W271-2 – Spring 2016 – **Lab 3**

Juanjo Carin, Kevin Davis, Ashley Levato, Minghu Song

April 22, 2016

Contents

Instructions

- Thoroughly analyze the given dataset or data series. Detect any anomalies in each of the variables. Examine if any of the variables that may appear to be top- or bottom-coded.
- Your report needs to include a comprehensive graphical analysis
- Your analysis needs to be accompanied by detailed narrative. Just printing a bunch of graphs and econometric results will likely receive a very low score.
- Your analysis needs to show that your models are valid (in statistical sense).
- Your rationale of using certian metrics to choose models need to be provided. Explain the validity / pros / cons of the metric you use to choose your "best" model.
- Your rationale of any decisions made in your modeling needs to be explained and supported with empirical evidence.
- All the steps to arrive at your final model need to be shown and explained clearly.
- All of the assumptions of your final model need to be thoroughly tested and explained and shown to be valid. Don't just write something like, "the plot looks reasonable", or "the plot looks good", as different people interpret vague terms like "reasonable" or "good" differently.

Part 1

Modeling House Values

In Part 1, you will use the data set `houseValue.csv` to build a linear regression model, which includes the possible use of the instrumental variable approach, to answer a set of questions interested by a philanthropist group. You will also need to test hypotheses using these questions.

The philanthropist group hires a think tank to examine the relationship between the house values and neighborhood characteristics. For instance, they are interested in the extent to which houses in neighbhorhood with desirable features command higher values. They are specifically interested in environmental features, such as proximity to water body (i.e. lake, river, or ocean) or air quality of a region.

The think tank has collected information from tens of thousands of neighborhoods throughout the United States. They hire your group as contractors, and you are given a small sample and selected variables of the original data set collected to conduct an initial, proof-of-concept analysis. Many variables, in their original form or transfomed forms, that can explain the house values are included in the dataset. Analyze each of these variables as well as different combinations of them very carefully and use them (or a subset of them), in its original or transformed version, to build a linear regression model and test hypotheses to address the questions. Also address potential (statistical) issues that may be casued by omitted variables.

Part 2

Modeling and Forecasting a Real-World Macroeconomic / Financial time series

Build a time-series model for the series in `lab3_series02.csv`, which is extracted from a real-world macroeconomic/financial time series, and use it to perform a 36-step ahead forecast. The periodicity of the series is purposely not provided. Possible models include AR, MA, ARMA, ARIMA, Seasonal ARIMA, GARCH, ARIMA-GARCH, or Seasonal ARIMA-GARCH models.

We argue that an XXXX model is the best choice. We explored this timeseries data and evaluted possible model types including AR, MA, ARMA, ARIMA, Seasonal ARIMA, GARCH, ARIMA-GARCH, or Seasonal ARIMA-GARCH models. A GARCHXXXXX model was identified as the best fitting model because the time-series is not stationary in the mean or variance and is conditional heteroskedasstic which makes modeling the data using AR, MA, ARMA, ARIMA, or SARIMA models not good fits.

First we load the data and conduct some exploratory analysis. This includes plotting the timeseries and examining key characteristics that can help us elimate unlikly modeling choices instead of testing all model types.

```
#load the data
d<-read.csv('lab3_series02.csv')
str(d)
```

```
## 'data.frame':    2332 obs. of  2 variables:
##  $ X         : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ DXCM.Close: num  9.88 9.79 9.68 9.64 9.42 9.47 9.16 8.99 8.6 8.81 ...
```

```
all(d$X == 1:dim(d)[1]) # check if 1st column is just an incremental index
```

```
## [1] TRUE
```

```
d <- d[, -1]
```

The file has two columns but the first one is just an incremental index so we discard it. The second column (that is stored in a numeric vector called `DXCM.Close`) contains 2332 observations. No information about the time scale is given so for now we will just index it from 1 to 2332 (with frequency=1). The main descriptive statistics of the series, as well as its histogram and time-series plot are shown below. The histogram is positively skewed with a long tail and far from normal, but as typical with timeseries data the histrogram does not provide information about the dynamics of the series.

```
# Exploratory Data Analysis ------------------------------------------------
# See the definition of the function in ## @knitr Libraries-Functions-Constants
desc_stat(d, 'Time series', 'Descriptive statistics of the time series.')
```

Table 1: Descriptive statistics of the time series.

	Time series
Mean	23.21
St. Dev	23.44
1st Quartile	8.19
Median	12.36

	Time series
3rd Quartile	32.56
Min	1.39
Max	101.91

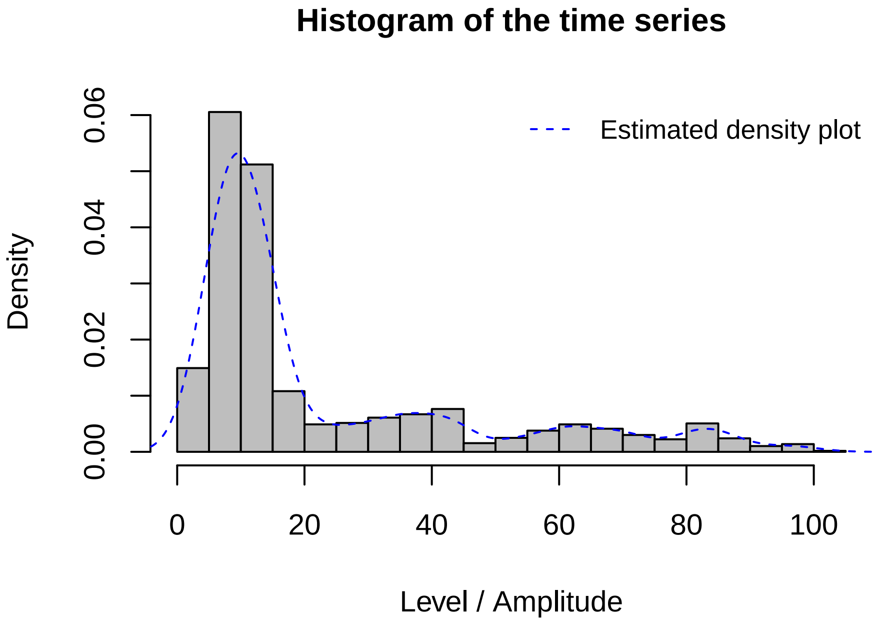
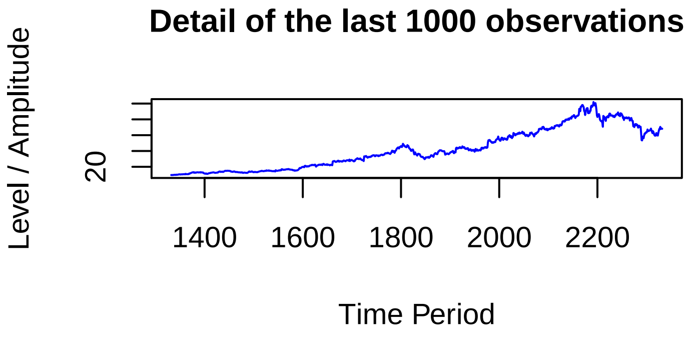
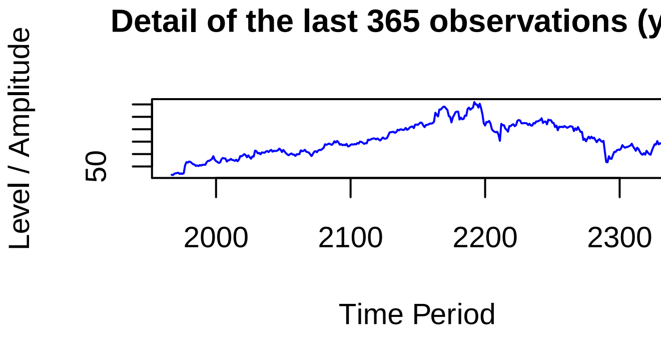

Histogram of the time series

Figure 1: Histogram of the data.

We look at the timeseries itself and some zoomed in plots covering approximately the last 1000, 365, 91, and 60 units which could possibly correspond to a year, quarter, and two months. These plots inform us that the time series is not (mean) stationary, as the mean depends on time with an increasing (then slighly decreasing) trend.

```
d.ts <- ts(d)
```


Detail of the last 1000 observations


Detail of the last 365 observations (y

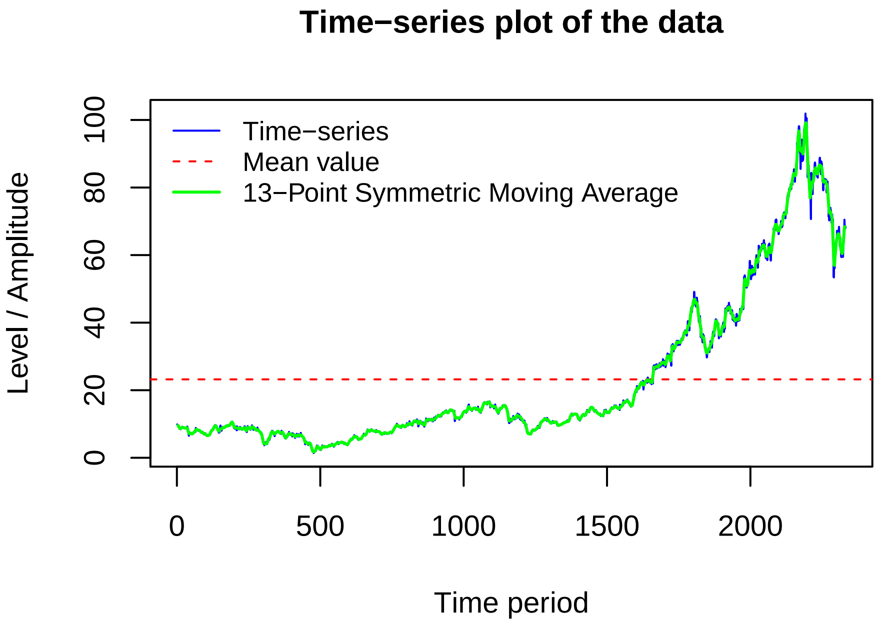
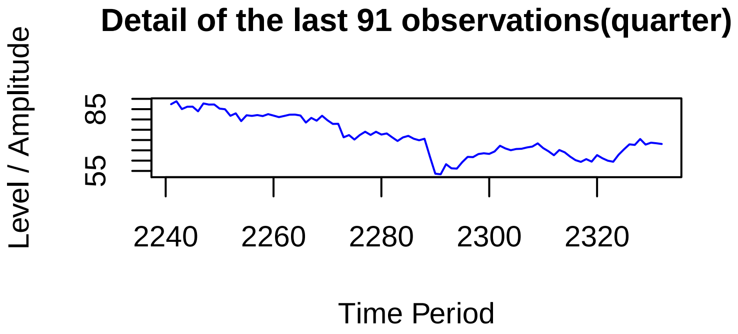
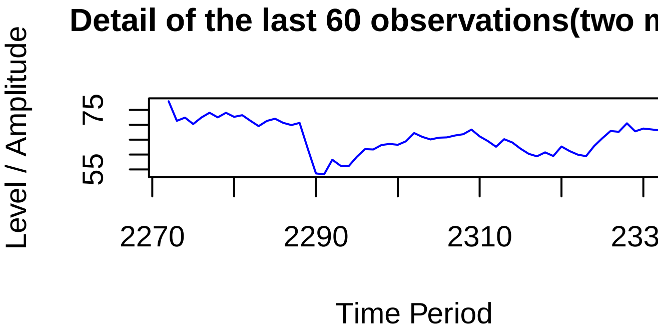


Figure 2: Time-series plot





We then take a look at the box and whisker plot of the timeseries in order to see if the variance is stationary over time. As we can see in the plot below the variance is not stationary in time and starts to increase over time especially after group 16 in the plot. Note since we do not know the unit of observation we choose to group the data into 22 equal bins of 106 observations in order to be able to identify any changes in variance. We can also plot the square of the timeseries to see if the variance is conditional on the prior and as we can see in the figure below there is a change at around obervation 1700.

```r
plot.ts(d.ts*d.ts, col = 'purple', type = 'l',
        xlab = "Time period", ylab = "Level / Amplitude",
        main = "Squared Time-series plot of the data")
```

From the variance analysis above we choose to subset the data and only work with the data from about 1500 to 2332, becasue the data prior to 1500 is likely to not be useful in forcasting the data.

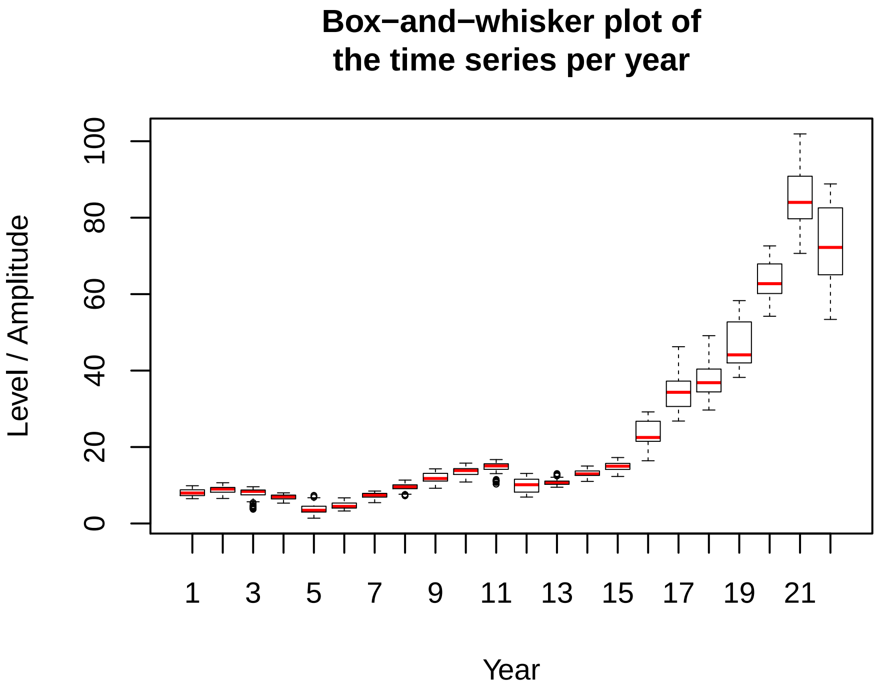


Figure 3: Boxplot of the series (every 106 observations).

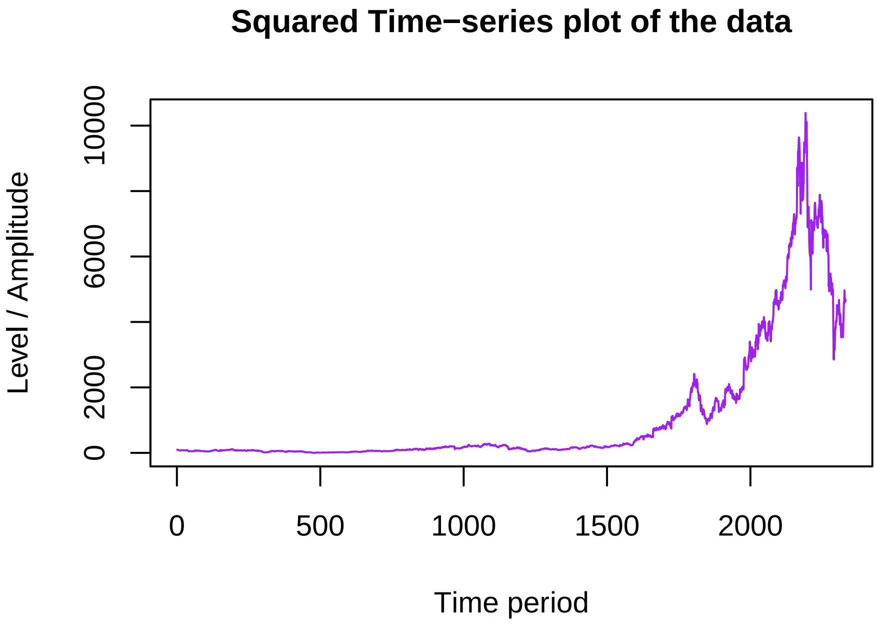


Figure 4: Squared Time-series plot

```
#subset the data to only use observation 1700-2332
d_sub<-d[1500:2332]
d_sub.ts<-ts(d_sub)
```

```
plot.ts(d_sub.ts, col = 'blue', type = 'l',
        xlab = "Time period", ylab = "Level / Amplitude",
        main = "Time-series plot of the data")
```

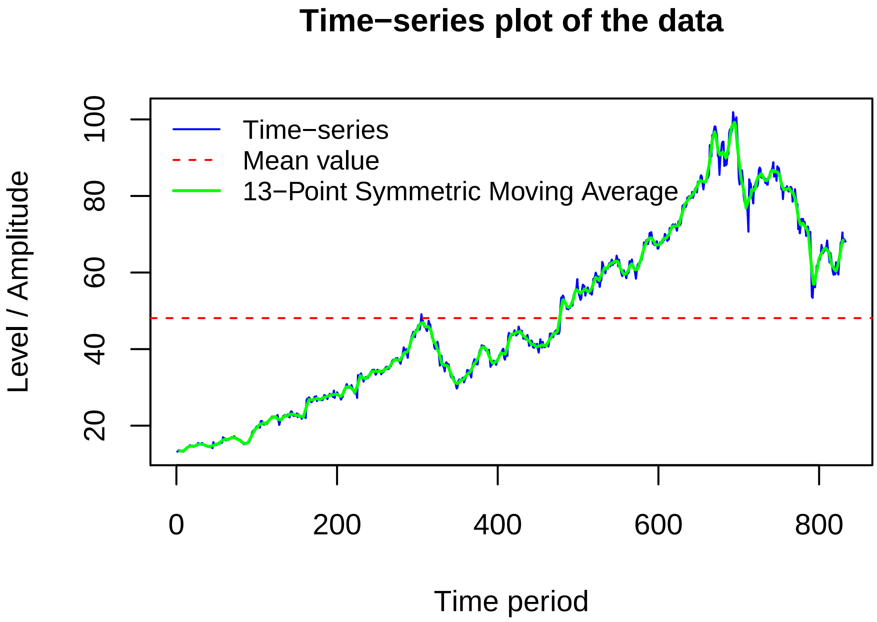
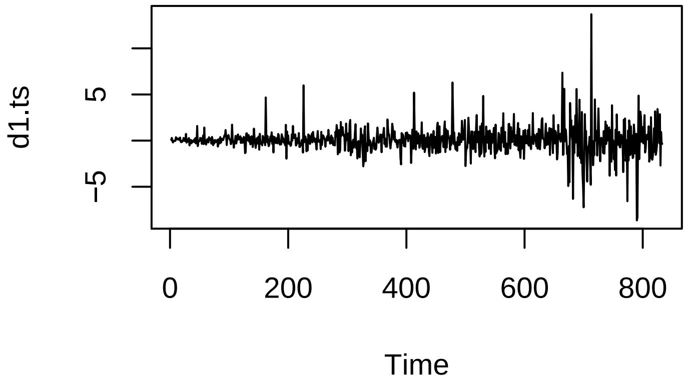


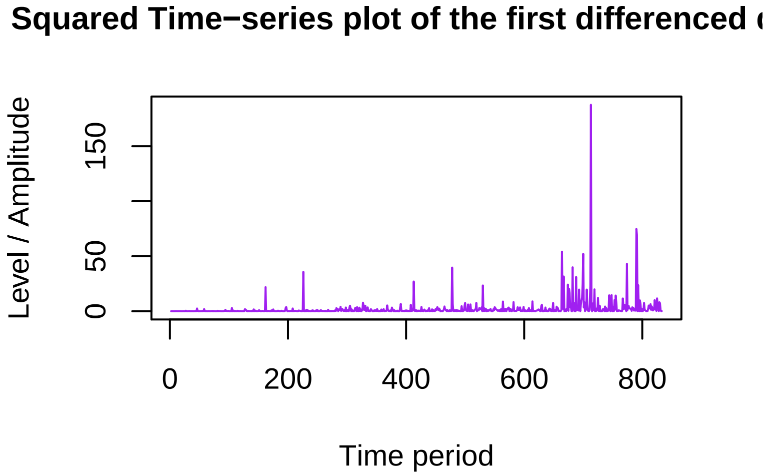
Figure 5: Subset Time-series plot

Since this data is not stationary in the mean or variance; MA, AR, and ARMA models will not be good models to forcast the data. In order to model this non-staionary data we need to difference the data. The plot of the differenced data looks stationary in the mean; however, we also plot the squared ts and notice the variance is still volitile. To confirm our suspicion of conditionally heteroskedastic we plot the ACF of the squared values with adjusted mean. The results in the plot below confirm our suspicion which indicatese that a GARCH style model would be better suited to model the data because they allow for conditional changes in variance.

```
d1.ts<-diff(d_sub.ts)
plot(d1.ts)
```



```
#Now we take a look at the variance of this differenced model by looking at the squared ts
plot.ts(d1.ts*d1.ts, col = 'purple', type = 'l',
        xlab = "Time period", ylab = "Level / Amplitude",
        main = "Squared Time-series plot of the first differenced data")
```



We now model the data using a GARCH model and look at the confidence intervals, acf, and acf of squared values to examine the residuals. We see that just the GARCH model coefficients are not all statistically significant since zero falls within the the confidence interval of b1.

```
library(tseries)
d1.garch<-garch(d1.ts,  trace=FALSE)
confint(d1.garch)
```

```
##          2.5 %     97.5 %
## a0  1.77323008 2.11335546
```

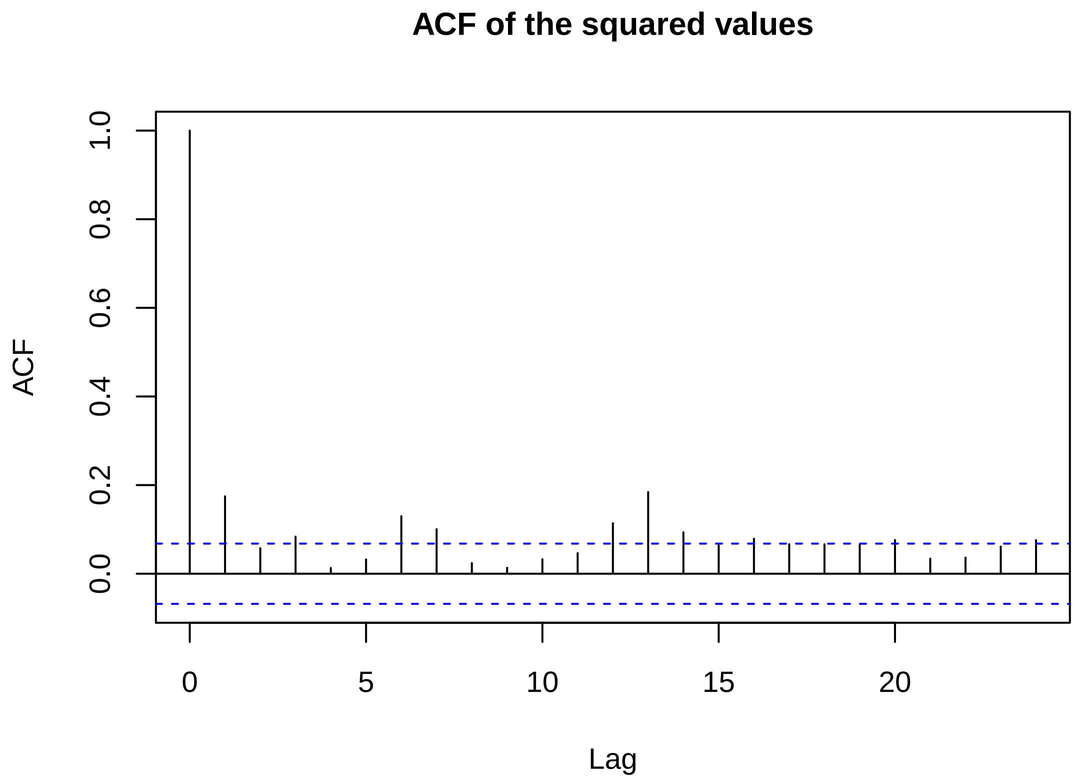


Figure 6: Autocorrelation Plot

```
## a1  0.16757260 0.24695469
## b1 -0.04585174 0.04585174
```

```
d1.res <- d1.garch$res[-1]
```

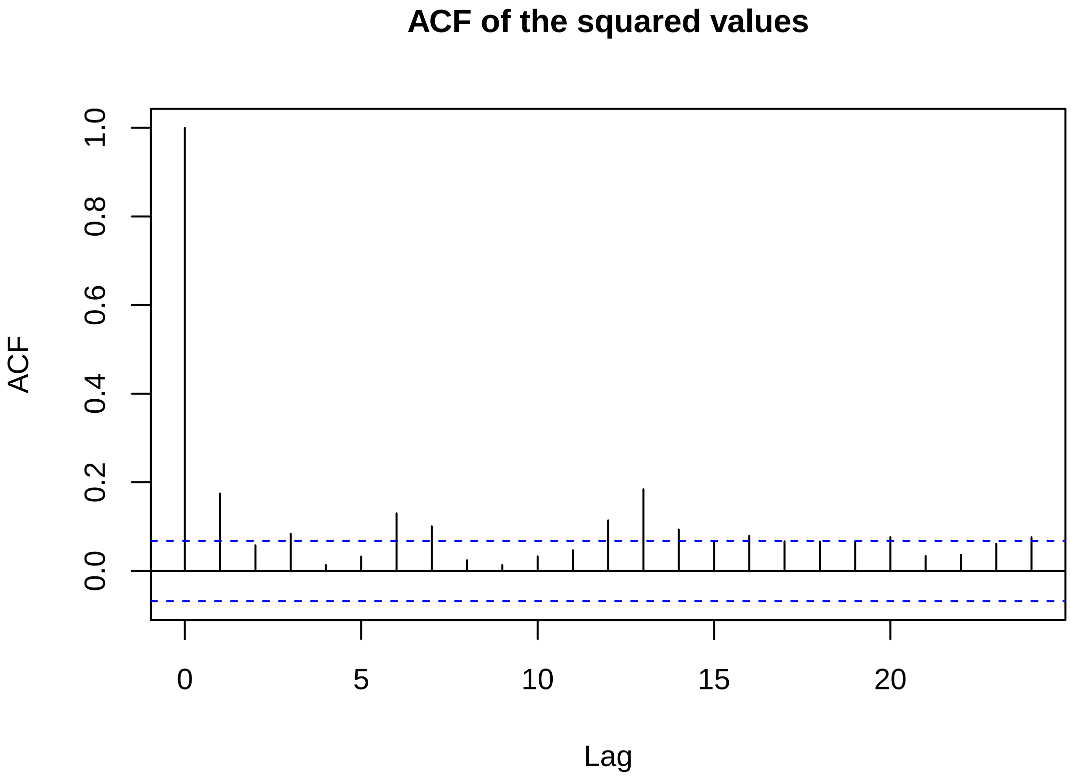


Figure 7: Autocorrelation Plots

Part 3

Forecast the Web Search Activity for global Warming

Imagine that you group is part of a data science team in an appreal company. One of its recent products is Global-Warming T-shirts. The marketing director expects that the demand for the t-shirts tends to increase when global warming issues are reported in the news. As such, the director asks your group to forecast the level of interest in global warming in the news. The dataset given to your group captures the relative web search activity for the phrase, "global warming" over time. For the purpose of this exercise, ignore the units reported in the data as they are unimportant and irrelevant. Your task is to produce the weekly forecast for the *next 3 months* for the relative web search activity for global warming. For the purpose of this exercise, treat it as a *12-step ahead forecast*.

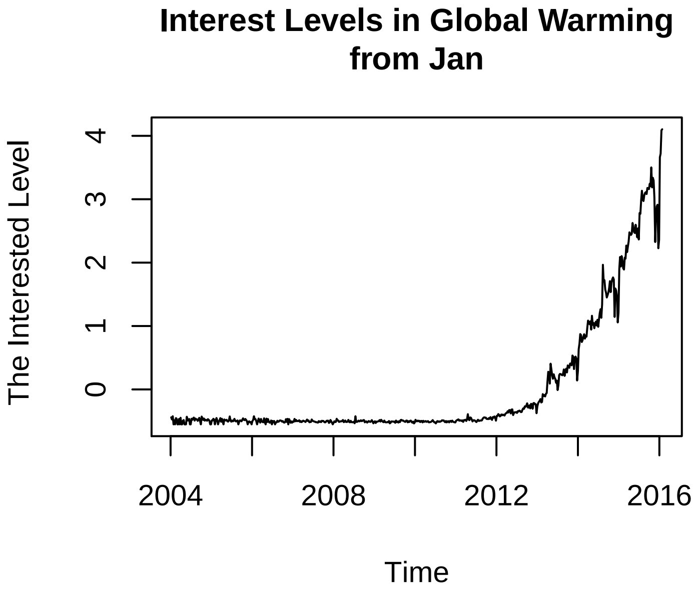
The dataset for this exercise is provided in `globalWarming.csv`. Use only models and techniques covered in the course (up to lecture 13). Note that one of the modeling issues you may have to consider is whether or not to use the entire series provided in the data set. Your choice will have to be clearly explained and supported with empirical evidence. As in other parts of the lab, the general instructions in the *Instruction Section* apply.

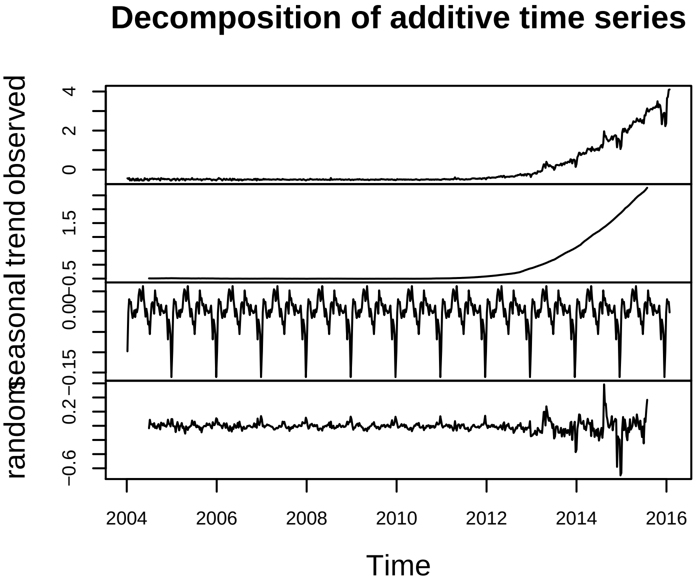
```
# Loading the Global Warming Data
gw<- read.csv('globalWarming.csv', header = TRUE)
head(gw)
```

```
##      Date data.science
## 1  1/4/04       -0.440
## 2 1/11/04       -0.474
## 3 1/18/04       -0.423
## 4 1/25/04       -0.551
## 5  2/1/04       -0.486
## 6  2/8/04       -0.551
```

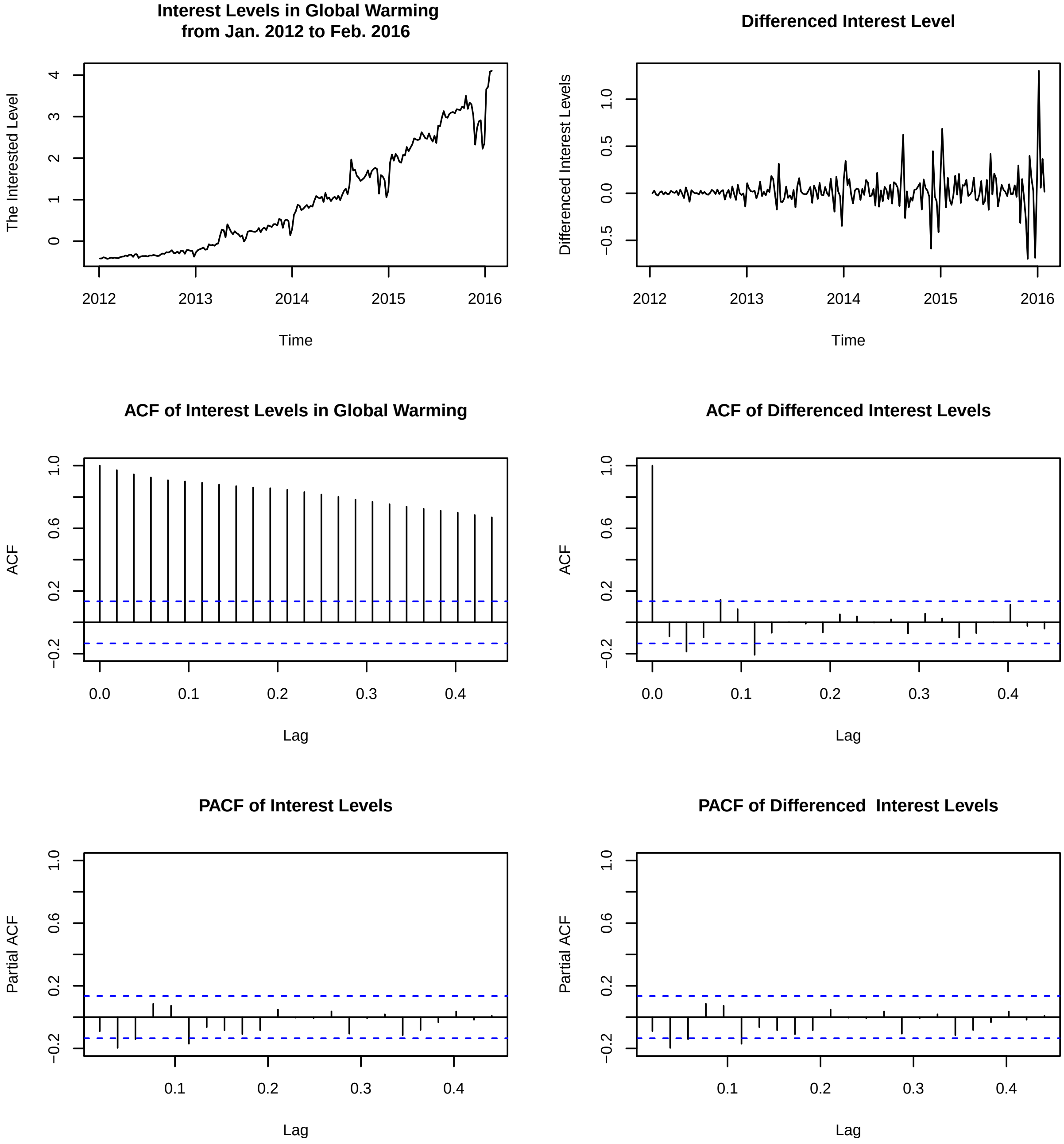
```
# Create ts object with non-integer frequencies and
gw_ts <- ts(gw[,2], start = 2004 + 4/265.25, frequency = 365.25/7)
str(gw_ts)
```

```
##  Time-Series [1:630] from 2004 to 2016: -0.44 -0.474 -0.423 -0.551 -0.486 -0.551 -0.453 -0.462 -0.55
```





As shown in the above figure, there is no significant change of the interested levels between 2004 and 2012. After 2012, there is a claer upward trend for the interested level for the globle warming. We extract a subset of data (from 2012 to 2016) for the subsequent study. In term of anaual seasonal change, there are abrupt changes of interest level at the end of June and December.



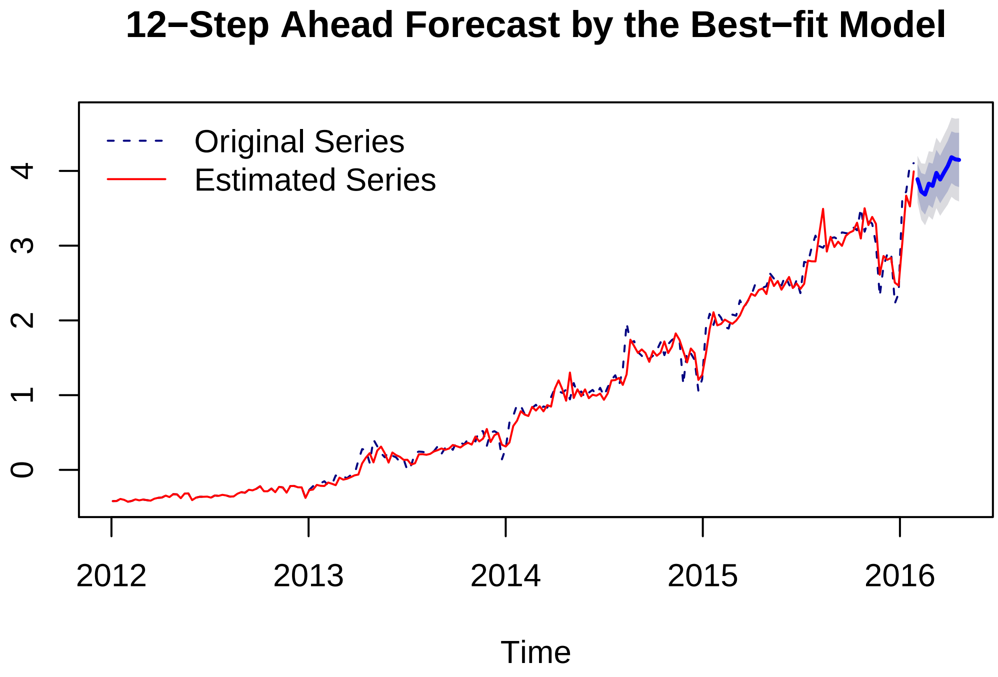
Instead, similar to the random walk, its ACF does not tail off quickly and many lags are highly correlated. Moreover, the variation in the differenced data series is not stationary as well. Therefore, the global warming data series itself is not stationary. We need to remove the trend and perform transformation of first-order differencing. Besides ACF and PACF plots, we also checked the first-order difference data as well as their ACF and PACF plots. There are still several significant lags observed in ACF and PACF plots of both original and differenced data series. we decided to use arima models to fit the data including certain seasonal effects.

first, we fit the data with the auto.arima function and obtain a complex model (ARIMA(1,1,1)(0,1,1)[52]). We also manually tried serval simple arima models with few parameters and select the best-fit model based on reported AICs.

```
# Best fit arima model and 12-step ahead forecast
gw_ts_fit <- auto.arima(gw_ts_new)
gw_ts_fit.fcast<-forecast.Arima(gw_ts_fit, h = 12)
gw_ts_fit
```

```
## Series: gw_ts_new
## ARIMA(1,1,1)(0,1,1)[52]
##
## Coefficients:
##          ar1      ma1     sma1
##       0.4724  -0.7973  -0.1720
## s.e.  0.1608   0.1201   0.0856
##
## sigma^2 estimated as 0.0255:  log likelihood=65.56
## AIC=-123.12   AICc=-122.86   BIC=-110.82
```

```
plot(gw_ts_fit.fcast, main="12-Step Ahead Forecast by the Best-fit Model",
     xlab="Time", xlim=c(), lty=2, col="navy")
lines(fitted(gw_ts_fit),col="red" )
leg.txt <- c("Original Series", "Estimated Series")
legend("topleft", legend=leg.txt, lty=c(2,1),
       col=c("navy","red"), bty='n', cex=1)
```



```
# Compare the AIC values of manually selected arima models:
# arima(0,1,0), arima(1,1,0), arima(0,1,1), arima(1,1,1), arima(1,1,1)
# arima(0,1,2), arima(1,1,2), arima(2,1,0), arima(2,1,2)
arima010<- arima(gw_ts_new, order=c(0,1,0))
```

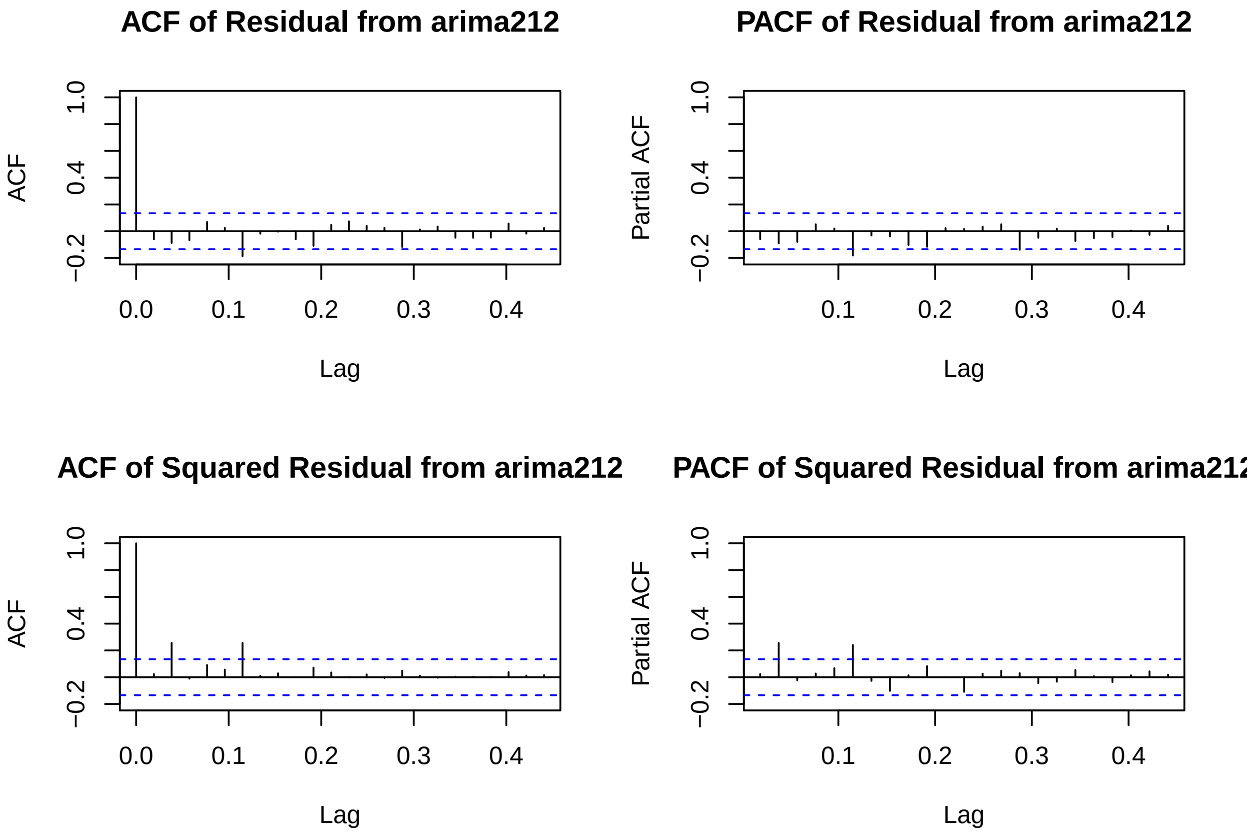
```
arima110<- arima(gw_ts_new, order=c(1,1,0))
arima011<- arima(gw_ts_new, order=c(0,1,1))
arima111<- arima(gw_ts_new, order=c(1,1,1))
arima012<- arima(gw_ts_new, order=c(0,1,2))
arima112<- arima(gw_ts_new, order=c(1,1,2))
arima210<- arima(gw_ts_new, order=c(2,1,0))
arima212<- arima(gw_ts_new, order=c(2,1,2))
aic_values <- c(arima010$aic, arima110$aic, arima011$aic, arima111$aic, arima012$aic,
                arima012$aic, arima112$aic, arima210$aic, arima212$aic)
aic_values
```

```
## [1] -125.3747 -124.5368 -125.2054 -127.9934 -129.4611 -129.4611 -127.8911
## [8] -129.3998 -137.0199
```

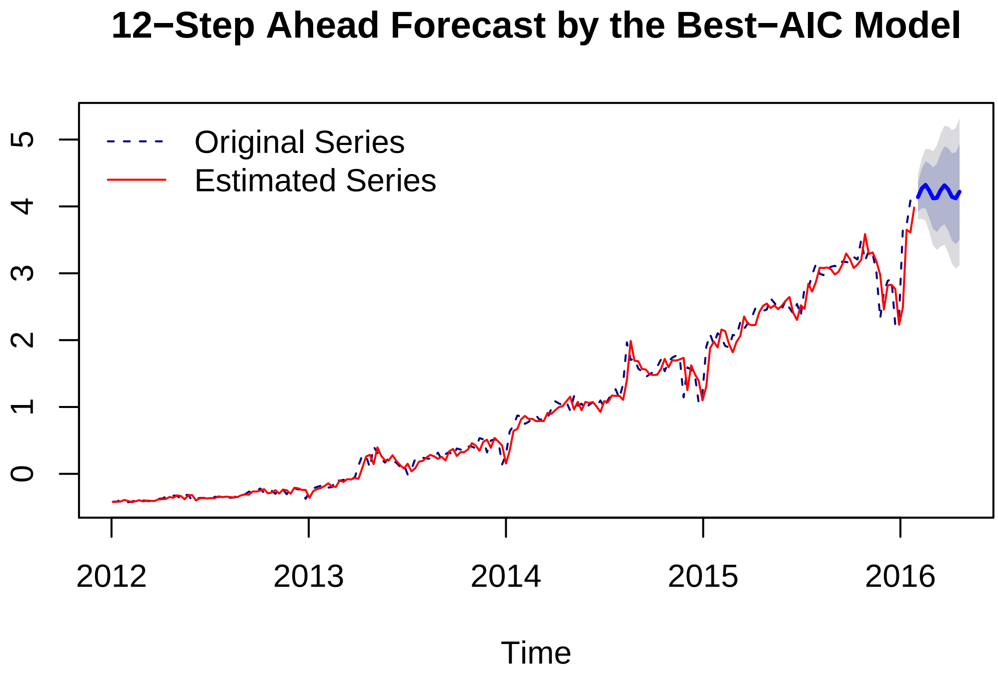
```
summary(arima212)
```

```
##
## Call:
## arima(x = gw_ts_new, order = c(2, 1, 2))
##
## Coefficients:
##          ar1      ar2      ma1     ma2
##       0.6944  -0.9734  -0.7785  1.0000
## s.e.  0.0175   0.0196   0.0191  0.0202
##
## sigma^2 estimated as 0.02852:  log likelihood = 73.51,  aic = -137.02
##
## Training set error measures:
##                      ME      RMSE       MAE      MPE     MAPE      MASE
## Training set 0.02275938 0.1684955 0.0990352 6.734563 22.47369 0.9822898
##                     ACF1
## Training set -0.05977991
```

According to the above aic values, we select arima(2,1,2) for further diagnostic checking. Both ACF and PACF of arima(2,1,2) model residues show no significant lags. Moreover, there is no significant labs observed within ACF and PACF plots of its square resudue. At this satge, we assume the selected simple arima(2,1,2) model is apprpropiate for the final prediction as well.



```r
# Forecast using the arima model with the best AIC value
gw_ts_bestaic.fcast<-forecast.Arima(arima212, h = 12)
plot(gw_ts_bestaic.fcast, main="12-Step Ahead Forecast by the Best-AIC Model",
     xlab="Time", xlim=c(), lty=2, col="navy")
lines(fitted(arima212),col="red" )
leg.txt <- c("Original Series", "Estimated Series")
legend("topleft", legend=leg.txt, lty=c(2,1),
       col=c("navy","red"), bty='n', cex=1)
```


12−Step Ahead Forecast by the Best−AIC Model
Original Series
Estimated Series
0
1
2
3
4
5
2012
2013
2014
2015
2016
Time

Part 4

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 exericse, 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.