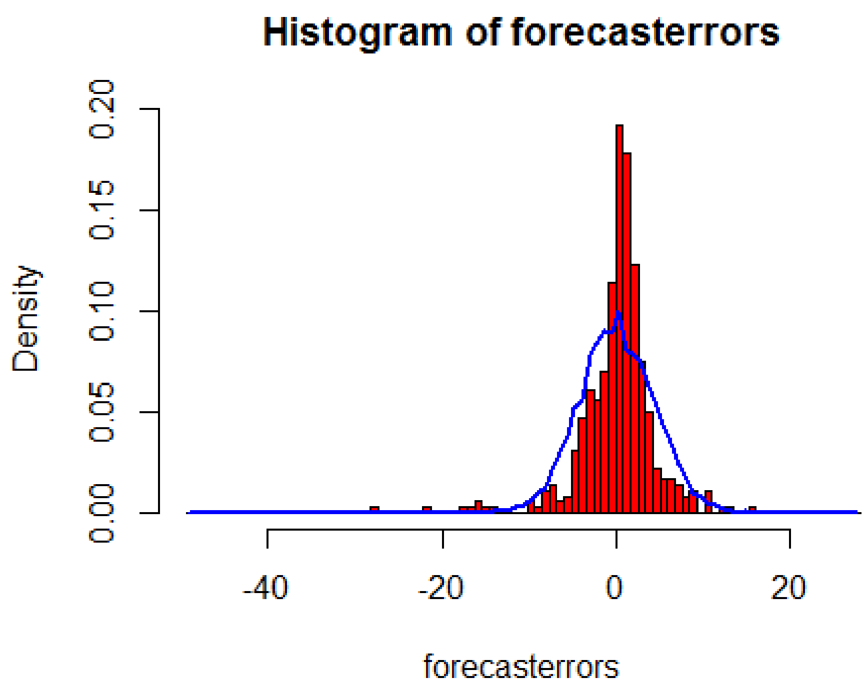
The correlogram shows that the autocorrelations for the in-sample forecast errors do exceed the significance bounds for lags 1-10. Furthermore, the p-value for Ljung-Box test is 2.2e-16, indicating that there is enough evidence of non-zero autocorrelations at lags 1-10.

To check whether the forecast errors are normally distributed with mean zero, we can plot a histogram of the forecast errors, with an overlaid normal curve that has mean zero and the same standard deviation as the distribution of forecast errors. To do this, we can define an R function "plotForecastErrors()", below:

First, let's define a function plotForecastErrors()

Then, we make a time plot and a histogram





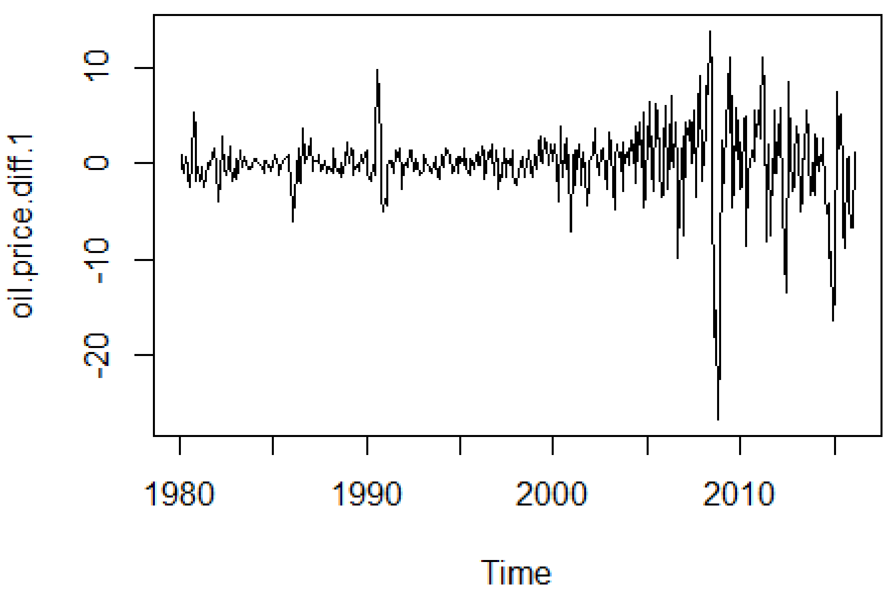
From the time plot, it appears that in 2009, the forecast errors starts to fluctuate a lot more than before. In addition, from the histogram of forecast errors, it shows that the forecast errors are roughly normally distributed with mean zero and constant variance. Due to the contradiction of these two plots and the result of Ljung-Box test, we shall modify current Holt-winters exponential smoothing model.

ARIMA Models

Since the forecast errors show more drift from zero than before, let's consider Autoregressive Integrated Moving Average(ARIMA) models which allow correlated error terms.

Because ARIMA models are defined for stationary time series. Therefore, if oil price is non-stationary, you will first need to 'difference' the time series until you obtain a stationary time series. If you have to difference d times to obtain a stationary time series, then you have an ARIMA(p,d,q) mode, where d is the order of differencing used.

Let's first difference oil price once, and plot the differenced series:



It seems than the first difference of oil price is stationary in mean, which centered around 0.

(a) Selecting a suitable ARIMA model

Since we have loosely obtained stationarity by observing the differencing plot, it's time to figure out an appropriate ARIMA model, ARIMA(p,d,q), where q and q are undetermined. Luckily, auto.arima() function can be used to find the appropriate ARIMA model and its output will suggest the value of p,d and q.

```
## Series: oil.price.ts
## ARIMA(2,1,2)(0,0,2)[12]
```

```
##
## Coefficients:
##          ar1      ar2      ma1     ma2    sma1     sma2
##       1.4723  -0.5867  -1.0334  0.1538  0.0914  -0.1301
## s.e.  0.0870   0.0785   0.1040  0.0959  0.0535   0.0515
##
## sigma^2 estimated as 11.56:  log likelihood=-1141.68
## AIC=2297.37   AICc=2297.63   BIC=2325.86
```

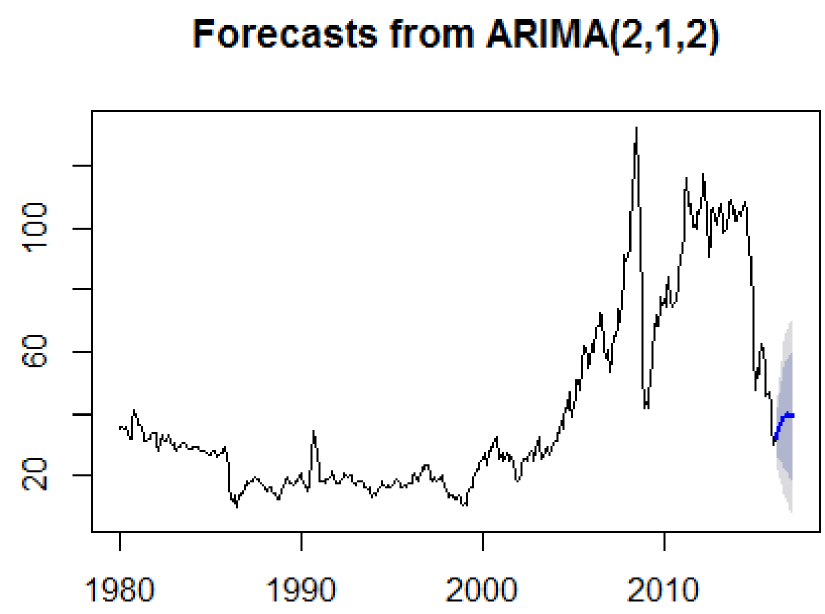
Since p,d,q are 2,1,2 respectively, we can conclude that the first difference of oil price is indeed a stationary time series.

(b) Forecasting using ARIMA(2,1,2) model

First, we fit an ARIMA(2,1,2) model to oil price

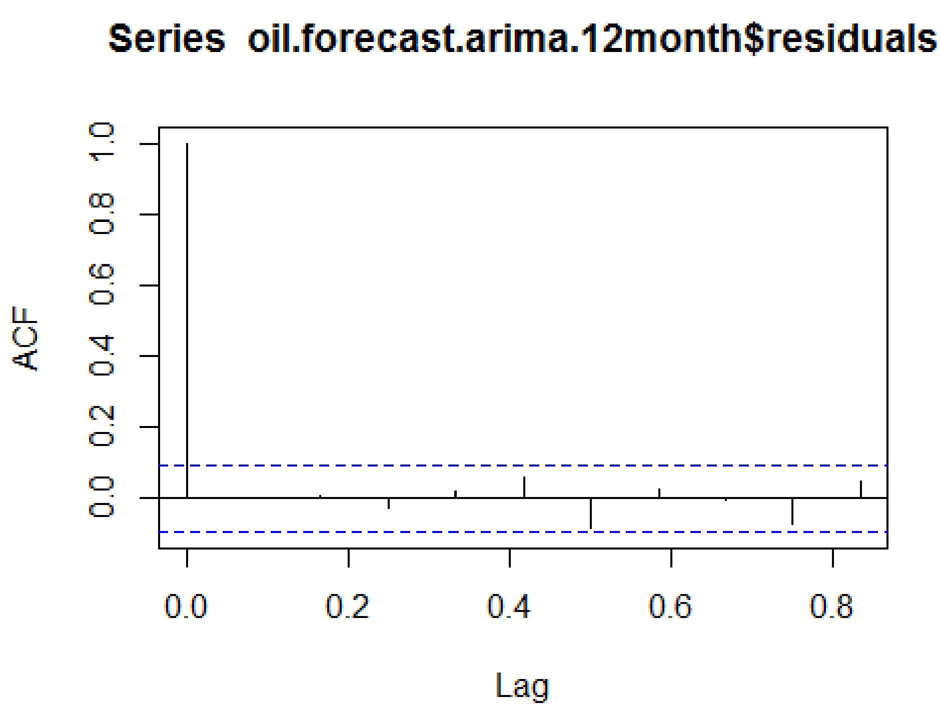
Now let's forecast the price of oil using the model we just obtained.

```
##          Point Forecast    Lo 80    Hi 80     Lo 95    Hi 95
## Mar 2016       32.23757 27.85420 36.62094 25.533785 38.94135
## Apr 2016       33.95556 26.35205 41.55907 22.326999 45.58413
## May 2016       35.71643 25.30002 46.13285 19.785908 51.64696
## Jun 2016       37.22481 24.48382 49.96580 17.739142 56.71048
## Jul 2016       38.34568 23.75411 52.93726 16.029798 60.66157
## Aug 2016       39.06030 23.02036 55.10024 14.529330 63.59127
## Sep 2016       39.42165 22.24430 56.59900 13.151164 65.69214
## Oct 2016       39.51687 21.42585 57.60789 11.849046 67.18470
## Nov 2016       39.43952 20.58625 58.29280 10.605929 68.27312
## Dec 2016       39.27223 19.75317 58.79130  9.420398 69.12407
## Jan 2017       39.07810 18.95052 59.20567  8.295626 69.86057
## Feb 2017       38.89871 18.19355 59.60386  7.232905 70.56451
```



(c) Checking correlations between successive forcast errors

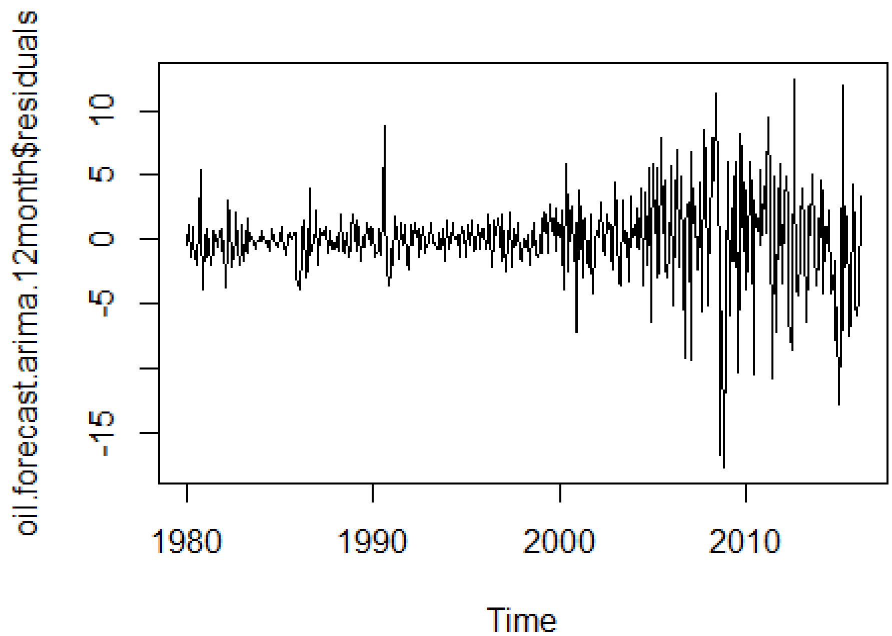
Again, let's use Ljung-Box test:

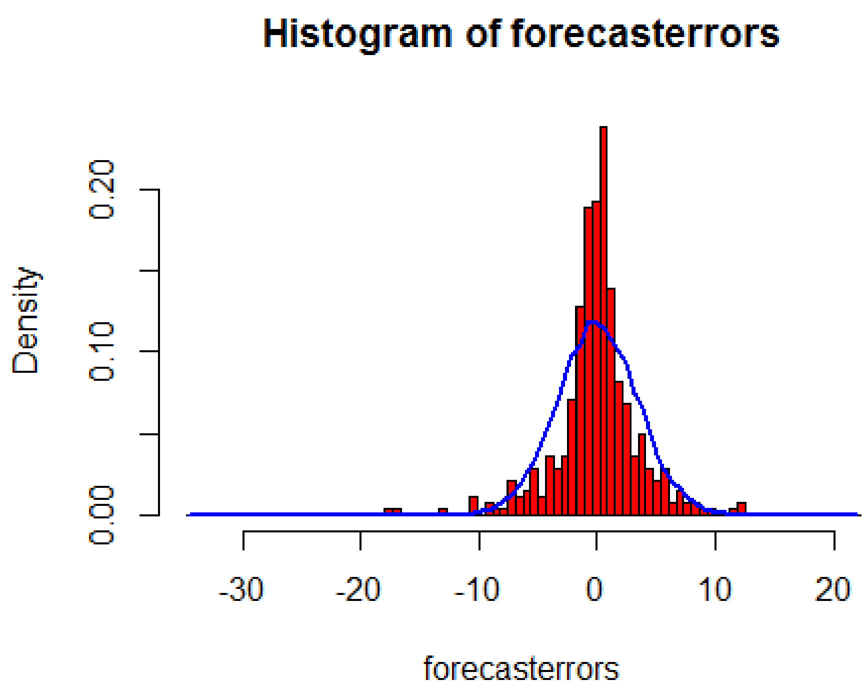


```
##
##  Box-Ljung test
##
## data:  oil.forecast.arima.12month$residuals
## X-squared = 8.7212, df = 10, p-value = 0.5588
```

The p-value for the Ljung-Box test is 0.06429, indicating that there is little evidence suggests that there is correlations in the forecast errors for lags 1-10.

To check whether forecast errors are normally distributed with mean zero and constant variance, we make a time plot of the forecast errors and a histogram.



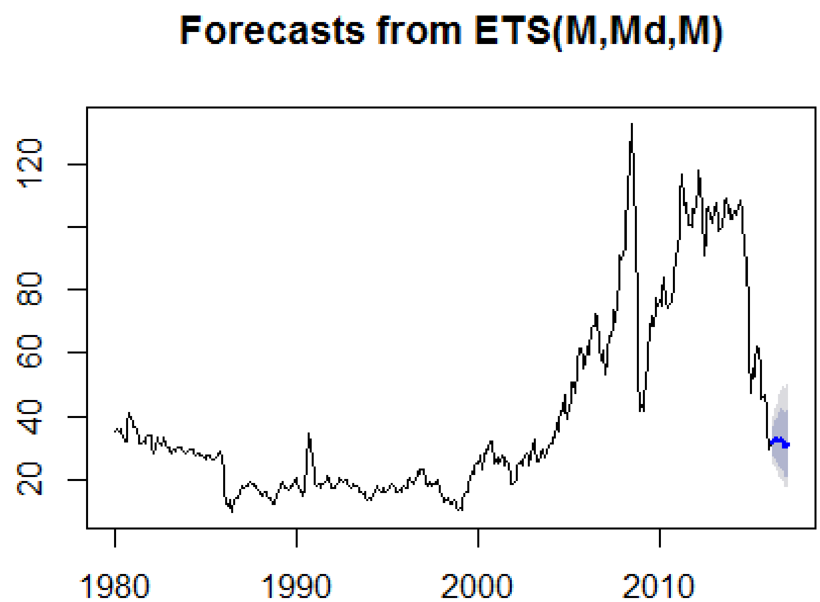


The time plot shows that the forecast errors are roughly centered on zero with constant variance, and the histogram looks like a normal distribution with zero mean and constant variance.

ETS

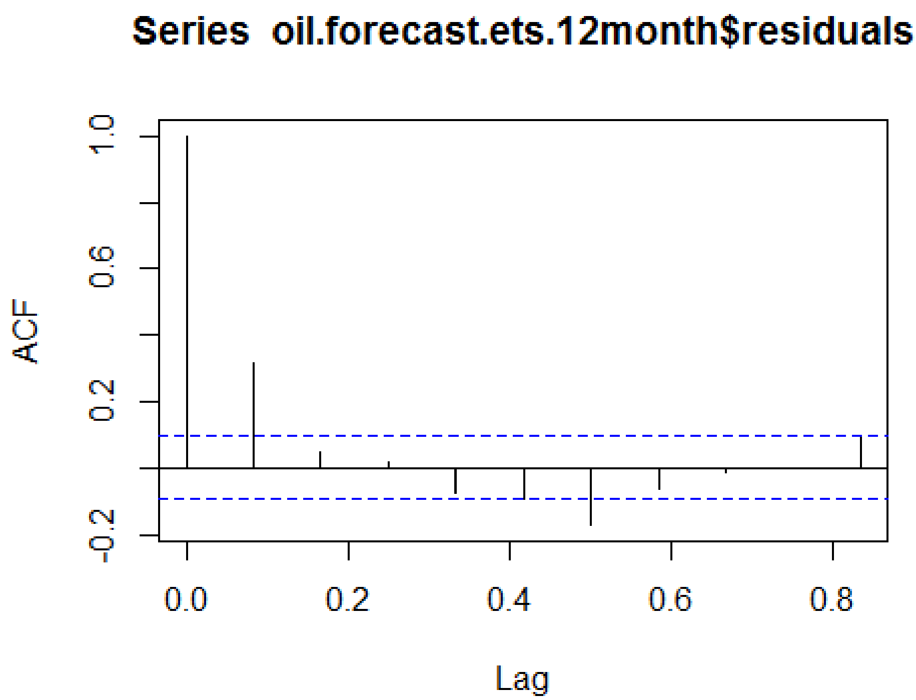
ETS is a more recent R package developed by Dr. Rob Hyndman, ETS(Exponential smoothing state space model) gained its popularity over the year through Dr. Haydnman's famous paper: Automatic Time Series Forecasting: The forecast Package for R.

(a) Forcasting using ETS()



(b) Checking correlations between successive forecast errors

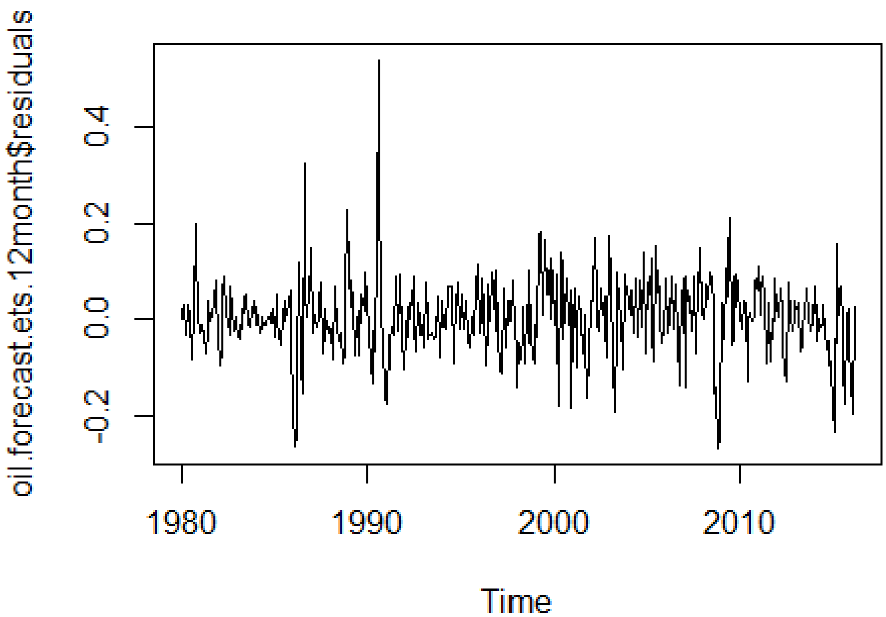
Let's use Ljung-Box test:

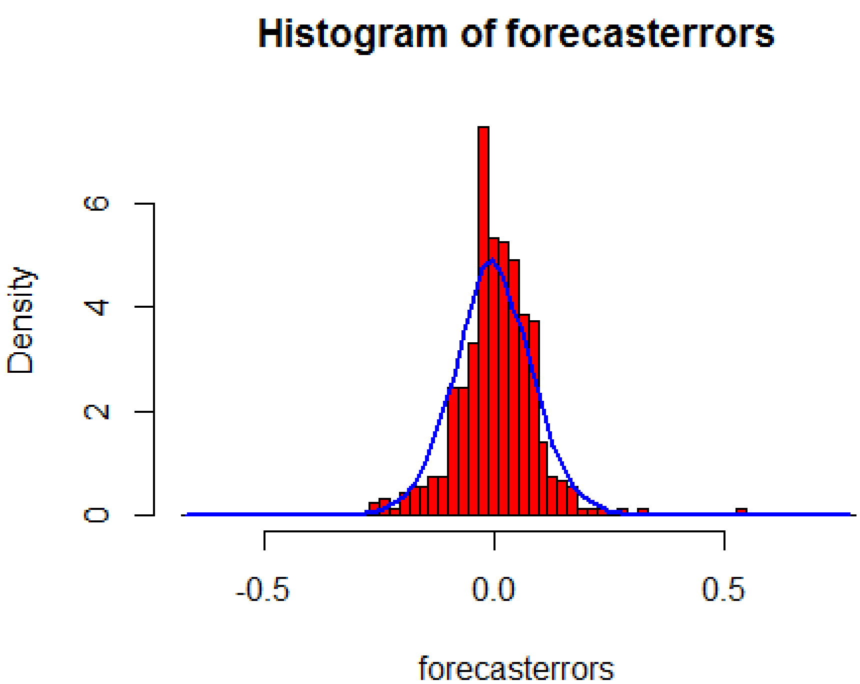



```
##
##  Box-Ljung test
##
## data:  oil.forecast.ets.12month$residuals
## X-squared = 69.715, df = 10, p-value = 5.033e-11
```

The p-value for the Ljung-Box test is 3.788e-10, indicating that there is little evidence suggests that there is correlations in the forecast errors for lags 1-10.

To check whether forecast errors are normally distributed with mean zero and constant variance, we make a time plot of the forecast errors and a histogram.





Similarly to previous models, the time plot shows that the forecast errors of ETS model are roughly centered on zero with constant variance, and the histogram looks like a normal distribution with zero mean and constant variance.

Model Selection

Time series cross-validation answers two important questions:

1. We have used Holtwinters, ARIMA and ETS model, which one is the best?
2. Some of models may require tuning during model training, which tuning parameter values should we choose?

Here I compare the Mean Absolute Error(MAE) of each model on different horizons.

According to Dr. Hyndman, time-series cross-validation follows the following steps:

Assume k is the minimum number of observations for a training set.

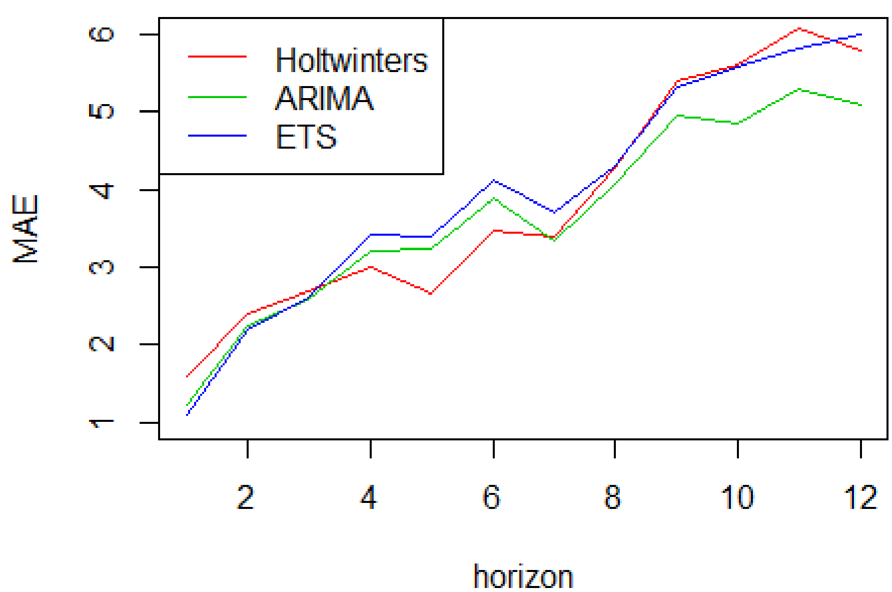
(1) Select observation k+i for test set, and use observations at times 1,2,..., k+i-1 to estimate model. Compute error on forecast for time k+i.

(2) Repeat for i = 0,1,..., T-k where T is total number of observations.

(3) Compute accuracy measure (MAE) over all errors.

```
## Warning in Arima(xshort, order = c(2, 1, 2), seasonal = list(order = c(0,
:
## No drift term fitted as the order of difference is 2 or more.

## Warning in Arima(xshort, order = c(2, 1, 2), seasonal = list(order = c(0,
:
## No drift term fitted as the order of difference is 2 or more.

## Warning in Arima(xshort, order = c(2, 1, 2), seasonal = list(order = c(0,
:
## No drift term fitted as the order of difference is 2 or more.

## Warning in HoltWinters(xshort): optimization difficulties: ERROR:
## ABNORMAL_TERMINATION_IN_LNSRCH

## Warning in Arima(xshort, order = c(2, 1, 2), seasonal = list(order = c(0,
:
## No drift term fitted as the order of difference is 2 or more.

## Warning in Arima(xshort, order = c(2, 1, 2), seasonal = list(order = c(0,
:
## No drift term fitted as the order of difference is 2 or more.

## Warning in Arima(xshort, order = c(2, 1, 2), seasonal = list(order = c(0,
:
## No drift term fitted as the order of difference is 2 or more.

## Warning in Arima(xshort, order = c(2, 1, 2), seasonal = list(order = c(0,
:
## No drift term fitted as the order of difference is 2 or more.
```

```
## Warning in Arima(xshort, order = c(2, 1, 2), seasonal = list(order = c(0,
:
## No drift term fitted as the order of difference is 2 or more.

## Warning in HoltWinters(xshort): optimization difficulties: ERROR:
## ABNORMAL_TERMINATION_IN_LNSRCH

## Warning in Arima(xshort, order = c(2, 1, 2), seasonal = list(order = c(0,
:
## No drift term fitted as the order of difference is 2 or more.

## Warning in Arima(xshort, order = c(2, 1, 2), seasonal = list(order = c(0,
:
## No drift term fitted as the order of difference is 2 or more.

## Warning in Arima(xshort, order = c(2, 1, 2), seasonal = list(order = c(0,
:
## No drift term fitted as the order of difference is 2 or more.

## Warning in HoltWinters(xshort): optimization difficulties: ERROR:
## ABNORMAL_TERMINATION_IN_LNSRCH

## Warning in Arima(xshort, order = c(2, 1, 2), seasonal = list(order = c(0,
:
## No drift term fitted as the order of difference is 2 or more.
```



The MAE plot shows that all three models (Holtwinters, ARIMA, ETS) has increasing MAE as horizon goes up, which makes perfect sense because the further into the future, the less forecasting power a model has. In terms of selecting model, ETS has the smallest MAE before 4 horizons, while holtwinters has between 4 and 7, ETS between 8 and 9, and ARIMA after 9. As a result, when forecasting short term(less than 4 horizons(month)), ETS beats all other model after we compare their MAE after cross validation.

```
##           Jan      Feb      Mar      Apr      May      Jun      Jul
## 2016                   31.37415 32.13658 32.77486 32.52968 32.13680
## 2017 30.61036 30.95599
##           Aug      Sep      Oct      Nov      Dec
## 2016 32.48553 32.64153 32.36447 31.50622 30.27365
## 2017
```

ETS model predicts that oil price will level off in March, April, May and slightly decrease in June, in conclusion, oil price will be facing continuous downward pressure, hence investor should carefully consider investment decision and lean towards selling.