

ISyE6402__HW2

Jim Liu

10/4/2020

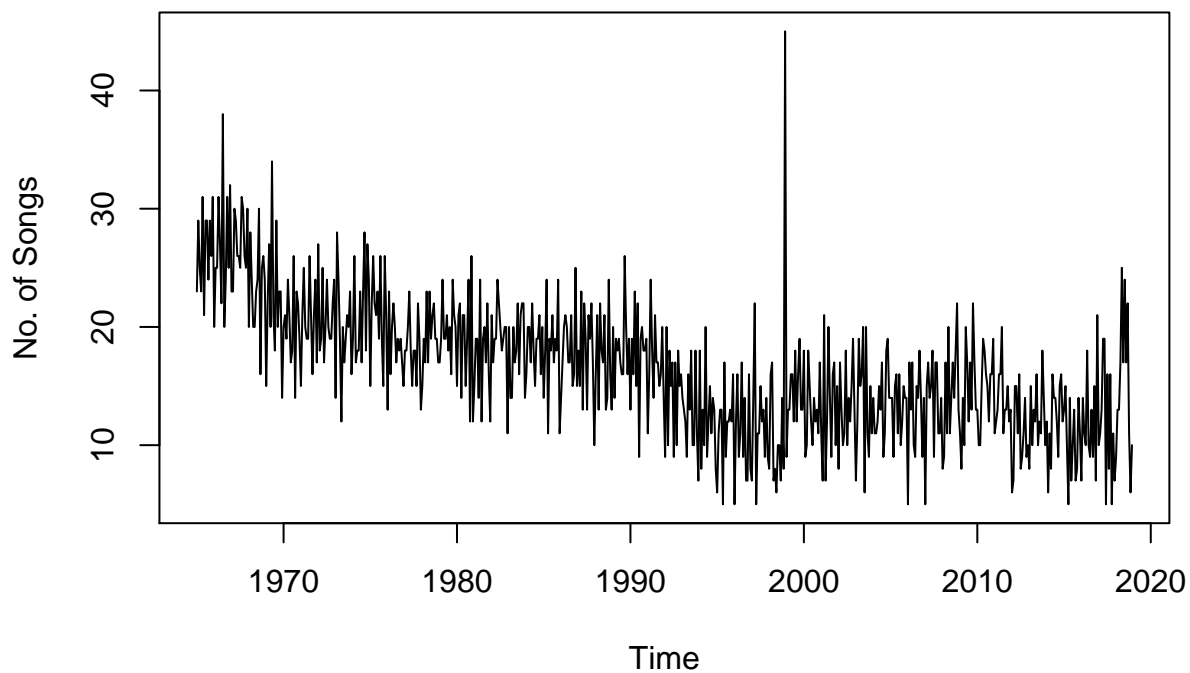
Question 1: Forecasting Music Charts (30 Pts)

Above is a dataset containing the number of unique songs to enter the billboard top 40 charts each month between 1965 and 2018 (every full year the charts have existed). We will use these charts to build forecasting models to estimate how many new songs will chart in the future.

1a. Exploratory Analysis

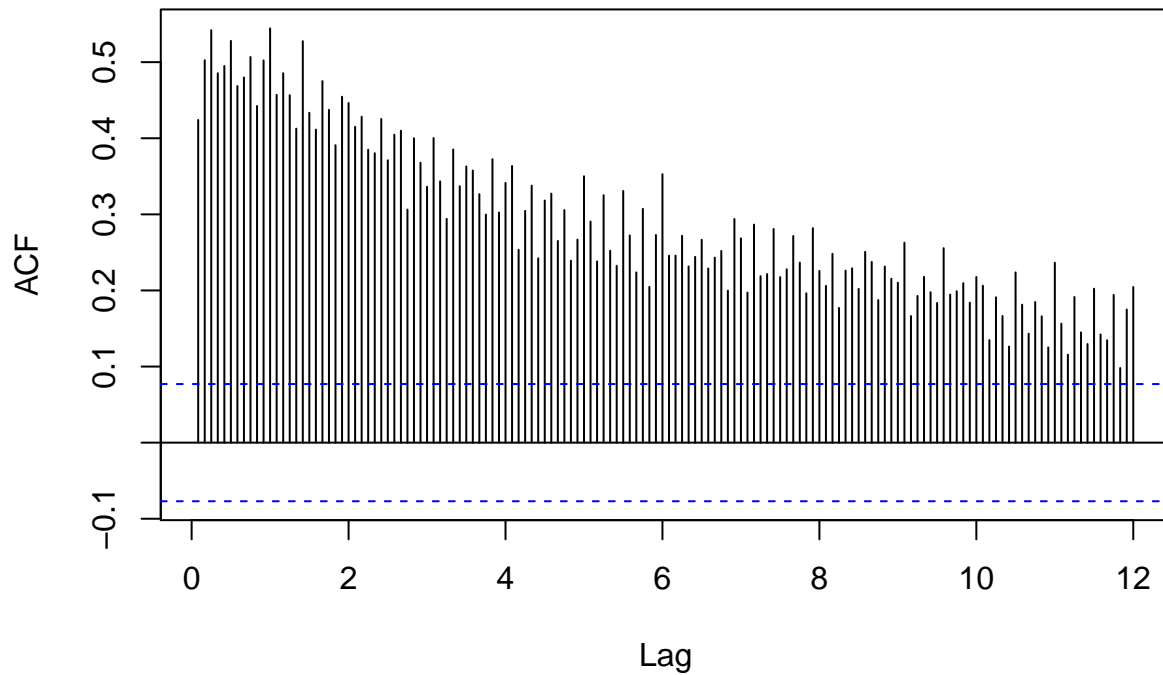
- (i) Plot the Time Series and ACF plots. Comment on the main features, and identify what (if any) assumptions of stationarity are violated.

```
ts.plot(chart, ylab="No. of Songs")
```



```
acf(chart, lag.max = 12 * 12, cex=0.3)
```

Series chart

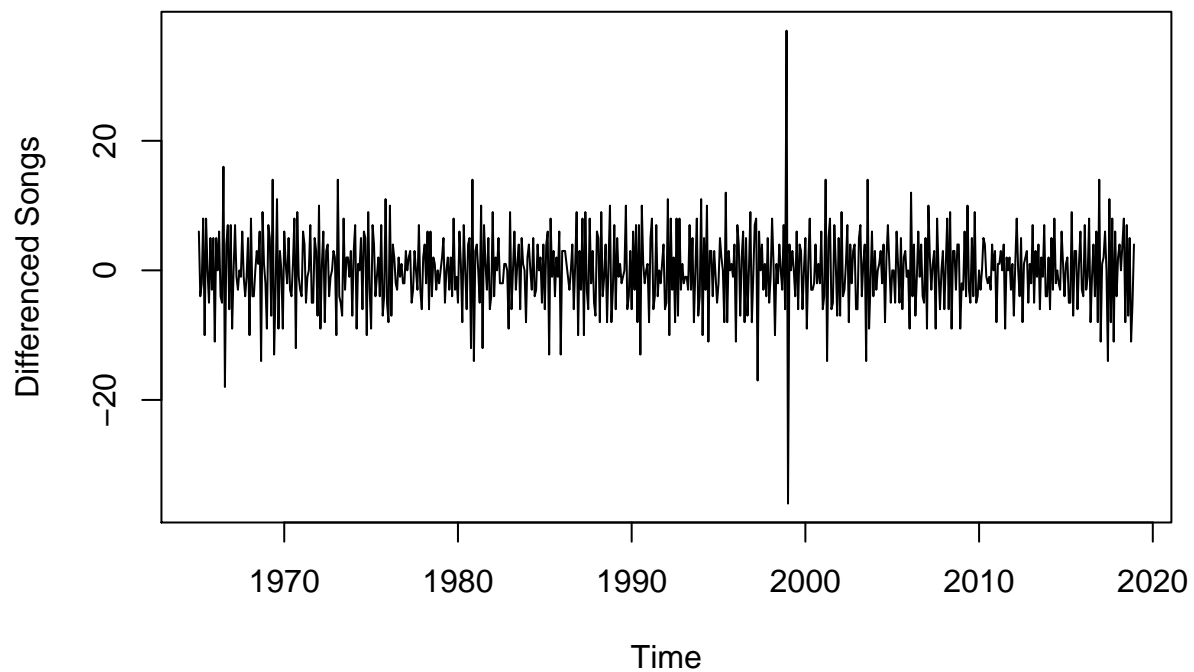


Comment:

From the time series plot and ACF plot, we could clearly see there is a decreasing trend and some seasonality is still remaining in the data. Therefore, it may violate constant mean and auto-covariance may not be independent of time. In the ACF plot, all spikes within lag=12 are all outside significant bands. And in the time series plot, from the beginning year to 2020, decreasing trend is here and there is a big spike around 1999.

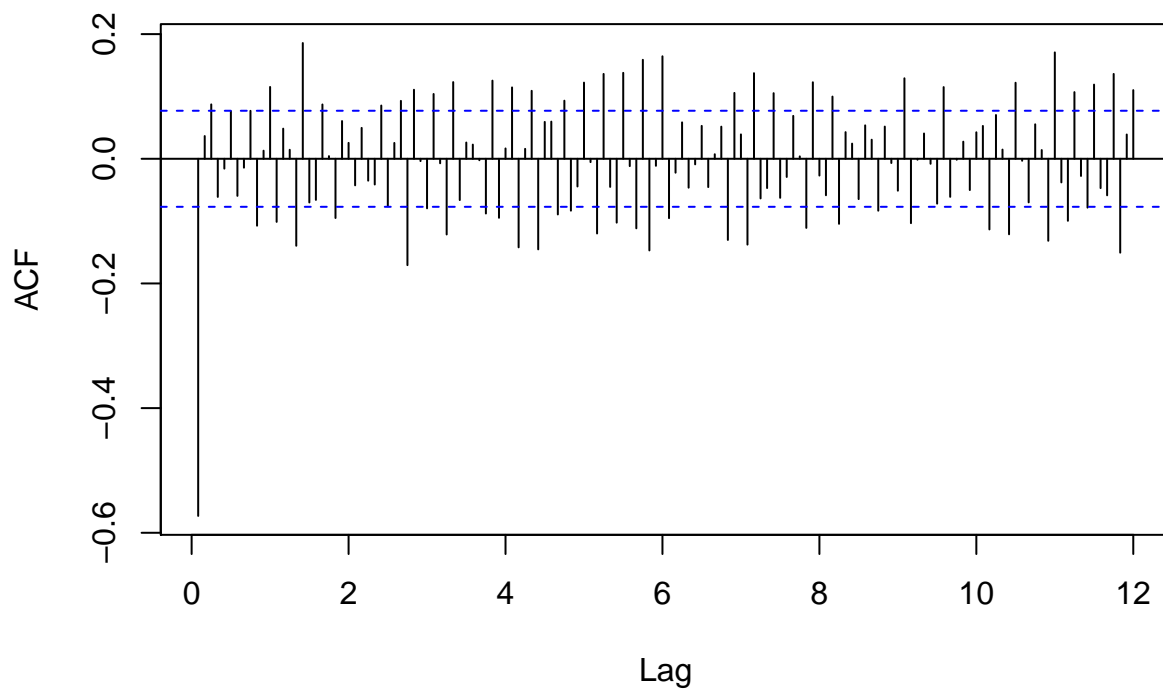
-
- (ii) Perform a differencing on the data, and perform the same analysis. How do assumptions of stationarity hold for the differenced data? Do you expect the differencing data be suitable for ARMA forecasting?

```
ts.plot(chart.dif, ylab="Differenced Songs")
```



```
acf(chart.dif, lag.max = 12 * 12, cex=0.3)
```

Series chart.dif



Comment:

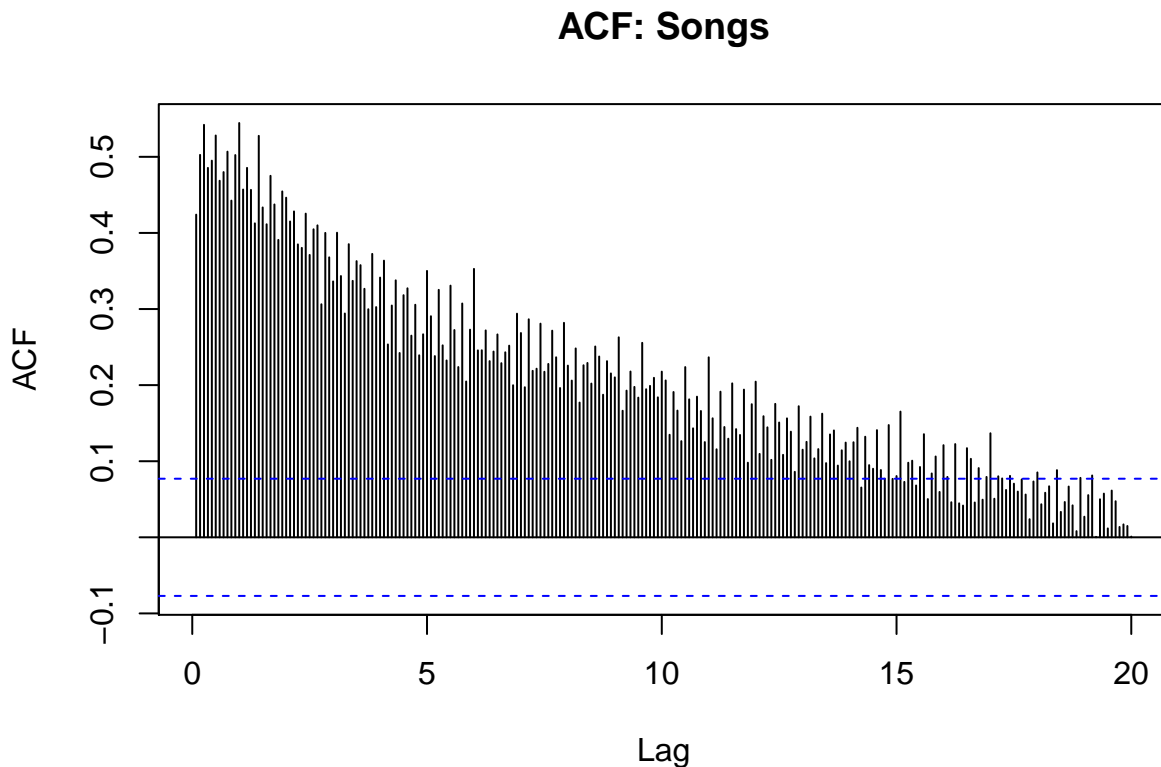
- (1) How do assumptions of stationarity hold for the differenced data? From the time series plot, the assumption of constant mean and constant variance hold, although there is one spike in the plot. But other than that, it shows constant mean and variance.

- (2) Do you expect the differencing data be suitable for ARMA forecasting? No. From the ACF plot, we could see most spikes are outside significant bands. ARMA models requires the stationary assumption. Therefore, the ARMA model on differenced data is not suitable for this dataset.
-

1b. ARIMA Modelling

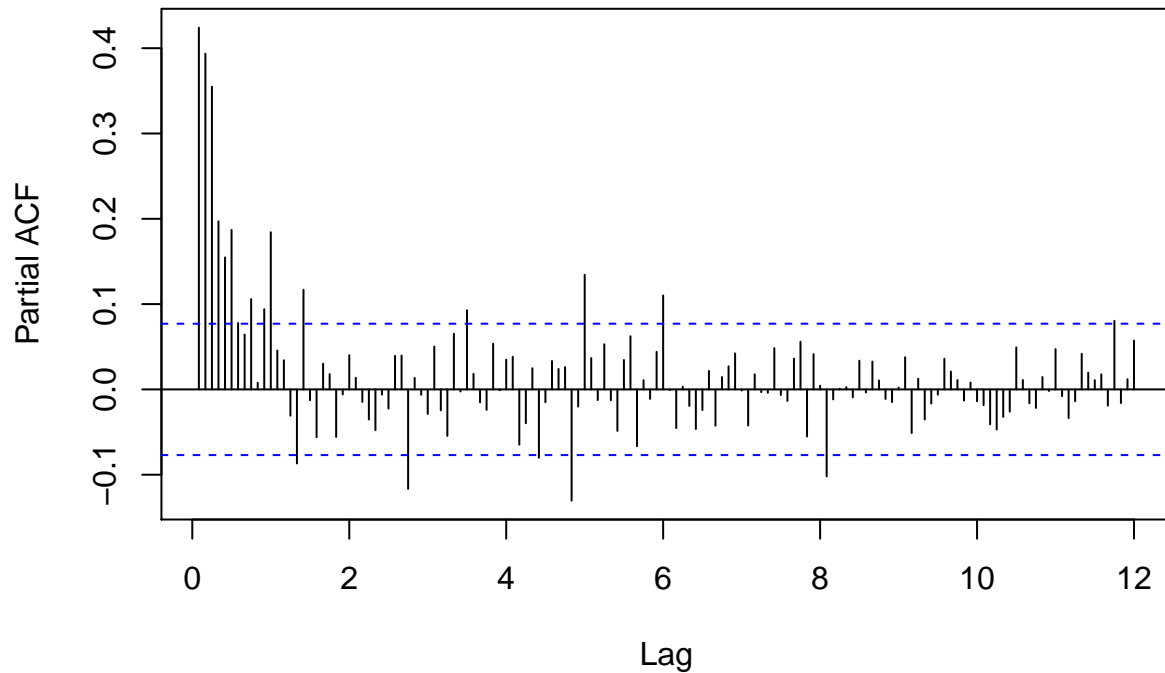
Using graphical analysis of ACF and PACF plots as well as the iterative approach, fit the an ARIMA(p,d,q) model with max order = 3, max differencing = 1. Using an AIC significance threshold of 2, choose which order to use and explain your reasoning for choosing it. Evaluate the model residuals with relevant plots and tests.

```
acf((chart),main="ACF: Songs", lag.max = 12*20)
```



```
pacf((chart),main="PACF: Songs", lag.max = 12*12)
```

PACF: Songs



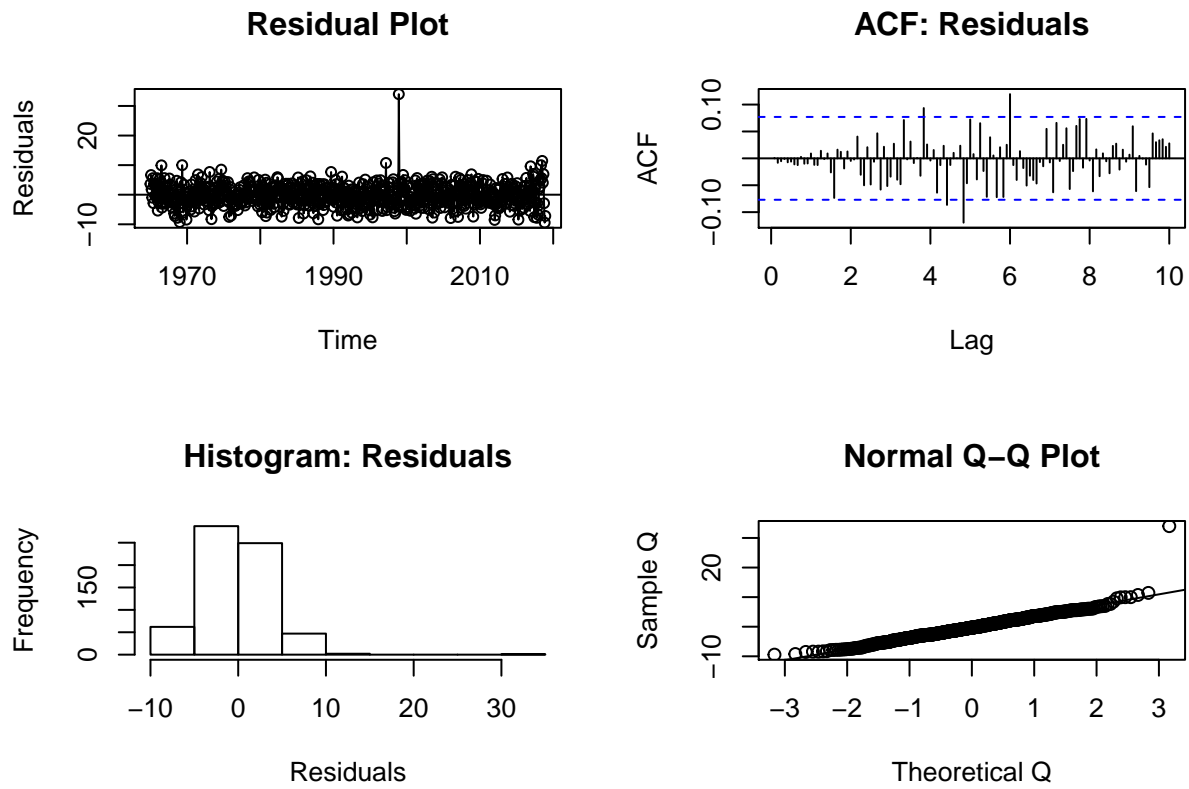
From the ACF and PACF plots, the assumption of constant mean is violated through seeing the ACF plot. In the ACF plot, when $q = 18$, it seems that all spikes are within significant bands. And, the PACF plots seems to give us the p value in AR(p) process. I decide to choose $p = 8$ and $q = 18$ to fit the ARMA(p, q) model.

```
mod.ACFPACF = arima(chart, order=c(8,0,18), method="ML")
AIC(mod.ACFPACF)
```

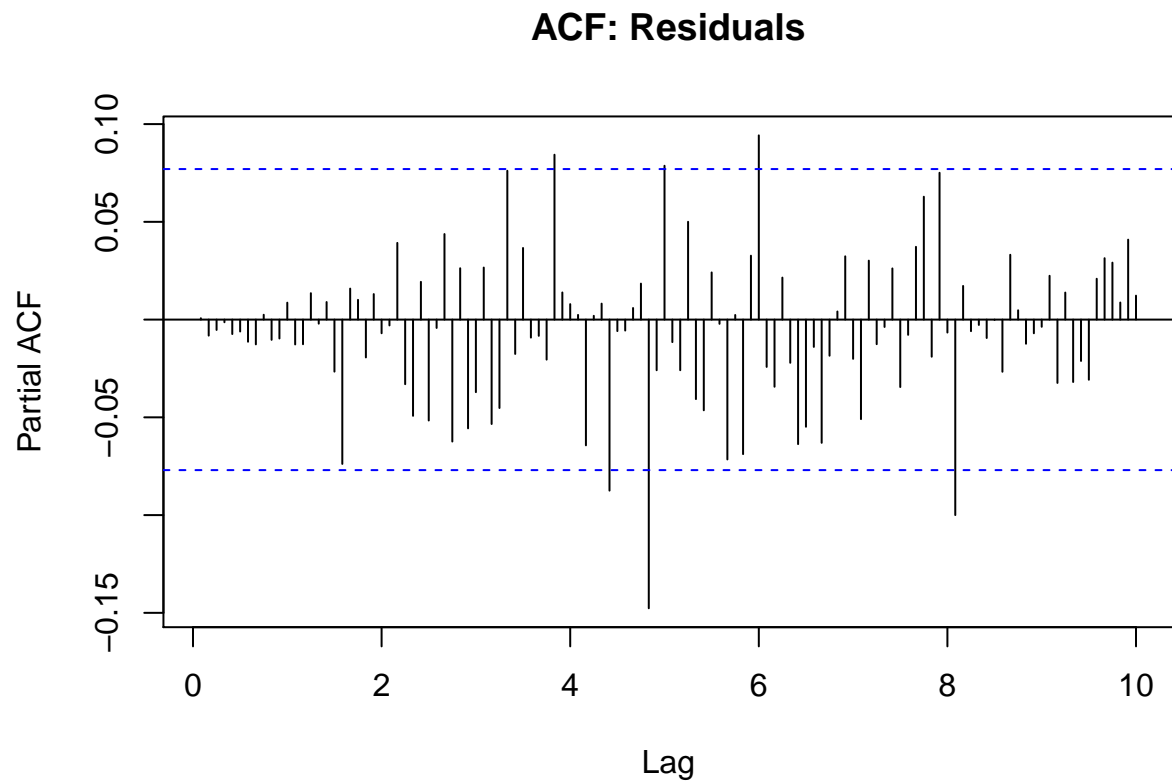
```
## [1] 3640.973
```

The AIC for ARMA(8, 18) is 3640.973.

```
par(mfrow=c(2,2))
plot(resid(mod.ACFPACF), ylab='Residuals',type='o',main="Residual Plot")
abline(h=0)
acf(resid(mod.ACFPACF),main="ACF: Residuals", lag.max = 10 * 12)
hist(resid(mod.ACFPACF),xlab='Residuals',main='Histogram: Residuals')
qqnorm(resid(mod.ACFPACF),ylab="Sample Q",xlab="Theoretical Q")
qqline(resid(mod.ACFPACF))
```



```
pacf(resid(mod.ACFPACF),main="ACF: Residuals", lag.max = 10 * 12)
```



```
Box.test(mod.ACFPACF$resid, lag = (3+1+3), type = "Box-Pierce", fitdf = (3+3))
```

```
##  
## Box-Pierce test  
##  
## data: mod.ACFPACF$resid  
## X-squared = 0.20338, df = 1, p-value = 0.652
```

```
Box.test(mod.ACFPACF$resid, lag = (3+1+3), type = "Ljung-Box", fitdf = (3+3))
```

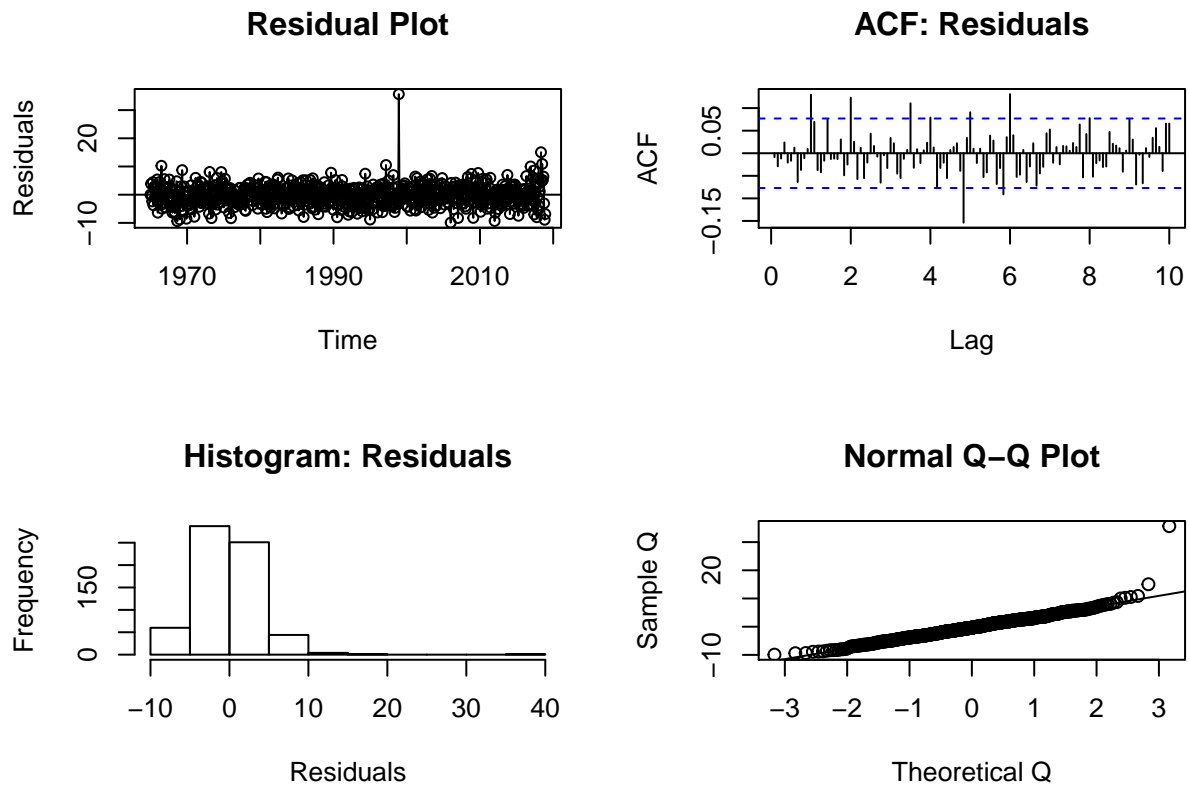
```
##  
## Box-Ljung test  
##  
## data: mod.ACFPACF$resid  
## X-squared = 0.20562, df = 1, p-value = 0.6502
```

Following values are the AIC selection process to determine which order I chose.

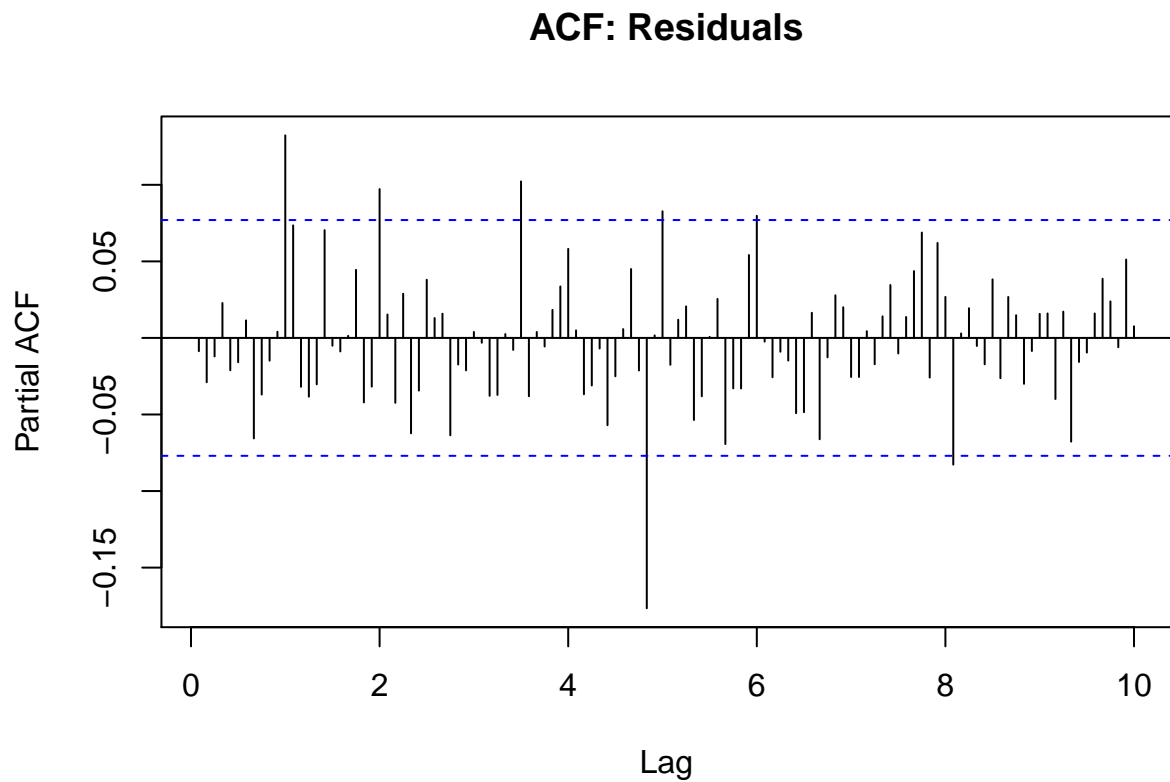
```
orders <- orders[order(-orders$AIC),]  
tail(orders)
```

```
##      p d q      AIC  
## 24 2 1 2 3648.002  
## 31 3 1 1 3647.687  
## 23 2 1 1 3647.237  
## 8  0 1 2 3646.283  
## 29 3 0 3 3636.940  
## 33 3 1 3 3612.688
```

```
par(mfrow=c(2,2))  
plot(resid(final_model), ylab='Residuals',type='o',main="Residual Plot")  
abline(h=0)  
acf(resid(final_model),main="ACF: Residuals", lag.max = 10 * 12)  
hist(resid(final_model),xlab='Residuals',main='Histogram: Residuals')  
qqnorm(resid(final_model),ylab="Sample Q",xlab="Theoretical Q")  
qqline(resid(final_model))
```



```
pacf(resid(final_model), main="ACF: Residuals", lag.max = 10 * 12)
```




```
Box.test(final_model$resid, lag = (3+1+3), type = "Box-Pierce", fitdf = (3+3))
```

```
##  
## Box-Pierce test  
##  
## data: final_model$resid  
## X-squared = 1.6037, df = 1, p-value = 0.2054
```

```
Box.test(final_model$resid, lag = (3+1+3), type = "Ljung-Box", fitdf = (3+3))
```

```
##  
## Box-Ljung test  
##  
## data: final_model$resid  
## X-squared = 1.6181, df = 1, p-value = 0.2034
```

- (1) Choose which order to use and explain your reasoning for choosing it

From the graphical analysis, I used ACF and PACF to determine p and q . Although the plots shows small difference from iterative approach, AIC statistics (≈ 3640) from the ACF and PACF is higher than one (≈ 3612) from the iterative approach. Therefore, in this assignment, I chose less complex model ARIMA(3, 1, 3), instead of ARMA(8, 18).

- (2) Evaluate the model residuals with relevant plots and tests.

From the two plots, there is small difference from two models.

In the ACF plots, ARMA(8, 18) seems to have fewer spikes outside the significant bands than ARIMA(3, 1, 3).

For the residual plots, two models shows the same big spike. The histograms from two models shows nearly the same shape of frequency.

As for Normal Q-Q plots, ARMA(8, 18) have more points lying on the straight line. For the PACF plots on residuals, ARMA(8, 18) shows that after lag=8, most spikes are within the significance band. And, ARIMA(3, 1, 3) also shows that after lag=8, most spikes are within the band.

For independence tests, Box-Pierce test and Box-Ljung test, ARMA(8, 18) have higher p-values (≈ 0.652) than one (≈ 0.20) from ARIMA(3, 1, 3). It means that the residuals from ARMA(8, 18) are possibly correlated because p-values are greater than values from 90% confidence intervals. On the contrary, p-values from ARIMA(3, 1, 3) are around 0.20, less than ARMA(8, 18). However, from the two models, residuals are most likely correlated because of higher p-values.

1c. Forecasting

Build an ARIMA(2,1,4) model. Display model coefficients, comment on anything that stands out about them, and write out the model formula in full form. Then, keep the last 6 data points for testing. Generate forecasts of those 6 months and compare the predicted values to the actual ones. Include 95% confidence interval for the forecasts and provide plots. Calculate Mean Absolute Prediction Error (MAE), Mean Absolute Percentage Error (MAPE), and Precision Measure (PM); comment on the accuracy of predictions.

```
final_model_1c = arima(chart, order = c(2,1,4), method = "ML")
final_model_1c
```

```
##
## Call:
## arima(x = chart, order = c(2, 1, 4), method = "ML")
##
## Coefficients:
##          ar1          ar2          ma1          ma2          ma3          ma4
##      -1.1589  -0.9991   0.0915  -0.0969  -0.8893   0.1602
## s.e.    0.0017   0.0013   0.0397   0.0186   0.0190   0.0387
##
## sigma^2 estimated as 15.06:  log likelihood = -1799.41,  aic = 3610.81
```

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
## Warning: package 'zoo' was built under R version 3.6.2
```

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

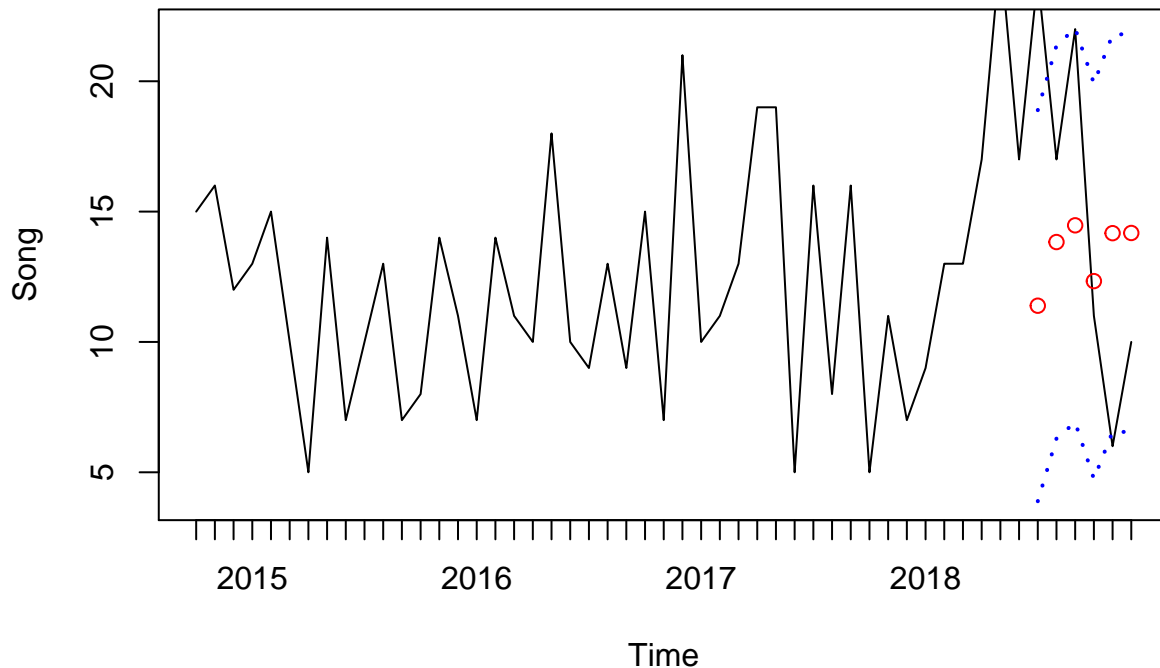
```
## The following objects are masked from 'package:base':
##
##      as.Date, as.Date.numeric
```

```
coeftest(final_model_1c)
```

```
##
## z test of coefficients:
##
##      Estimate Std. Error  z value Pr(>|z|)
## ar1 -1.1589131  0.0016626 -697.0688 < 2.2e-16 ***
## ar2 -0.9991219  0.0012734 -784.6384 < 2.2e-16 ***
## ma1  0.0914596  0.0397442   2.3012  0.02138 *
## ma2 -0.0969066  0.0186010  -5.2097  1.891e-07 ***
## ma3 -0.8892608  0.0190499 -46.6807 < 2.2e-16 ***
## ma4  0.1602295  0.0386646   4.1441  3.412e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- (1) Display model coefficients, comment on anything that stands out about them, and write out the model formula in full form. There are six coefficients. Two of them are for the AR process and the other are for the MA process. The full form of this model is $X_t = -1.15X_{t-1} - 0.99X_{t-2} + 0.09Z_{t-1} - 0.09Z_{t-2} - 0.88Z_{t-3} - 0.16Z_{t-4} + Z_t$, where $Z_t \sim WN(0, \sigma^2 = 15.06)$. From this model, I found we have smaller AIC value compared to previous ones. And, I also performed testing to check coefficients and found all coefficients are significant.
- (2) Keep the last 6 data points for testing. Generate forecasts of those 6 months and compare the predicted values to the actual ones. Include 95% confidence interval for the forecasts and provide plots.

Songs Prediction



From the prediction plot, it seems that prediction a little catch the shapes of true values.

- (3) Calculate Mean Absolute Prediction Error (MAE), Mean Absolute Percentage Error (MAPE), and Precision Measure (PM); comment on the accuracy of predictions.

```
## Compute Accuracy Measures
obssong = chart[(nfit+1):n]
predsong = outpred$pred
### Mean Absolute Prediction Error (MAE)
mean(abs(predsong-obssong))
```

```
## [1] 6.164941
```

```
### Mean Absolute Percentage Error (MAPE)
mean(abs(predsong-obssong)/obssong)
```

```
## [1] 0.4925023
```

```
### Precision Measure (PM)
sum((predsong-obssong)^2)/sum((obssong-mean(obssong))^2)
```

```
## [1] 1.217731
```

MAE places emphasis on this is the mean signed difference. We could generally know that it ceters around the value,6.16. Actually, it is a little higher if we want to make prediction more accurate.

MAPE gives us a model evaluation, because of its very intuitive interpretation in terms of relative error. For the prediction situation, we hope we could get smaller MAPE because from the definition if two values are close, the percentage would be smaller. Therefore, we could only get around 0.5 accruacy.

PM is also a similar indicator which use the sum of variance of observed values to show ratio between prediction and true values. If prediction is close to observed values, the value of PM would be close to 0. But, this value is not only close to 0 but also outside of 1. So it means most predictions are greater than one from the observed values to the mean. We know that the variance captures the spread of data. Therefore, if the values of PM is within 1, it shows that prediction is close to true values.

Question 2: ARIMA Model Comparisons for Currency Conversion Rates (25 Pts)

Using the USD to EURO conversion rate data that we already explored in Question 2 of Homework 1, we will attempt to build a predictive model to forecast what rates will be like in the future. Additionally, we build a similar model for USD to GBP conversion rate data for the same time period, though we'll be a bit reckless and skip proper exploratory data analysis this time.

2a. ARIMA Fitting

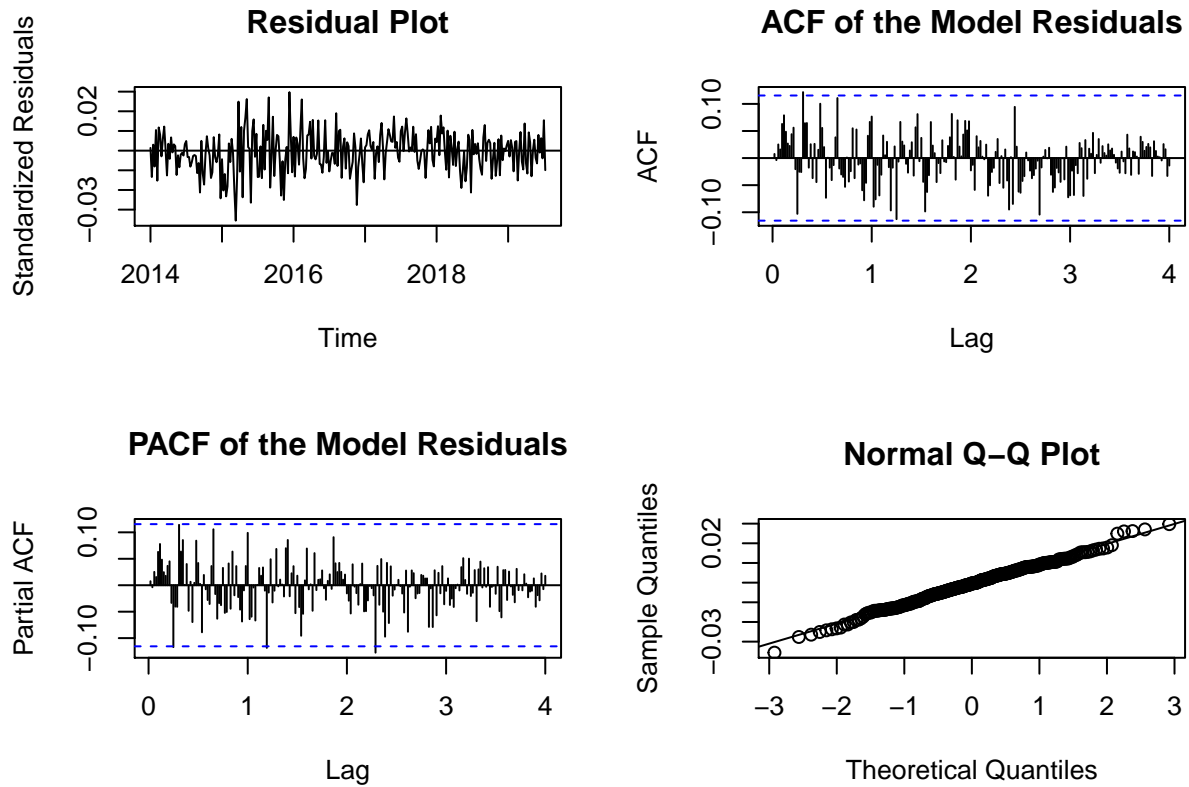
For both time series, use the iterative model to fit an ARIMA(p,d,q) model with max order = 3, max differencing = 2. Evaluate the models for ACF and PACF plots as well as relevant tests.

- EU

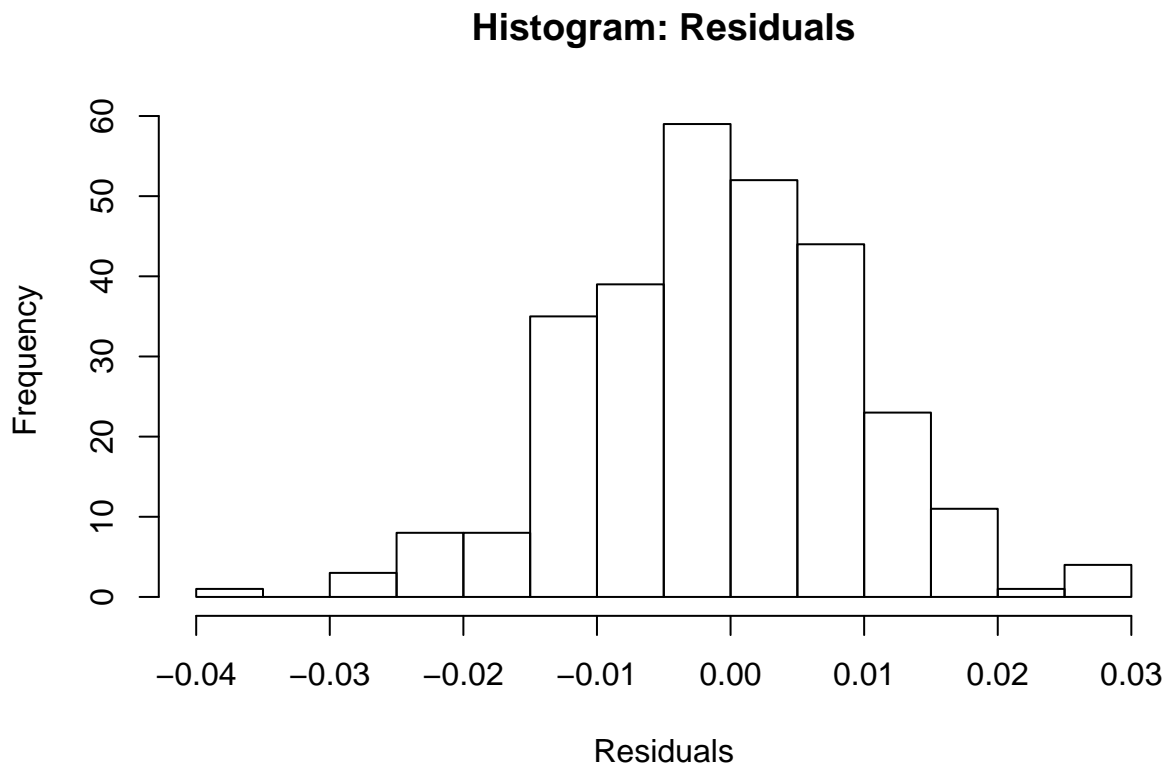
```
orders <- orders[order(-orders$AIC),]  
tail(orders)
```

```
##      p d q      AIC  
## 17 1 1 0 -1788.229  
## 7  0 1 1 -1788.708  
## 42 3 1 2 -1796.099  
## 31 2 1 3 -1796.100  
## 43 3 1 3 -1796.761  
## 30 2 1 2 -1798.008
```

```
final_EU = arima(EU, order = c(2,1,2), method = "ML")  
resids.EU = resid(final_EU)  
## Residual Analysis  
par (mfrow=c(2,2))  
plot(resids.EU, ylab='Standardized Residuals', main="Residual Plot")  
abline(h=0)  
acf(resids.EU, main= 'ACF of the Model Residuals', lag.max = 52*4)  
pacf(resids.EU, main='PACF of the Model Residuals', lag.max = 52*4)  
qqnorm(resids.EU)  
qqline(resids.EU)
```



```
hist(resid(final_EU),xlab='Residuals',main='Histogram: Residuals')
```



```
## Test for Independence for final model
Box.test(resids.EU, lag = (2+2+1), type = "Box-Pierce", fitdf = (2+2))
```

```
##
## Box-Pierce test
##
## data:  resids.EU
## X-squared = 1.4178, df = 1, p-value = 0.2338
```

```
Box.test(resids.EU, lag = (2+2+1), type = "Ljung-Box", fitdf = (2+2))
```

```
##
## Box-Ljung test
##
## data:  resids.EU
## X-squared = 1.4511, df = 1, p-value = 0.2284
```

(1) Evaluate the models for ACF and PACF plots as well as relevant tests.

For AIC selection, ARIMA(2, 1, 2) has smaller AIC compared to the other. But, we should also make sure AIC threshold (< 2). So other candidates like ARIMA(3, 1, 3), ARIMA(2, 1, 3) and ARIMA(3, 1, 2) are also options. However, less parameters means simpler models. Therefore, the numbers of parameters for ARIMA(2, 1, 2) is smaller than others. Thus, ARIMA(2, 1, 2) is our final model.

From the ACF plot, most spikes are within significant bands which means the assumptions of stationary process are not violated.

From the residual plot, we could also see constant variance.

For the PACF plot, in theory partial autocorrelations are equal to 0 beyond that point. So in this plot, it follows the theory and if we take expectation on the PACF plot, the value would be close to 0.

As for the Normal Q-Q plot, we could clearly see that it follows the assumption of normality which means points are almost on the straight line.

For residual independent tests, we could see that the p-values are around 0.23. So they indicate residuals are possibly correlated.

- GBP

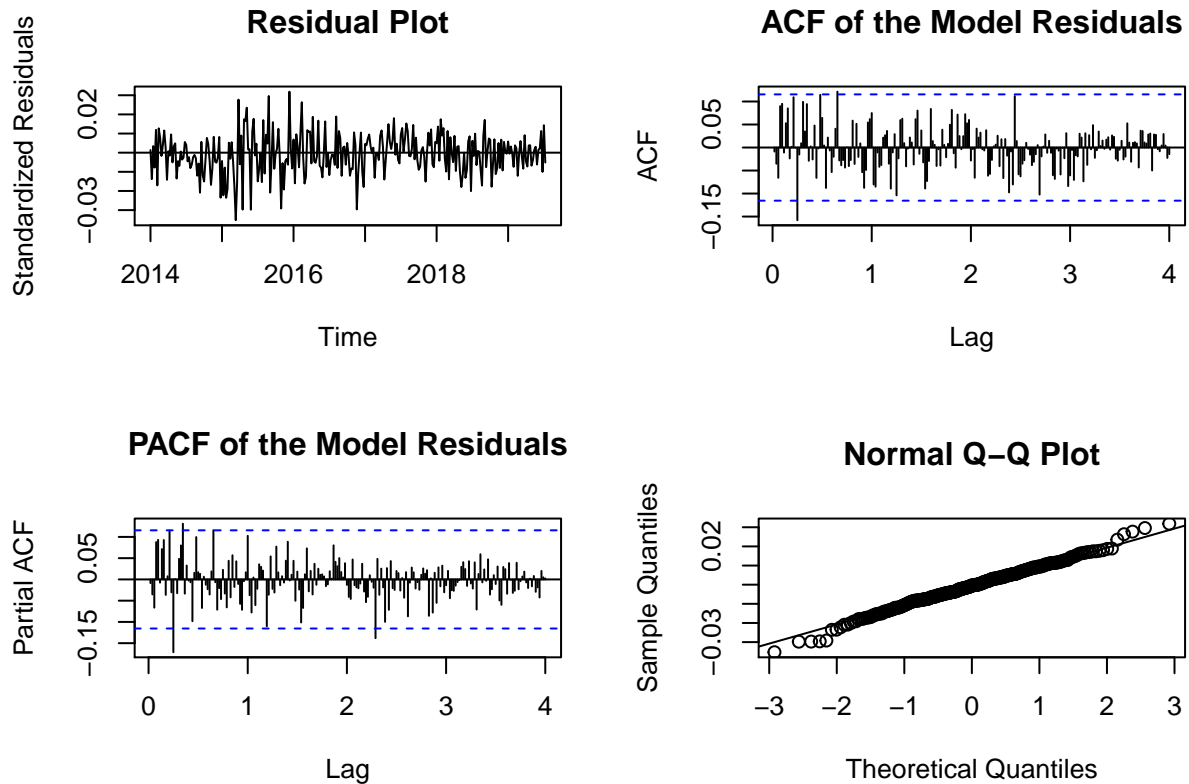
```
orders <- orders[order(-orders$AIC),]
tail(orders)
```

```
##      p d q      AIC
## 9   0 1 3 -1630.533
## 19  1 1 1 -1630.534
## 28  2 0 2 -1630.642
## 8   0 1 2 -1631.018
## 7   0 1 1 -1631.801
## 30  2 1 0 -1632.058
```

```

final_GBP = arima(EU, order = c(0,1,1), method = "ML")
resids.GBP = resid(final_GBP)
## Residual Analysis
par (mfrow=c(2,2))
plot(resids.GBP, ylab='Standardized Residuals', main="Residual Plot")
abline(h=0)
acf(resids.GBP,main= 'ACF of the Model Residuals', lag.max = 52*4)
pacf(resids.GBP,main='PACF of the Model Residuals', lag.max = 52*4)
qqnorm(resids.GBP)
qqline(resids.GBP)

```

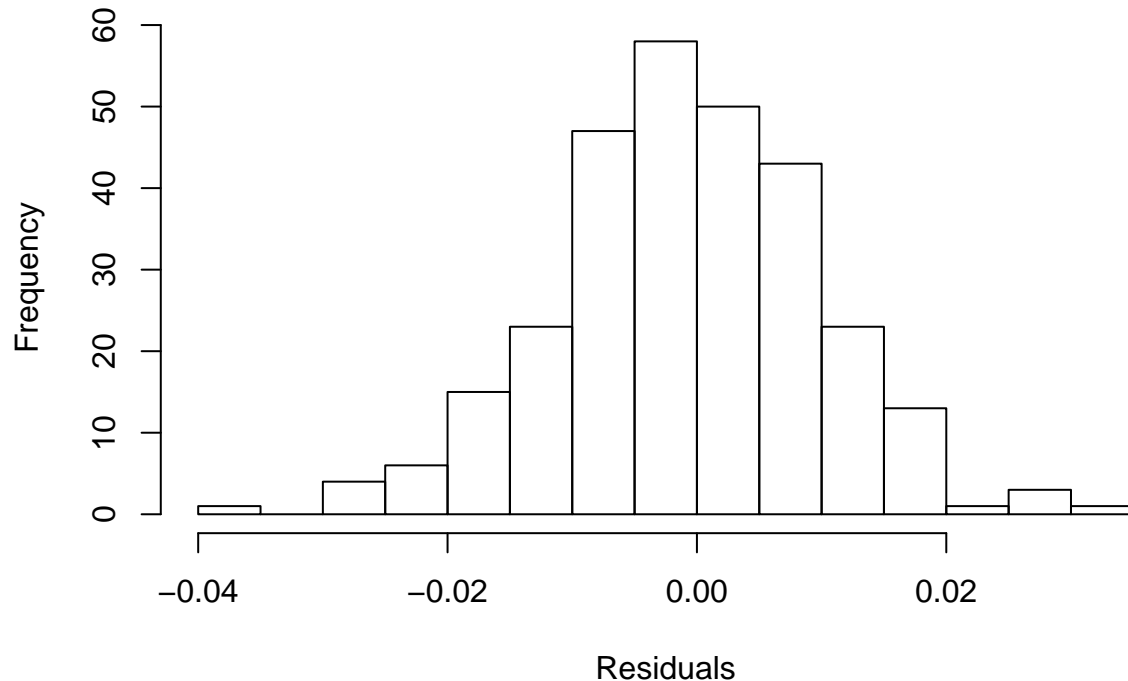


```

hist(resid(final_GBP), xlab='Residuals', main='Histogram: Residuals')

```

Histogram: Residuals



```
## Test for Independence for final model
```

```
Box.test(resids.EU, lag = (0+1+1), type = "Box-Pierce", fitdf = (0+1))
```

```
##  
## Box-Pierce test  
##  
## data:  resids.EU  
## X-squared = 0.020333, df = 1, p-value = 0.8866
```

```
Box.test(resids.EU, lag = (0+1+1), type = "Ljung-Box", fitdf = (0+1))
```

```
##  
## Box-Ljung test  
##  
## data:  resids.EU  
## X-squared = 0.020559, df = 1, p-value = 0.886
```

For AIC selection, ARIMA(0, 1, 1) has little larger AIC compared than ARIMA(2, 1, 0). The AIC of ARIMA(2, 1, 0) is -1632.058. However, we set the AIC threshold as 2. So we need to consider other values as well. Other candidates are ARIMA(0, 1, 2), ARIMA(2, 0, 2), ARIMA(1, 1, 1) and ARIMA(0, 1, 3). However, less parameters means simpler models. Therefore, the numbers of parameters for ARIMA(0, 1, 1) is smaller than others. Thus, ARIMA(0, 1, 1) is our final model.

From the residual plot, we could also see constant variance.

From the ACF plot, most spikes are within significant bands which means the assumptions of stationary process are not violated. But there is one spike which is outside the significance band.

For the PACF plot, some of spikes are outside the significance bands.

From the residual plot, we could also see constant variance.

As for the Normal Q-Q plot, we could clearly see that it follows the assumption of normality which means points are almost on the straight line.

For histograms of residual, they do not totally follow normal shapes.

For residual independent tests, we could see that the p-values are around 0.886. So they strongly indicates residuals are possibly correlated.

2b. Forecasting

Show coefficients for both models and compare significance of coefficients. Next, for each series keep the last 12 data points for testing. Generate forecasts of those 12 weeks and compare the predicted values to the actual ones. Include 95% confidence interval for the forecasts and provide plots. Calculate Mean Absolute Percentage Error (MAPE) and Precision Measure (PM) for each; compare the models.

- (1) Show coefficients for both models and compare significance of coefficients.

```
library(lmtest)
coeftest(final_EU)
```

```
##
## z test of coefficients:
##
##      Estimate Std. Error  z value  Pr(>|z|)
## ar1 -0.450384   0.079828  -5.6419 1.681e-08 ***
## ar2 -0.849137   0.058169 -14.5979 < 2.2e-16 ***
## ma1  0.595798   0.061439   9.6973 < 2.2e-16 ***
## ma2  0.928562   0.057338  16.1945 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

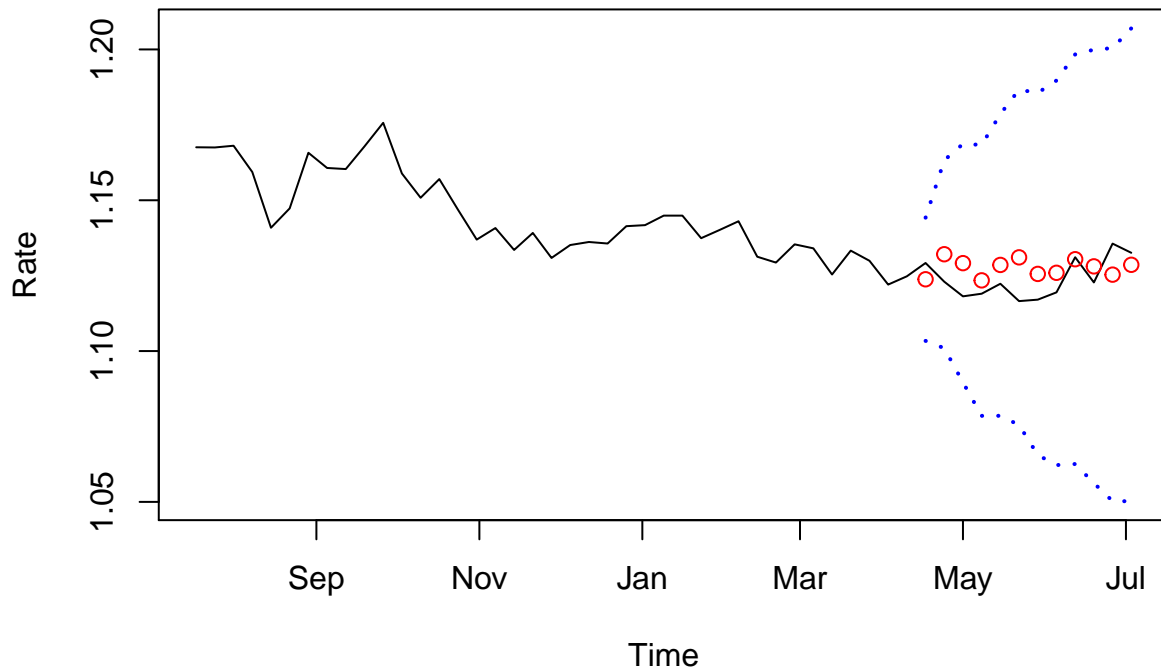
```
library(lmtest)
coeftest(final_GBP)
```

```
##
## z test of coefficients:
##
##      Estimate Std. Error z value Pr(>|z|)
## ma1  0.14683   0.05903   2.4874 0.01287 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

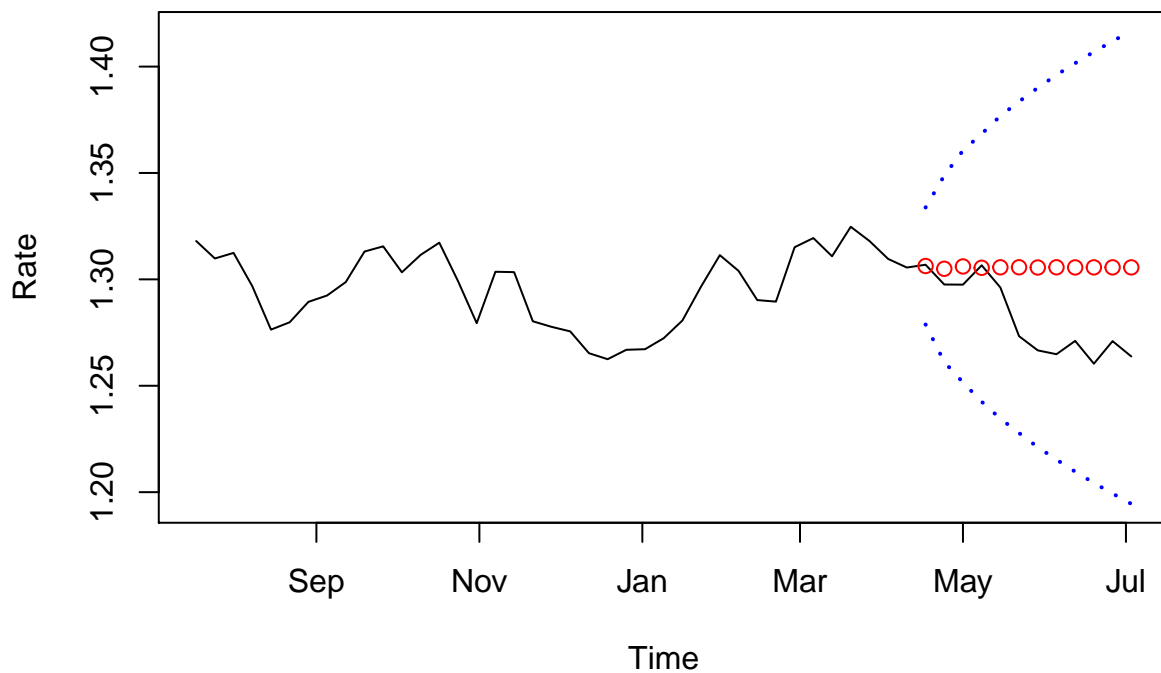
For EU arima model, we could see all coefficient are significant. However, GBP arima model is 90% significant compared to EU model which is more significant.

- (2) For each series keep the last 12 data points for testing. Generate forecasts of those 12 weeks and compare the predicted values to the actual ones. Include 95% confidence interval for the forecasts and provide plots.

EU Forecasting



GBP Forecasting



From the prediction plots, the EU prediction captures better shapes of observed values than GBP.

From GBP prediction, it is almost a straight line which is not ideal.

- (3) Calculate Mean Absolute Percentage Error (MAPE) and Precision Measure (PM) for each; compare the models.

```
## Compute Accuracy Measures
obsEU = EU[(nfit+1):n]
predEU = outpred.EU$pred
### Mean Absolute Percentage Error (MAPE)
mean(abs(predEU-obsEU)/obsEU)
```

```
## [1] 0.006380015
```

```
### Precision Measure (PM)
sum((predEU-obsEU)^2)/sum((obsEU-mean(obsEU))^2)
```

```
## [1] 1.616309
```

```
## Compute Accuracy Measures
obsGBP = GBP[(nfit+1):n]
predGBP = outpred.GBP$pred
### Mean Absolute Percentage Error (MAPE)
mean(abs(predGBP-obsGBP)/obsGBP)
```

```
## [1] 0.01940747
```

```
### Precision Measure (PM)
sum((predGBP-obsGBP)^2)/sum((obsGBP-mean(obsGBP))^2)
```

```
## [1] 2.999087
```

From two MAPEs and PMs, we could see what EU model does is better than the GBP model.

For the prediction of MAPE, we hope we could get smaller MAPE because from the definition if two values are close, the percentage would be smaller. Therefore, the EU model has the value close to 0 but the MAPE of the GBP model is a little higher.

In Precision Measure, if prediction is close to observed values, the value of PM would be within 1. However, both values from these two models are outside 1. But, the EU model is better than the GBP model.

Question 3: Reflection on ARIMA (5 Pts)

Considering your understanding of the model as well as what you experiences completing the above questions, how would you personally regard the effectiveness of ARIMA modelling? Where would it be appropriate to use for forecasting and where would you recommend against? What are some specific points of caution one would need to consider when considering using it?

Answer:

- (1) How would you personally regard the effectiveness of ARIMA modelling?

In my opinion, I think ARIMA is more effective when we need to consider more orders of difference of raw observations to make this time series plot stationary. If we just use MA, AR or ARMA models, sometimes we could not perfectly get stationary data. Therefore, ARIMA beats other models.

(2) Where would it be appropriate to use for forecasting and where would you recommend against?

When it is appropriate,

- Short term forecasting: From the above forecasting, we could see only short period works well.
- Current data points rely on past data points. It is because the factor for changing forecasting is simple. And, ARIMA could perform well.
- Data is stationary: the assumptions for ARIMA is stationary.

When I would not recommend,

- Scarce Data: If the d is higher, it means we need more time lags. If we only have small amounts of data. It is hard to use ARIMA.
- Unstable:
- Long-time observation: ARIMA only perform well in short periods.

(3) What are some specific points of caution one would need to consider when considering using it?

- If the series has a strong and consistent seasonal pattern, then you should use an order of seasonal differencing.
- It can be stationarized by a combination of differencing and other mathematical transformations such as logging
- We have a substantial amount of data to work with: at least 4 full seasons in the case of seasonal data.