ENV 790.30 - Time Series Analysis for Energy Data | Spring 2022

Assignment 7 - Due date 03/25/22

Rob Kravec

Set up

```
library(tidyverse)
library(forecast)
library(tseries)
library(patchwork)
library(Kendall)
```

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**.

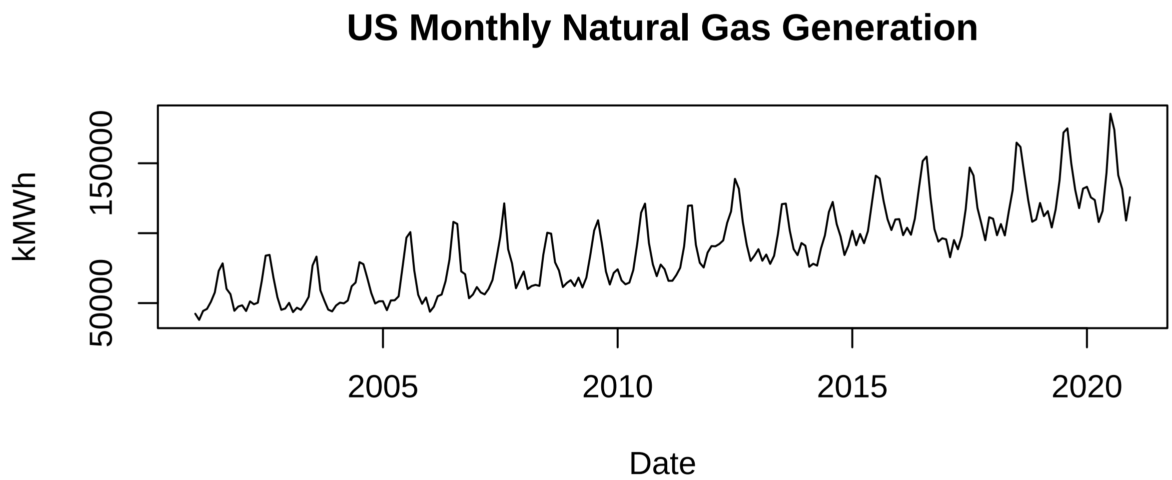
Q1

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.

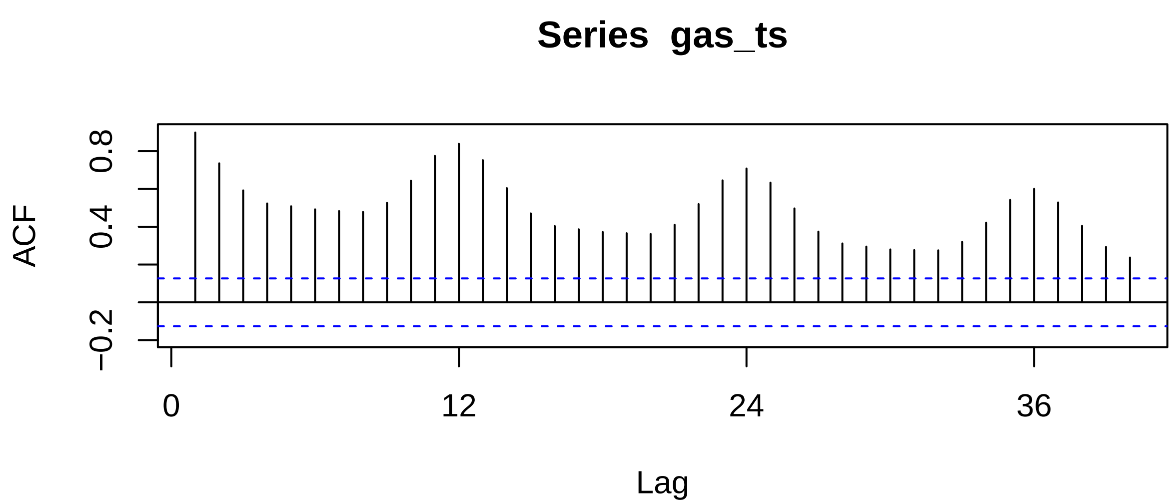
```
# Read in data
path <- "../Data/Net_generation_United_States_all_sectors_monthly.csv"
data <- read_csv(file = path, skip = 4)

# Extract relevant column
gas <- data %>%
  rename(gas = `natural gas thousand megawatthours`) %>%
  select(Month, gas) %>%
  map_df(rev)

# Create ts object
gas_ts <- ts(gas$gas, start = c(2001, 1), frequency = 12)
head(gas_ts)
```

```
##           Jan      Feb      Mar      Apr      May      Jun
## 2001 42388.66 37966.93 44364.41 45842.75 50934.21 57603.15
```

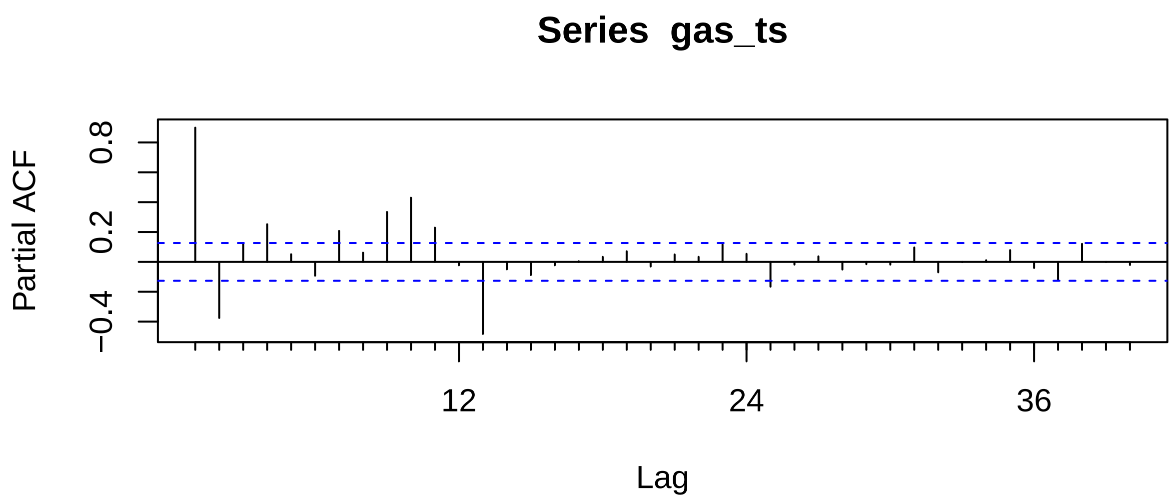
```
### Plots
# Time series plot
plot(gas_ts, main = "US Monthly Natural Gas Generation",
     xlab = "Date",
     ylab = "kMWh")
```




```r
# ACF and PACF
Acf(gas_ts, lag.max = 40)
```



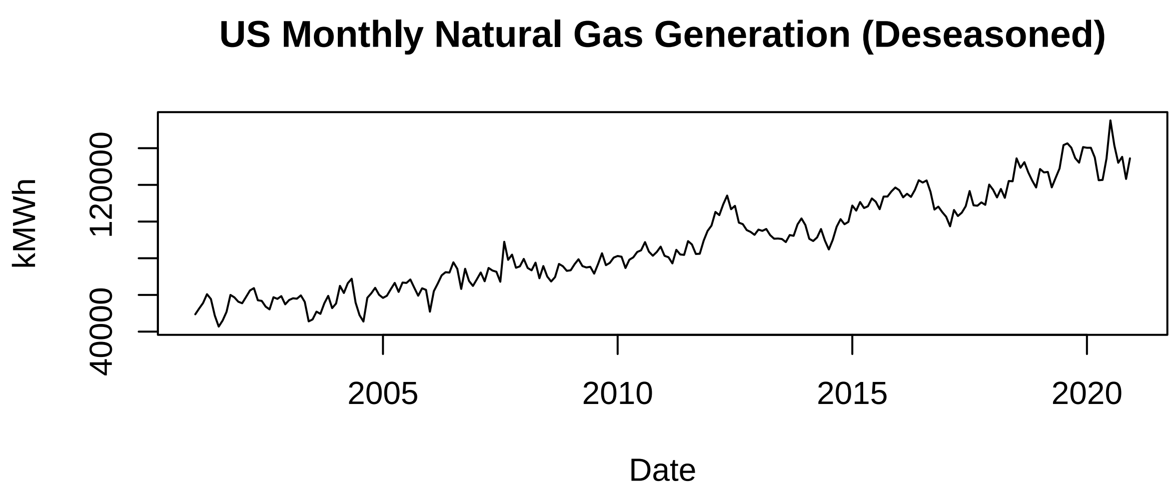

```r
Pacf(gas_ts, lag.max = 40)
```



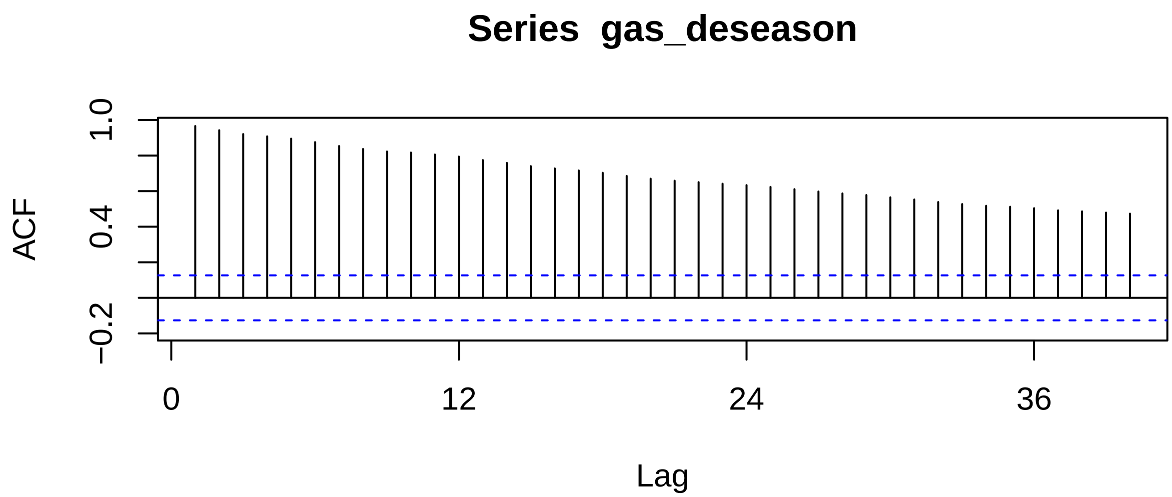
Q2

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.

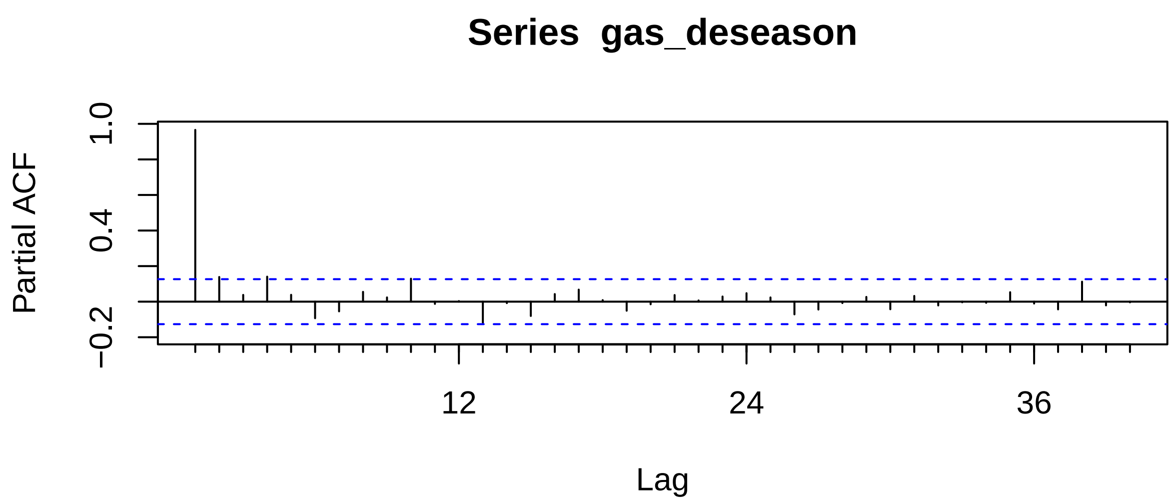
```
# Create deseasoned time series
gas_decomp <- decompose(gas_ts)
gas_deseason <- seasadj(gas_decomp)

### Generate same plots as in q1
# Time series plot
plot(gas_deseason, main = "US Monthly Natural Gas Generation (Deseasoned)",
     xlab = "Date",
     ylab = "kMWh")
```



```r
# ACF and PACF
Acf(gas_deseason, lag.max = 40)
```




```r
Pacf(gas_deseason, lag.max = 40)
```



Comparison:

- **Time Series**: The original time series displays a clear seasonal pattern, while the deseasoned time series appears to be governed more be random variations. There is an upward trend present in both series
- **ACF**: The ACF of the original series shows clear seasonality with its wave-like pattern. In contrast, the deseasoned series ACF shows a slow decay
- **PACF**: The PACF of the original series has values after lag 1 that exceed the standard error threshold (lag 13 is clearest example). In contrast, the deseasoned PACF has a clear cutoff after lag 1

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.

```
# ADF test
print(adf.test(gas_deseason,alternative = "stationary"))
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  gas_deseason
## Dickey-Fuller = -4.0271, Lag order = 6, p-value = 0.01
## alternative hypothesis: stationary
```

```
# Mann Kendall
print(summary(MannKendall(gas_deseason)))
```

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

ADF test: We reject the null hypothesis that the series contains a unit root. Therefore, we conclude that the series does not contain a stochastic trend

Mann-Kendall: We reject the null hypothesis that the series is stationary, concluding that the series contains a deterministic 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.

```
# Check need for differencing
ndiffs(gas_deseason)
```

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

The ACF plot shows a slow decay, while the PACF plot shows a cutoff after lag 1. These results suggest an AR(1) process.

While the ADF test concluded that the series does not contain a unit root, the `ndiffs` function suggests that the series should be differenced one time.

Therefore, I would expect this series to be modeled by an ARIMA(1, 1, 0).

Q5

Use *Arima*() from package "forecast" to fit an ARIMA model to your series considering the order estimated in Q4. Should you allow for constants in the model, i.e., $include.mean = TRUE$ or $include.drift = TRUE$. **Print the coefficients** in your report. Hint: use the *cat*() function to print.

```
q5 <- Arima(gas_deseason, order = c(1,1,0),
            include.drift = T)
q5
```

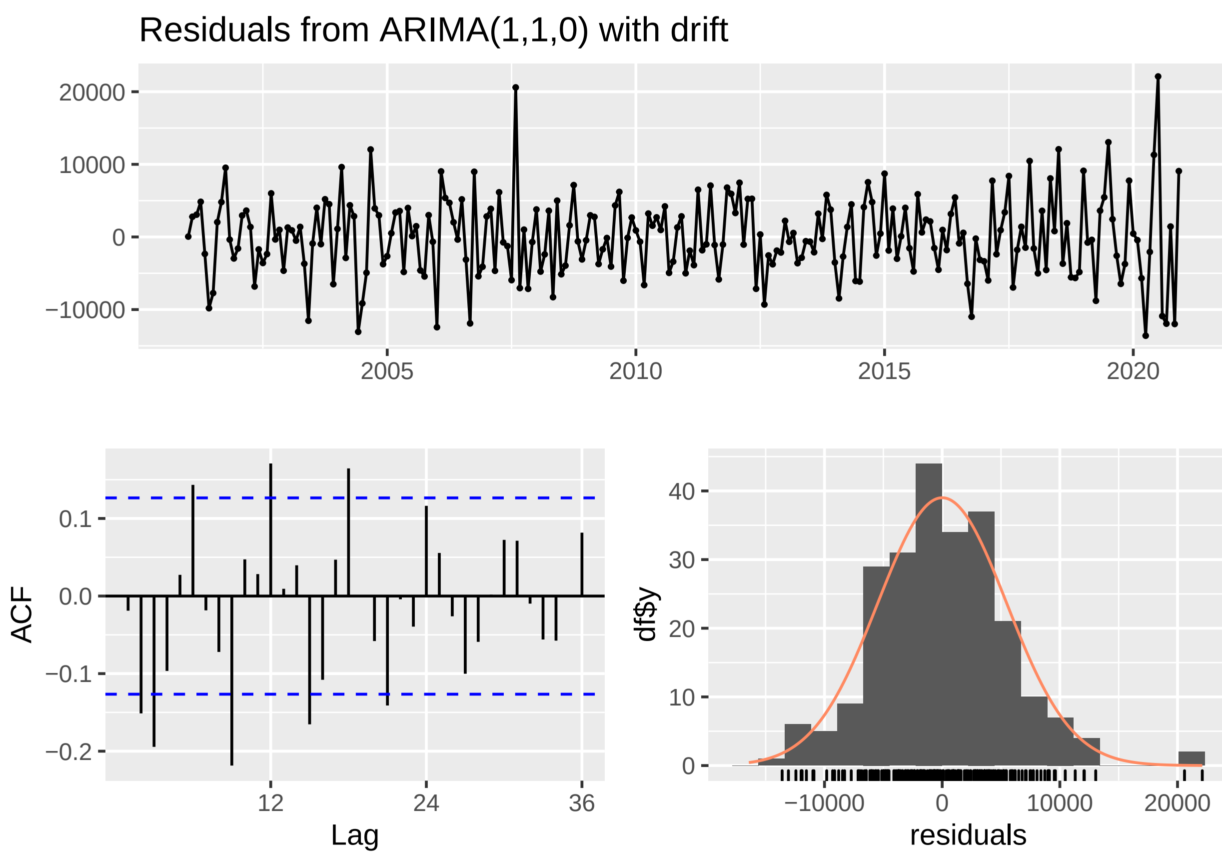
```
## Series: gas_deseason
## ARIMA(1,1,0) with drift
##
## Coefficients:
##           ar1     drift
##       -0.1479  348.3927
## s.e.   0.0644  308.8385
##
## sigma^2 = 30254066:  log likelihood = -2396.54
## AIC=4799.07   AICc=4799.18   BIC=4809.5
```

We don't need to include a mean term in the model because the series will be differenced once. Including a drift term allows for small deviations from a mean of zero, which is appropriate for a series that has only been differenced once.

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?

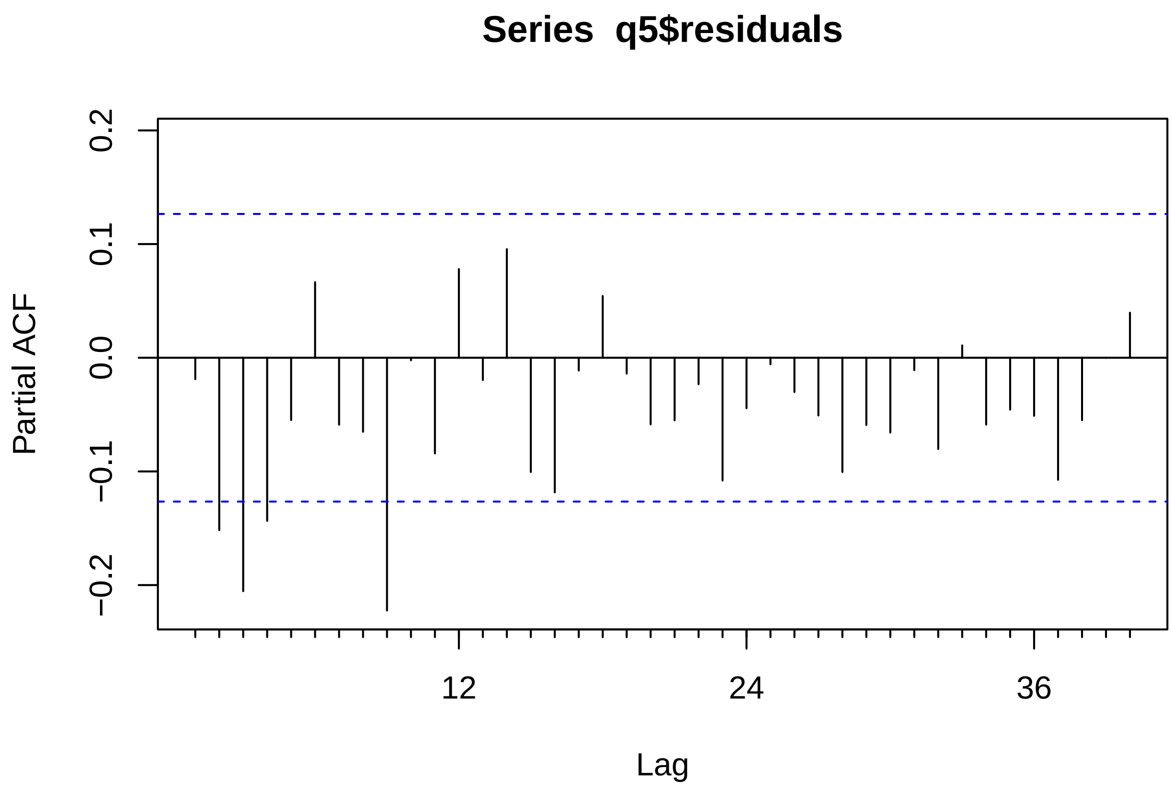
```
checkresiduals(q5)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(1,1,0) with drift
## Q* = 72.475, df = 22, p-value = 2.683e-07
```

```
##
## Model df: 2.   Total lags used: 24
```

```r
Pacf(q5$residuals, lag.max = 40)
```



While the residuals are approximately normally distributed, they don't quite look like random noise. The ACF plot has some values that are borderline larger than the standard error cutoffs, and the residuals vs. observation shows some "stickiness" (i.e., for every residual above zero the next residual is always below zero). If we truly had a random noise sequence, that likely wouldn't happen. As a result, I imagine that the model also needs a moving average term.

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.

```r
ndiffs(gas_ts)
```

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

```r
nsdiffs(gas_ts)
```

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

The `ndiffs` and `nsdiffs` functions indicate that $d = 1$ and $D = 1$, respectively.

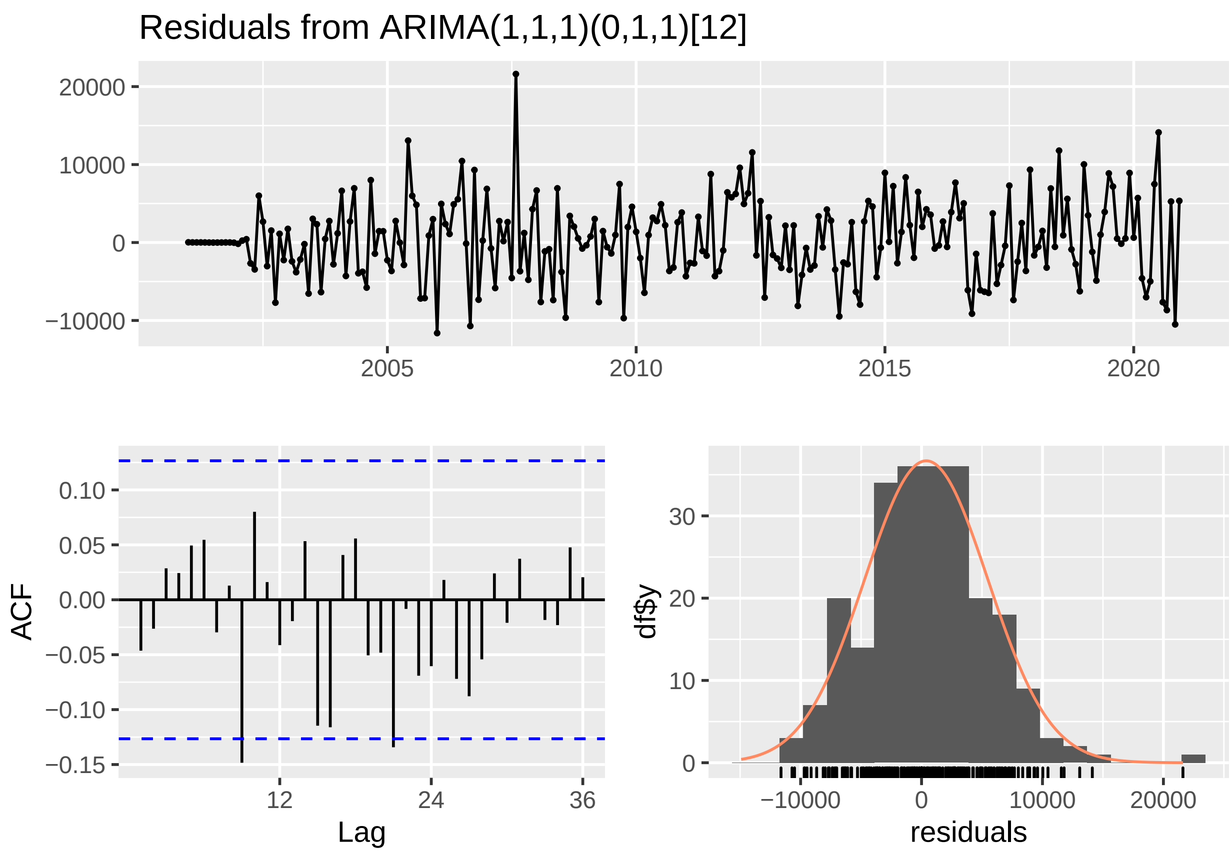
Given the residual analysis from question 6, I hypothesize that $p = 1$ and $q = 1$.

From the ACF plot in question 1, there appears to be a seasonal moving average term, as evidenced from the relatively large value at lag 12 ($Q = 1$).

```
# No drift included, per warning message from R about no drift term allowed
# for order of difference > 1
q7 <- Arima(gas_deseason, order = c(1,1,1), seasonal = c(0,1,1))
q7
```

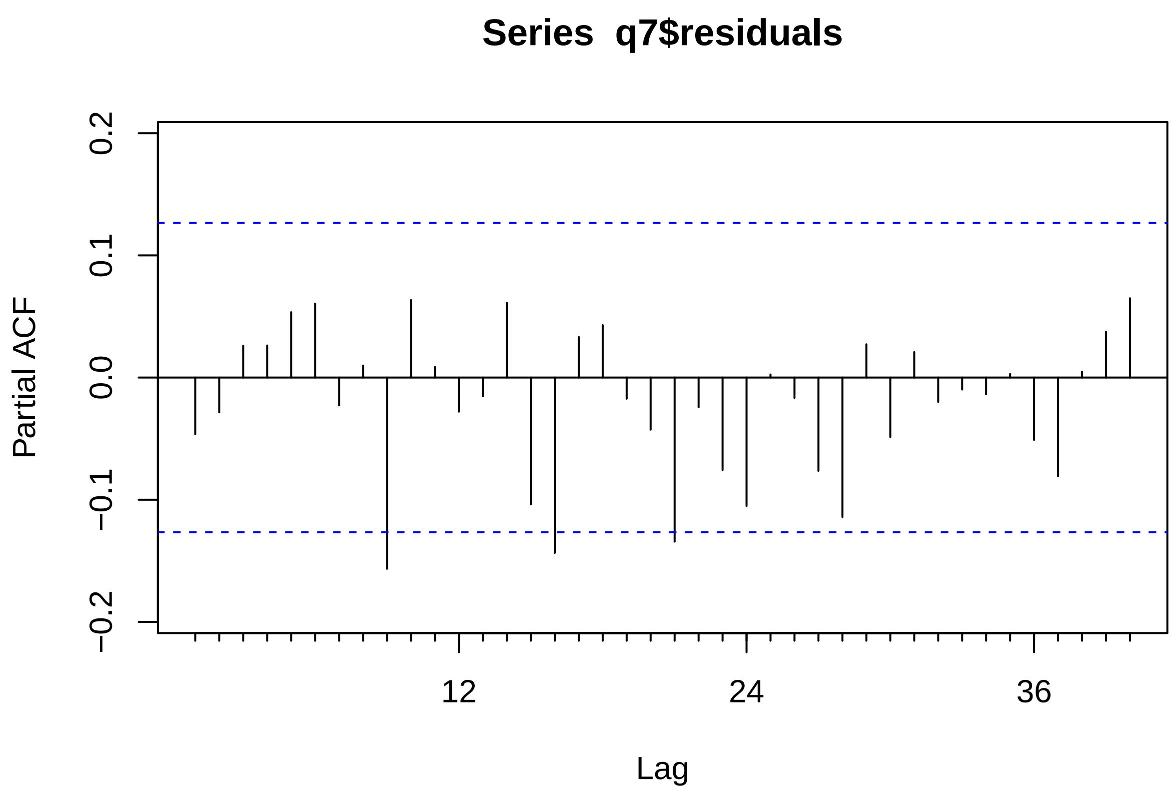
```
## Series: gas_deseason
## ARIMA(1,1,1)(0,1,1)[12]
##
## Coefficients:
##          ar1      ma1     sma1
##       0.7323  -0.9819  -0.7017
## s.e.  0.0504   0.0183   0.0563
##
## sigma^2 = 27922078:  log likelihood = -2272.2
## AIC=4552.39   AICc=4552.57   BIC=4566.09
```

```
checkresiduals(q7)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(1,1,1)(0,1,1)[12]
## Q* = 27.598, df = 21, p-value = 0.1519
##
## Model df: 3.   Total lags used: 24
```

```
Pacf(q7$residuals, lag.max = 40)
```



Relative to the residuals in question 6, these residuals look more like white noise, though they have some right skew in distribution. It's possible this model is also misspecified.

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.

I compared the two residual series in the previous question, noting that the q7 residuals look more like white noise, while the q6 residuals have a more normal-looking distribution. We cannot really use this comparision to determine which ARIMA model better represents the Natural Gas Series because the models were fit on two different versions of the series, one with the seasonality removed and one with the seasonality still present. To compare the performance of two different model specifications, the underlying series should be the same.

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 correct orders. The intention of the assignment is to walk you to the process and help you figure out what you did wrong (if you did anything wrong!).

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?

```r
q9 <- auto.arima(gas_deseason)
q9
```

```
## Series: gas_deseason
## ARIMA(1,1,1) with drift
##
## Coefficients:
##          ar1      ma1     drift
##       0.7065  -0.9795  359.5052
## s.e.  0.0633   0.0326   29.5277
##
## sigma^2 = 26980609:  log likelihood = -2383.11
## AIC=4774.21   AICc=4774.38   BIC=4788.12
```

The order is (1, 1, 1), which is not what I originally thought but is what I concluded after looking at the residuals of the (1, 1, 0) model.

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?

```r
q10 <- auto.arima(gas_ts)
q10
```

```
## Series: gas_ts
## ARIMA(1,0,0)(0,1,1)[12] with drift
##
## Coefficients:
##          ar1     sma1     drift
##       0.7416  -0.7026  358.7988
## s.e.  0.0442   0.0557   37.5875
##
## sigma^2 = 27569124:  log likelihood = -2279.54
## AIC=4567.08   AICc=4567.26   BIC=4580.8
```

The order is (1, 0, 0)(0, 1, 1)[12], which is certainly not what I expected! First off, the `ndiffs` function indicated that $d = 1$, so it is surprising to see a best fit with $d = 0$. Secondly, while I was able to identify the presence of the seasonal moving average term, I also included a non-seasonal moving average term that did not end up in the best model.