ENV 790.30 - Time Series Analysis for Energy Data | Spring 2023

Assignment 7 - Due date 03/20/23

Maggie O'Shea

Directions

You should open the .rmd file corresponding to this assignment on RStudio. The file is available on our class repository on Github. And to do so you will need to fork our repository and link it to your RStudio.

Once you have the file open on your local machine the first thing you will do is rename the file such that it includes your first and last name (e.g., "LuanaLima_TSA_A07_Sp23.Rmd"). Then change "Student Name" on line 4 with your name.

Then you will start working through the assignment by **creating code and output** that answer each question. Be sure to use this assignment document. Your report should contain the answer to each question and any plots/tables you obtained (when applicable).

When you have completed the assignment, **Knit** the text and code into a single PDF file. Submit this pdf using Sakai.

Set up

```
#Load/install required package here
library(lubridate)
library(ggplot2)
library(forecast)
library(Kendall)
library(tseries)
library(outliers)
library(tidyverse)

#install.packages("smooth")
library(smooth)
```

Importing and processing the data set

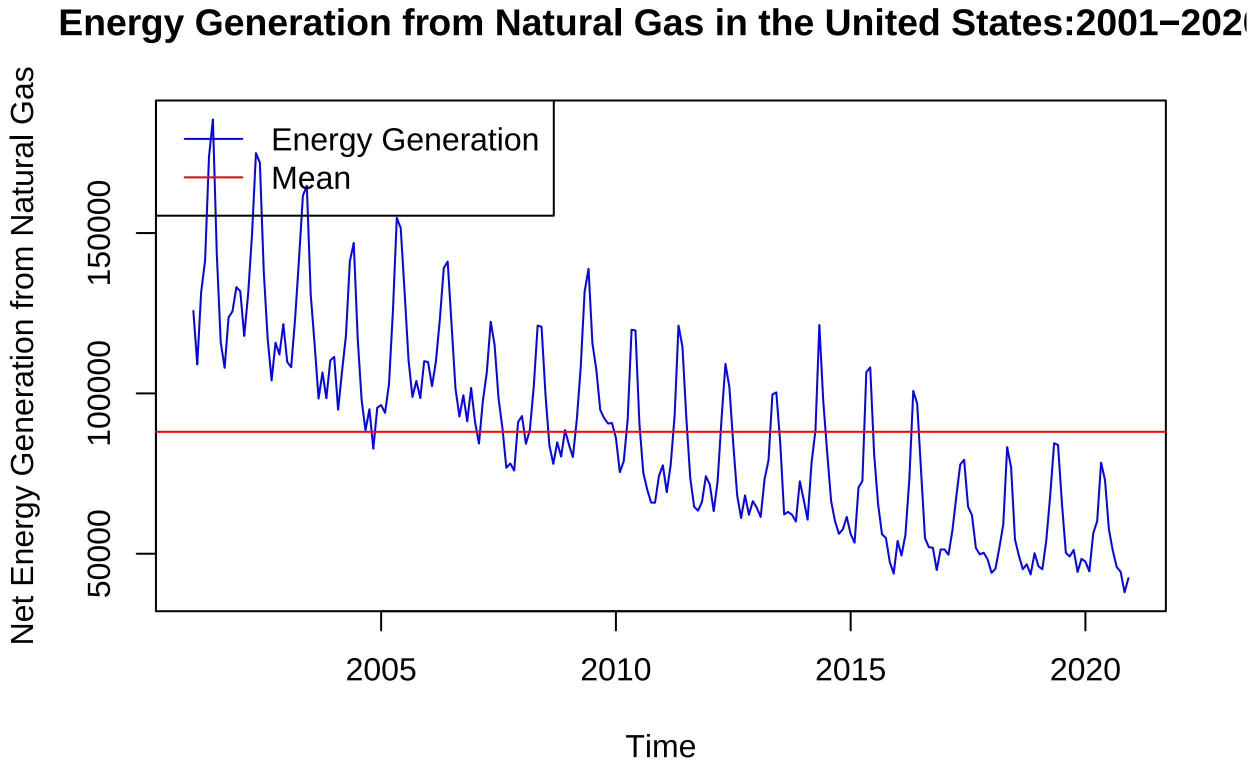
Consider the data from the file "Net_generation_United_States_all_sectors_monthly.csv". The data corresponds to the monthly net generation from January 2001 to December 2020 by source and is provided by the US Energy Information and Administration. **You will work with the natural gas column only**.

Packages needed for this assignment: "forecast","tseries". Do not forget to load them before running your script, since they are NOT default packages.\

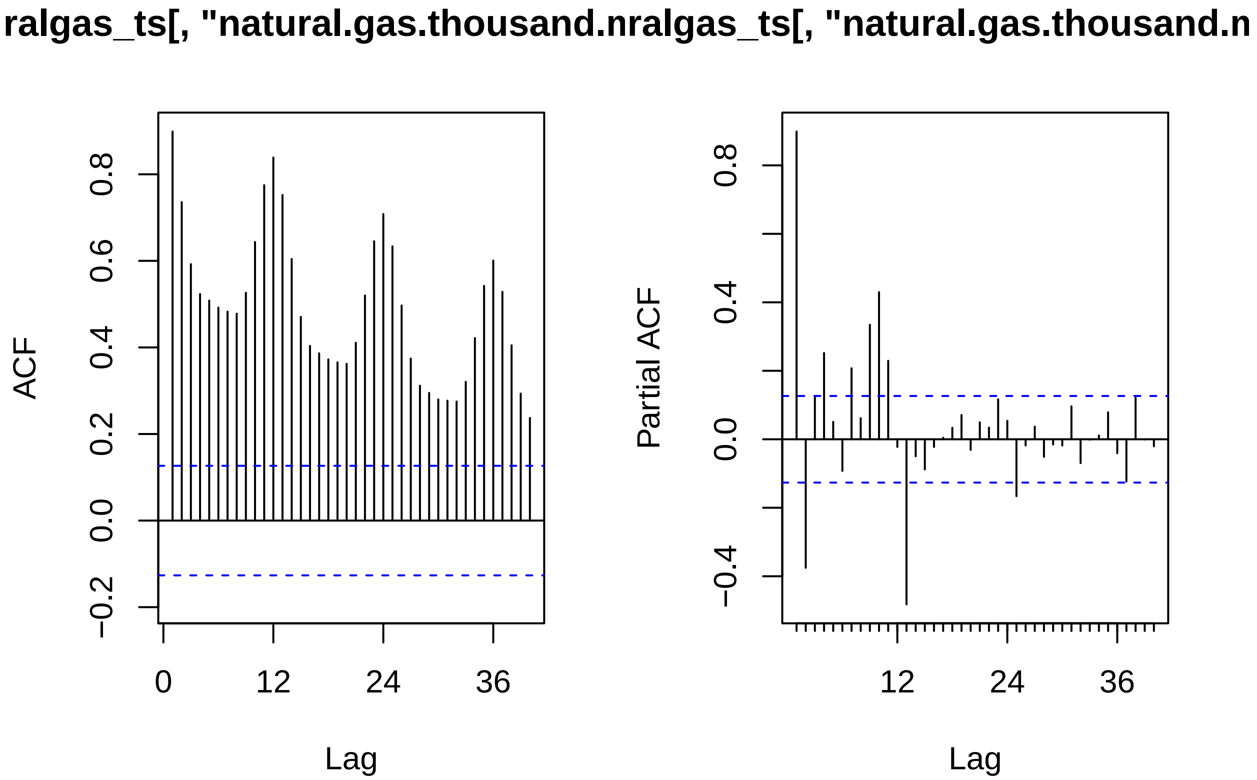
Q1

Import the csv file and create a time series object for natural gas. Make you sure you specify the **start=** and **frequency=** arguments. Plot the time series over time, ACF and PACF.

```
energy_generation <- read.csv("./Data/Net_generation_United_States_all_sectors_monthly.csv", skip = 4)%
  select(Month, natural.gas.thousand.megawatthours)
naturalgas_ts <- ts(energy_generation, start=c(2001,1),
                           frequency=12)

plot(naturalgas_ts[,"natural.gas.thousand.megawatthours"],type="l",col="blue",
     ylab="Net Energy Generation from Natural Gas",xlab="Time",
     main="Energy Generation from Natural Gas in the United States:2001-2020")
abline(h=mean(naturalgas_ts[,"natural.gas.thousand.megawatthours"]),col="red")
legend("topleft",legend=c("Energy Generation","Mean"),
       lty=c("solid","solid"),col=c("blue","red"))
```



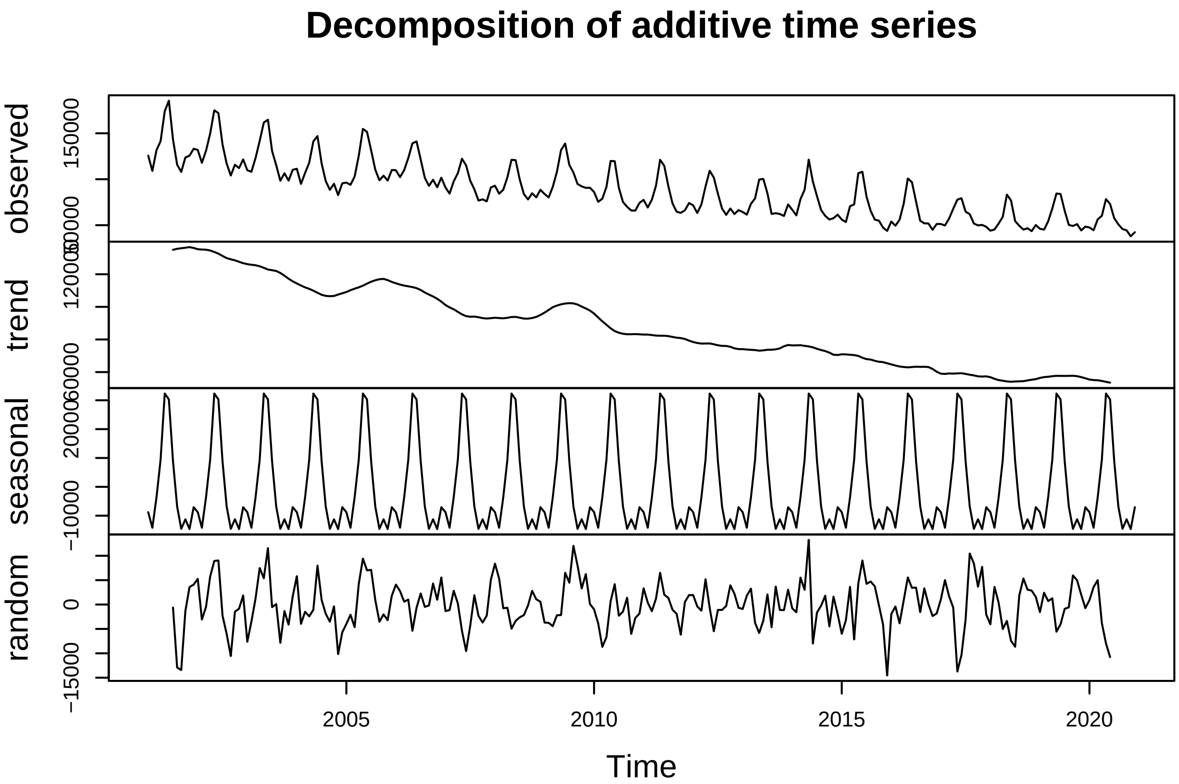
```
#ACF and PACF plots
par(mfrow=c(1,2))

ACF_Plot <- Acf(naturalgas_ts[,"natural.gas.thousand.megawatthours"], lag = 40, plot = TRUE)
PACF_Plot <- Pacf(naturalgas_ts[,"natural.gas.thousand.megawatthours"], lag = 40)
```



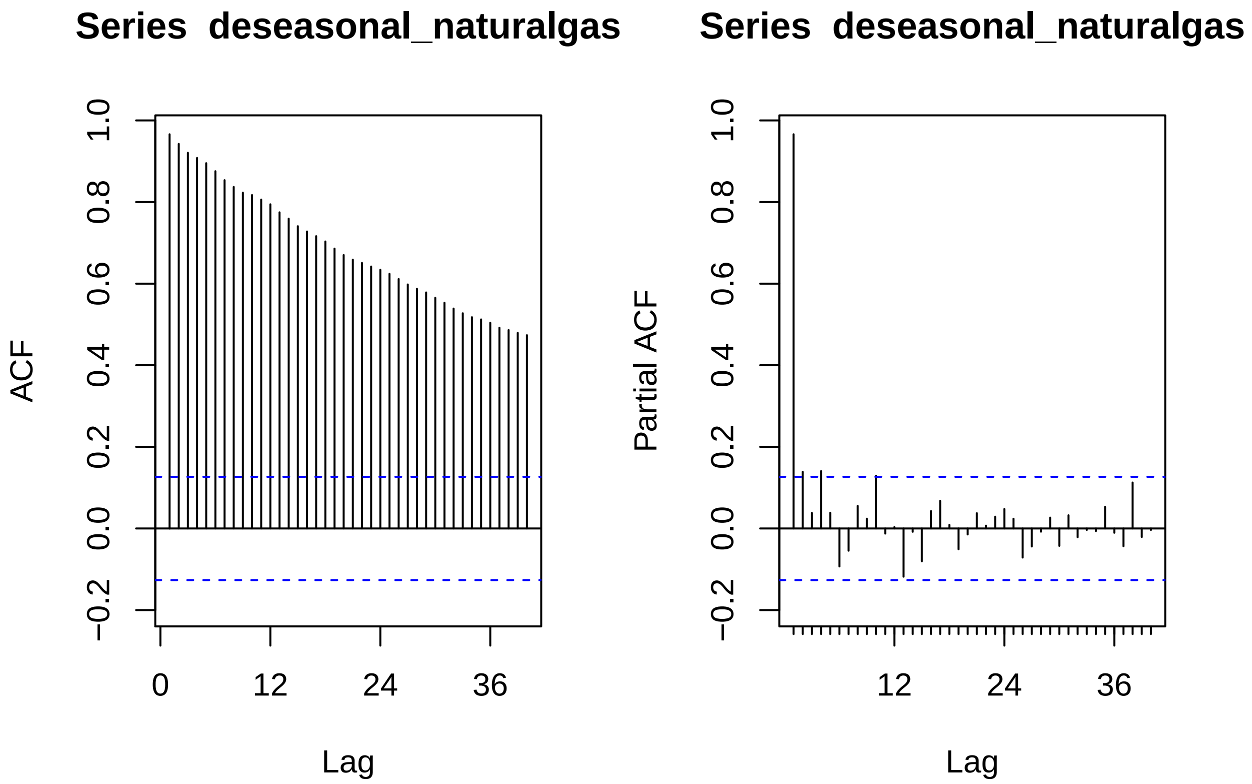
Q2

Using the *decompose*() or *stl*() and the *seasadj*() functions create a series without the seasonal component, i.e., a deseasonalized natural gas series. Plot the deseasonalized series over time and corresponding ACF and PACF. Compare with the plots obtained in Q1.

```
decompose_natgas <- decompose(naturalgas_ts[,"natural.gas.thousand.megawatthours"],"additive")
plot(decompose_natgas)
```



```
deseasonal_naturalgas <- seasadj(decompose_natgas)

par(mfrow=c(1,2))
ACF_Plot_deseas <- Acf(deseasonal_naturalgas, lag = 40, plot = TRUE)
PACF_Plot_deseas <- Pacf(deseasonal_naturalgas, lag = 40)
```



### Q2 Answer: In comparing the plots from Q2 with Q1, it is clear that the seasonality that was present in Q1 plots is no longer present. The ACF in Q1 had waves suggesting seasonality, and Q2 ACF does not have that. Additionally, the PACF in Q1 shows a high correlation at lag 12, that does not appear as present in the PACF for Q2. Finally, the new ACF plot (Q2) shows a slow decay which is a sign of non-stationarity.

Modeling the seasonally adjusted or deseasonalized series

Q3

Run the ADF test and Mann Kendall test on the deseasonalized data from Q2. Report and explain the results.

```r
#ADF Test
print(adf.test(deseasonal_naturalgas,alternative = "stationary"))
```

```
## Warning in adf.test(deseasonal_naturalgas, alternative = "stationary"): p-value
## smaller than printed p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  deseasonal_naturalgas
## Dickey-Fuller = -4.0574, Lag order = 6, p-value = 0.01
## alternative hypothesis: stationary
```

```r
#Mann Kendall
summary(MannKendall(deseasonal_naturalgas))
```

```
## Score =  -24186 , Var(Score) = 1545533
## denominator =  28680
## tau = -0.843, 2-sided pvalue =< 2.22e-16
```

Q3 Answer (Results) The ADF test had a p-value of <0.01 which indicates that we can reject the null hypothesis, and thus there is statistical evidence that the series is not stationary. The Mann Kendall aligned with this finding as it had a p-value of <0.01, suggesting again that the data does have a trend.

Q4

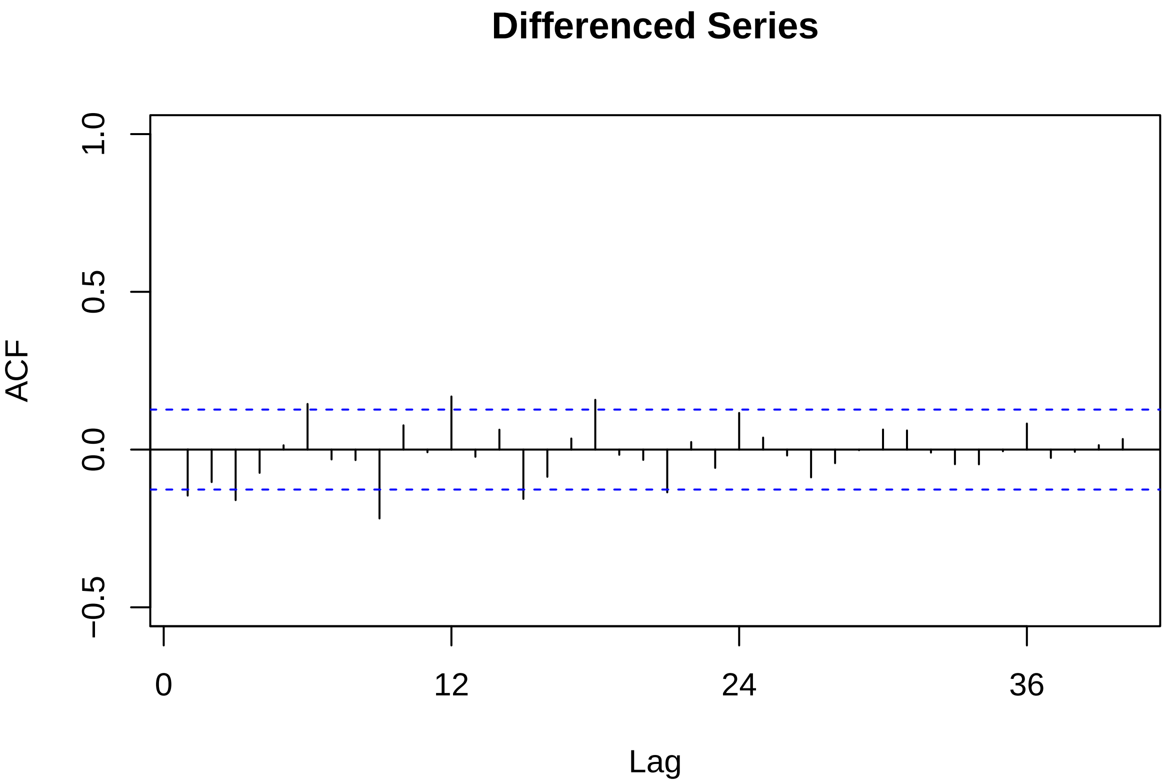
Using the plots from Q2 and test results from Q3 identify the ARIMA model parameters p, d and q. Note that in this case because you removed the seasonal component prior to identifying the model you don't need to worry about seasonal component. Clearly state your criteria and any additional function in R you might use. DO NOT use the *auto.arima*() function. You will be evaluated on ability to can read the plots and interpret the test results.

```
deseasonal_diff1 <- diff(deseasonal_naturalgas,differences=1,lag=1)
ndiffs(deseasonal_naturalgas)
```

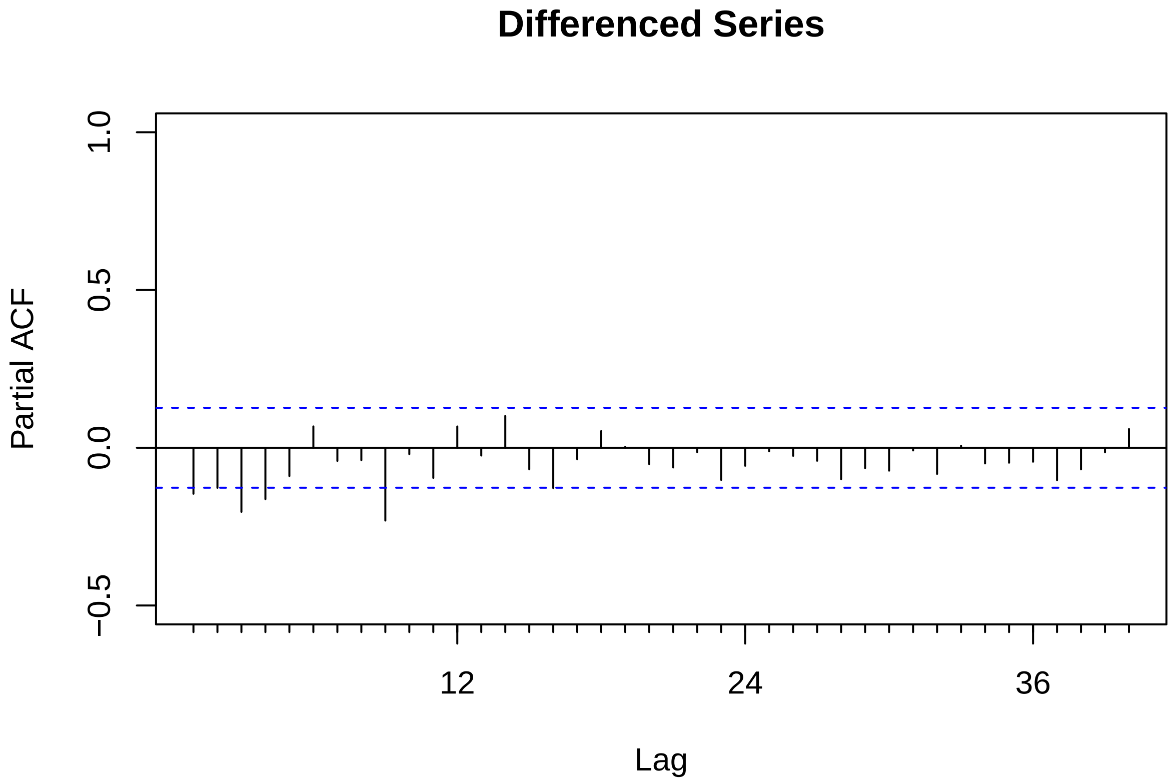
Q4 Answer/Discussion

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

```
Acf(deseasonal_diff1,lag.max=40,main="Differenced Series",ylim=c(-.5,1))
```



```
Pacf(deseasonal_diff1,lag.max=40,main="Differenced Series",ylim=c(-.5,1))
```



Q4 Discussion

- Parameter p is 3 based on the PACF of the differenced series – this plot shows that after lag 3 there is a cut off where the lags no longer exceed the blue dotted line boundary suggesting p should be 3.
- Parameter d is 1 because the stationary tests indicated that there is a trend that no longer was present when differenced once. At lag1 autocorrelation is near zero and negative, so the series does not need a higher order of differencing. This can be further validated by the R function ndiffs which also suggests d=1.
- Parameter q is determined based on the ACF of the differenced series. This plot did not show the first lag with a correlation much greater than the blue line, and in addition, has a relatively slow decay (rather than a cut off) especially evident in lags 1-5, which suggests an AR process making the MA term q to be 0. Given this evidence, I interpret this plot to mean q=0.

Q5

Use *Arima*() from package "forecast" to fit an ARIMA model to your series considering the order estimated in Q4. You should allow constants in the model, i.e., $include.mean = TRUE$ or $include.drift = TRUE$. **Print the coefficients** in your report.

```
ARIMA_manual <- Arima(deseasonal_naturalgas,order=c(3,1,0),include.drift=TRUE)
print(ARIMA_manual)
```

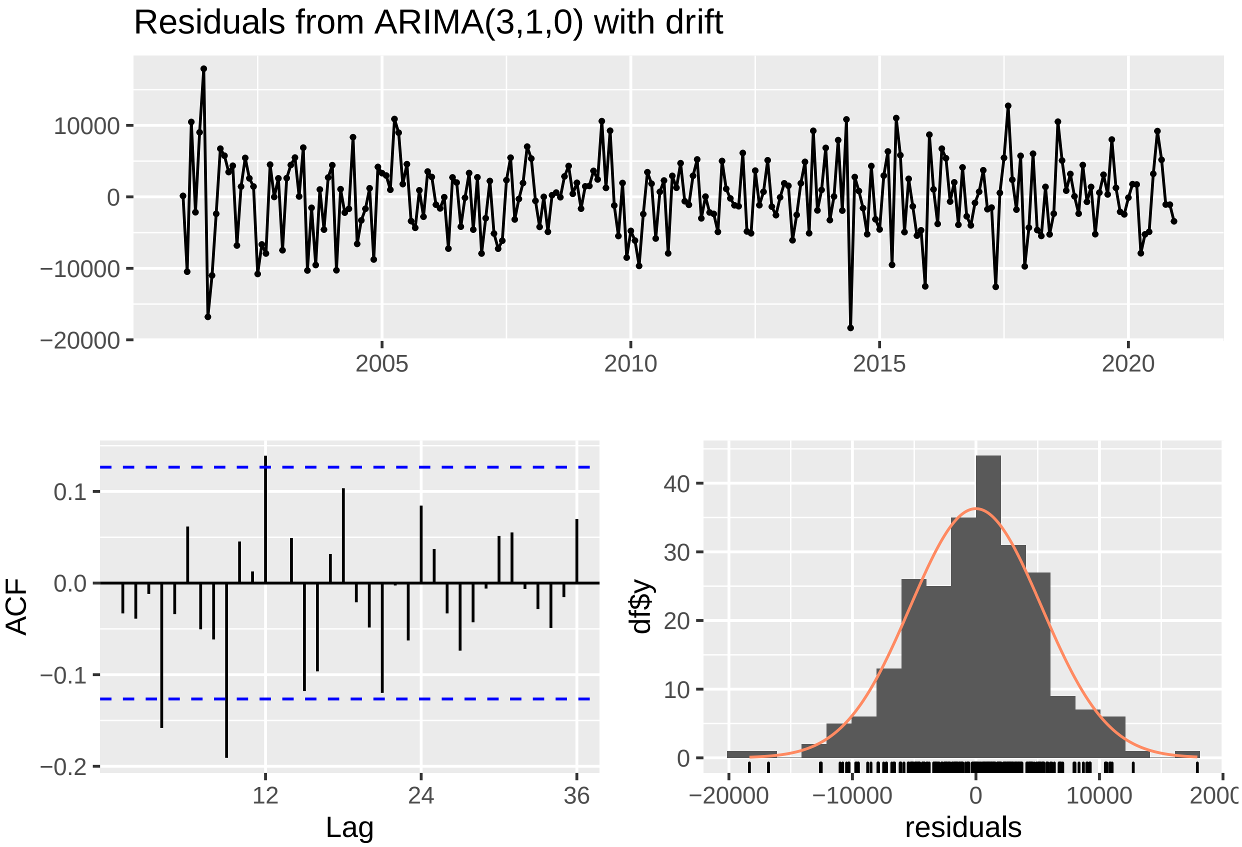
```
## Series: deseasonal_naturalgas
## ARIMA(3,1,0) with drift
```

```
##
## Coefficients:
##           ar1      ar2      ar3      drift
##       -0.1920  -0.1592  -0.2078  -340.6799
## s.e.   0.0637   0.0643   0.0642   221.2619
##
## sigma^2 = 28742263:  log likelihood = -2389.48
## AIC=4788.96   AICc=4789.22   BIC=4806.34
```

Q6

Now plot the residuals of the ARIMA fit from Q5 along with residuals ACF and PACF on the same window. You may use the *checkresiduals*() function to automatically generate the three plots. Do the residual series look like a white noise series? Why?

```
checkresiduals(ARIMA_manual)
```



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(3,1,0) with drift
## Q* = 41.469, df = 21, p-value = 0.004903
##
## Model df: 3.   Total lags used: 24
```

Q6 Answer The residuals do not look exactly like white noise - there are certainly spikes including, for example, before 2015. Also, at the start of the residuals plot they vary from high to low much more, and then seem to decrease in variation between 2005 and 2015. This potential irregularity also shows in the other plots - the ACF does not have all lines within the blue dotted boundaries, and the histogram shows what looks like an outlier (or a left tail) at -20000 and perhaps another one near +20000. They don't look like white noise possibly because there is irregularity that the model did not capture which means that the model may need to be improved.

Modeling the original series (with seasonality)

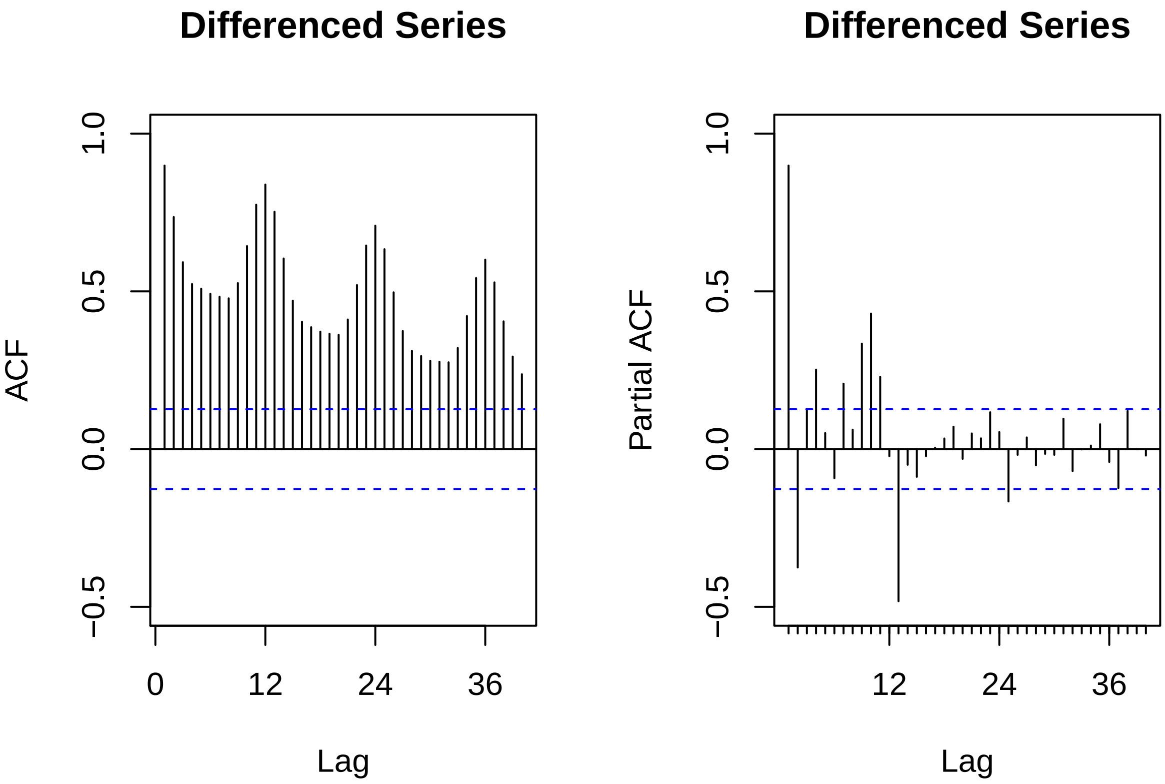
Q7

Repeat Q4-Q6 for the original series (the complete series that has the seasonal component). Note that when you model the seasonal series, you need to specify the seasonal part of the ARIMA model as well, i.e., P, D and Q.

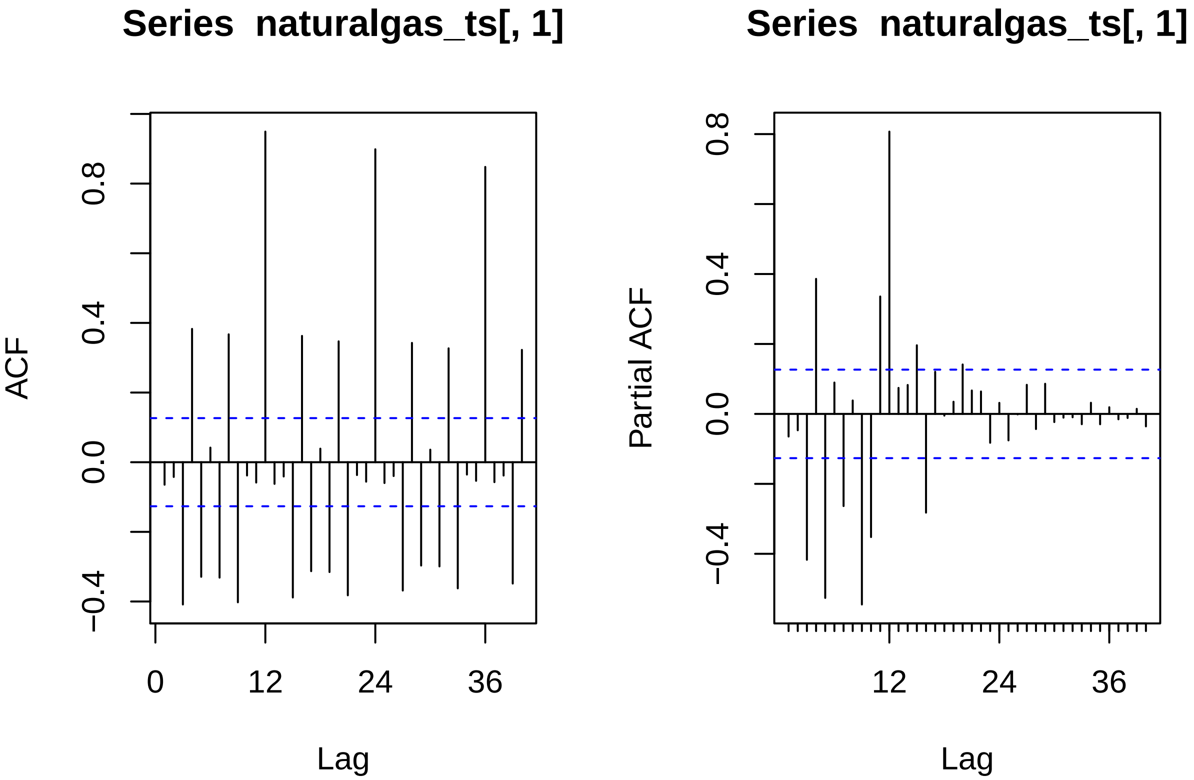
```
# First, identify p,d,q on original data
ndiffs(naturalgas_ts[,"natural.gas.thousand.megawatthours"])
```

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

```
#d=1
par(mfrow=c(1,2))
Acf(naturalgas_ts[,"natural.gas.thousand.megawatthours"],lag.max=40,main="Differenced Series",ylim=c(-.
Pacf(naturalgas_ts[,"natural.gas.thousand.megawatthours"],lag.max=40,main="Differenced Series",ylim=c(-
```



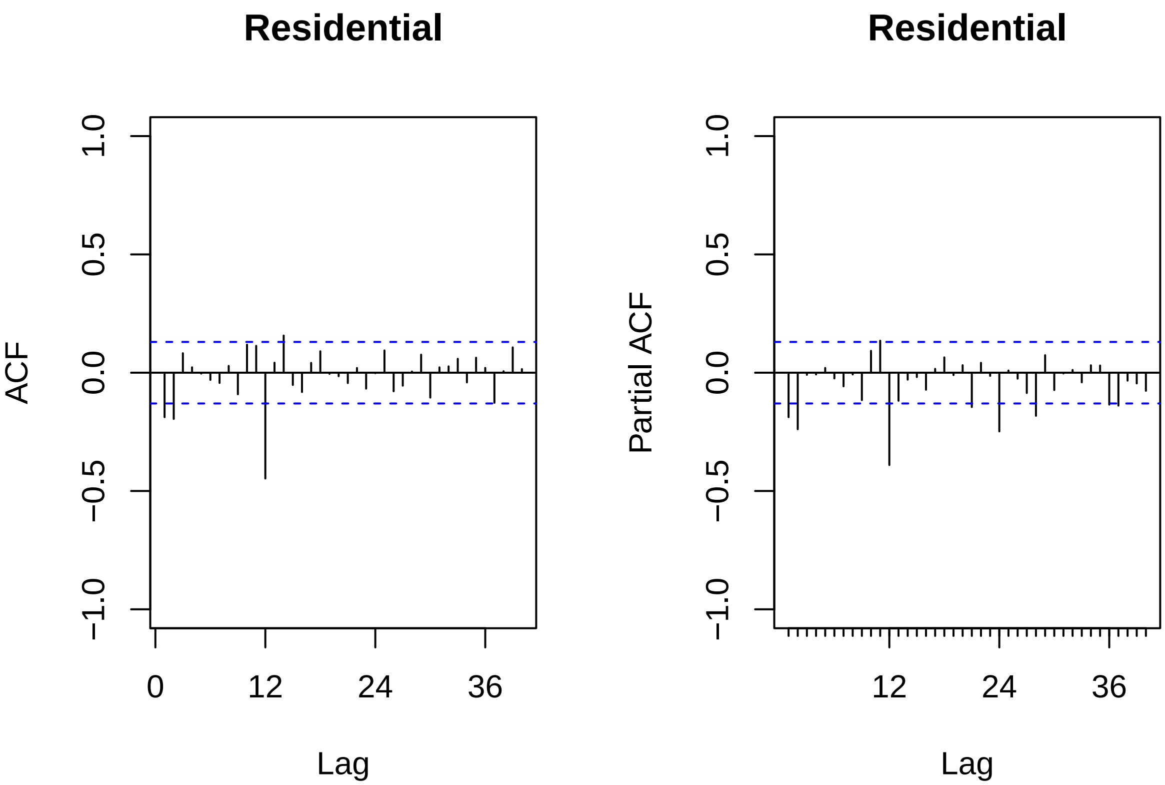
```
#ACF/PACF on original data to identify if seasonal data
par(mfrow=c(1,2))
Acf(naturalgas_ts[,1], lag = 40)
Pacf(naturalgas_ts[,1], lag = 40)
```



```
#nsdiffs indicates how many seasonal differences
nsdiffs(naturalgas_ts[,1], max.D=1)
```

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

```
#Difference for seasonal component from already trend-differenced data
trend_diff <- diff(naturalgas_ts[,"natural.gas.thousand.megawatthours"],lag =1, differences=1) #differe
both_diff <- diff(trend_diff,lag =12, differences=1)

#New ACF/PACF
par(mfrow=c(1,2))
Acf(both_diff,lag.max=40,main="Residential",ylim=c(-1,1))
Pacf(both_diff,lag.max=40,main="Residential",ylim=c(-1,1))
```



```
#Run Seasonal ARIMA
SARIMA_manual <- Arima(naturalgas_ts[,"natural.gas.thousand.megawatthours"],order=c(1,1,0),seasonal=c(0
#Drift = False because differenced series twice so not allowing a constant (mu)
print(SARIMA_manual)
```

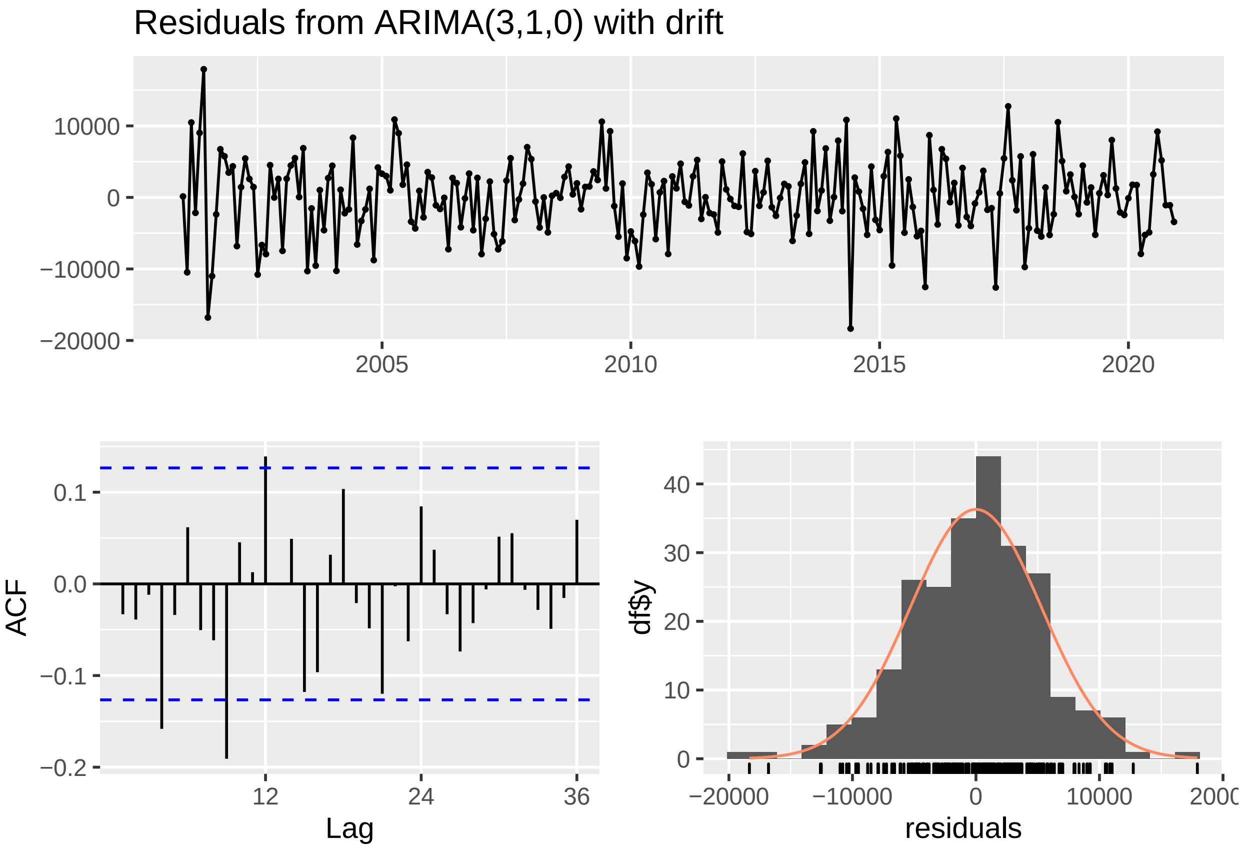
```
## Series: naturalgas_ts[, "natural.gas.thousand.megawatthours"]
## ARIMA(1,1,0)(0,1,1)[12]
##
## Coefficients:
##           ar1     sma1
##       -0.1808  -0.6898
## s.e.   0.0655   0.0557
##
## sigma^2 = 30627233:  log likelihood = -2281.43
## AIC=4568.86   AICc=4568.96   BIC=4579.13
```

Q7 Answer: First, I re-examined the plots to identify values for p, d, and q because the first ARIMA was on deseasoned data. I found that ndiffs on the original data suggested that d=1. Parameter p is based on the PACF which showed a cut off after lag 1 suggesting p=1. Parameter q shows a slow decay which again suggests an AR process indicating that q is 0. This would make p, d, q: 1, 1, 0. Next, addressing seasonality, the ACF/PACFs clearly show a seasonal component given that there are equally spaced lags especially visible in the ACF which are strong and stable over time. The function nsdiffs indicates the number of seasonal differences needed to achieve stationarity and when running this function the suggested number of seasonal differences is 1. Next, I performed the seasonal difference on the already trend differenced series to examine the ACF/PACF and ultimately determine what P and Q should be for the model. Because the ACF has just a single spike at lag 12, and the PACF has multiple spikes at seasonal lags (12, 24, etc.), this is likely an SMA process. This means that P is 0, and Q is likely 1. The result is a SARIMA model with p,d,q as 1, 1, 0 and P, D, Q, as 0, 1, 1.

Q8

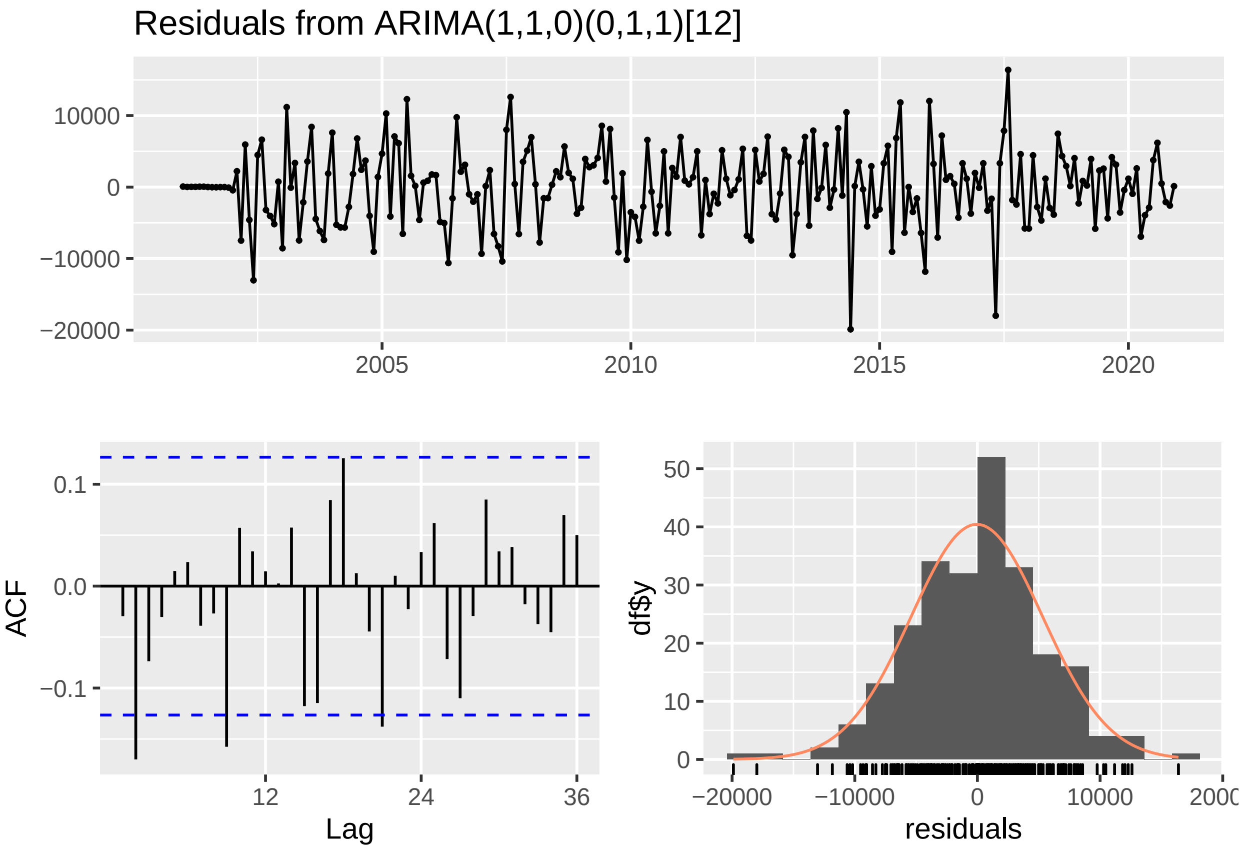
Compare the residual series for Q7 and Q6. Can you tell which ARIMA model is better representing the Natural Gas Series? Is that a fair comparison? Explain your response.

```
checkresiduals(ARIMA_manual)
```



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(3,1,0) with drift
## Q* = 41.469, df = 21, p-value = 0.004903
##
## Model df: 3.   Total lags used: 24
```

```
checkresiduals(SARIMA_manual)
```



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(1,1,0)(0,1,1)[12]
## Q* = 36.85, df = 22, p-value = 0.02457
##
## Model df: 2.   Total lags used: 24
```

Q8 Answer: The residuals for the SARIMA look slightly better particularly until 2015, however, the spikes are even more pronounced in the latter half of the time series (from 2015 and on). Still, based on these plots, the SARIMA is a better model for this data. I think it is possible to compare these because in any model we hope to find that the residuals do not suggest major irregularities in the model. Then, if one is diagnosed as a good model with this diagnostic test, and the other as a model is diagnosed as one that needs improvement then we can compare this finding and say perhaps the 'good' residuals plot may be a better fit of a model.

Checking your model with the auto.arima()

Please do not change your answers for Q4 and Q7 after you ran the *auto.arima*(). It is **ok** if you didn't get all orders correctly. You will not loose points for not having the same order as the *auto.arima*().

Q9

Use the *auto.arima*() command on the **deseasonalized series** to let R choose the model parameter for you. What's the order of the best ARIMA model? Does it match what you specified in Q4?

```
auto.arima(deseasonal_naturalgas)
```

```
## Series: deseasonal_naturalgas
## ARIMA(3,1,0)(1,0,1)[12] with drift
##
## Coefficients:
##           ar1      ar2      ar3    sar1     sma1      drift
##       -0.2028  -0.1851  -0.1378  0.6609  -0.4698  -331.8138
## s.e.   0.0645   0.0655   0.0682  0.1918   0.2120   328.8248
##
## sigma^2 = 27791547:  log likelihood = -2384.89
## AIC=4783.79   AICc=4784.27   BIC=4808.12
```

Q9 Answer: The auto.arima function did produce the same values as I did in Q4 for p,d, and q which were 3,1,0.

Q10

Use the *auto.arima*() command on the **original series** to let R choose the model parameters for you. Does it match what you specified in Q7?

```
auto.arima(naturalgas_ts[,"natural.gas.thousand.megawatthours"])
```

```
## Series: naturalgas_ts[, "natural.gas.thousand.megawatthours"]
## ARIMA(2,0,1)(2,1,2)[12] with drift
##
## Coefficients:
##          ar1      ar2      ma1     sar2     sar2     sma1    sma2      drift
##       1.1650  -0.2834  -0.4837  -0.0667  -0.0785  -0.6371  0.0072  -357.8435
## s.e.  0.4164   0.3180   0.3993   1.3218   0.1014   1.3211  0.9212    44.1285
##
## sigma^2 = 27958165:  log likelihood = -2278.46
## AIC=4574.91   AICc=4575.74   BIC=4605.77
```

Q10 Answer:

The findings from auto.arima with the original data did not match my findings from Q7. I had thought the parameters were ARIMA(1,1,0)(0,1,1) but auto.arima found that instead they should be (2,0,1)(2,1,2).