Time Series Analysis for Passenger Miles Flown in U.K.

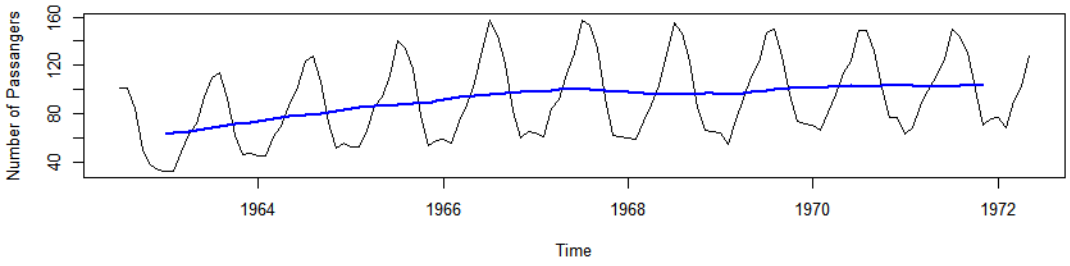
Yunfei Luo

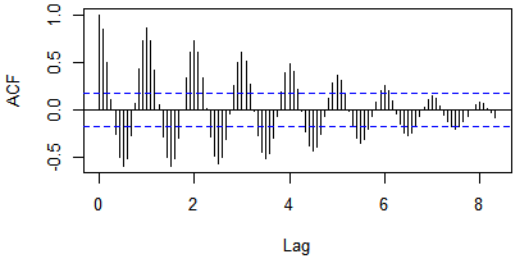
April 24th, 2018

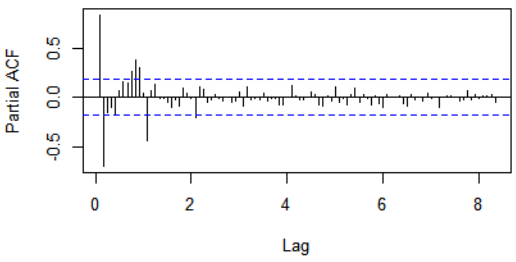
1 Introduction

The passenger miles flown in U.K. from July 1962 to May 1972 are shown in the following plot. We observe the seasonality with the frequency of 12 months. Within a year, the number of tourists tends to reach the peak in July and the trough in February. We use moving average smoothing to reveal the trend. The trend is slightly upward in the beginning and then turns to be horizontal from the middle to the end. The ACF and PACF plots demonstrate that this time series is not stationary.

```
passenger <- read.csv(file = "passenger.csv")
passenger = ts(passenger, start = c(1962,7), end = c(1972,5), frequency = 12)
plot(passenger, ylab = "Number of Passangers")
pass.filter = filter(passenger, sides=2, filter=c(.5, rep(1,11), .5)/12)
lines(pass.filter, lwd=2, col=4)
acf(passenger, lag.max = 100, main = "")
pacf(passenger, lag.max = 100, main = "")
```







2 ARMA Model

From the plots above, we can roughly determine that there exists a slight trend with the seasonal component. There are two ways to simulate the seasonality, by the trigonometric functions or by the monthly dummy variables. We will conduct both estimations to determine which one is better.

```
# define variables
t1 = time(passenger)
t2 = t1^2
z1 = sin(2*pi*t1)
z2 = cos(2*pi*t1)

# Method 1: quadratic for the trend and trigonometric for seasonality
modelA = lm(passenger ~ t1 + t2 + z1 + z2)
summary(modelA)

Call:
lm(formula = passenger ~ t1 + t2 + z1 + z2)

Residuals:
    Min      1Q  Median      3Q     Max
-17.993  -7.726   1.034   6.412  24.622

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.480e+06  4.561e+05  -5.437 3.13e-07 ***
t1           2.517e+03  4.637e+02   5.428 3.26e-07 ***
t2          -6.385e-01  1.178e-01  -5.419 3.39e-07 ***
z1          -1.974e+00  1.221e+00  -1.616    0.109
z2          -4.098e+01  1.225e+00 -33.458  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.419 on 114 degrees of freedom
Multiple R-squared:  0.9203,Adjusted R-squared:  0.9175
F-statistic: 329.2 on 4 and 114 DF,  p-value: < 2.2e-16
```

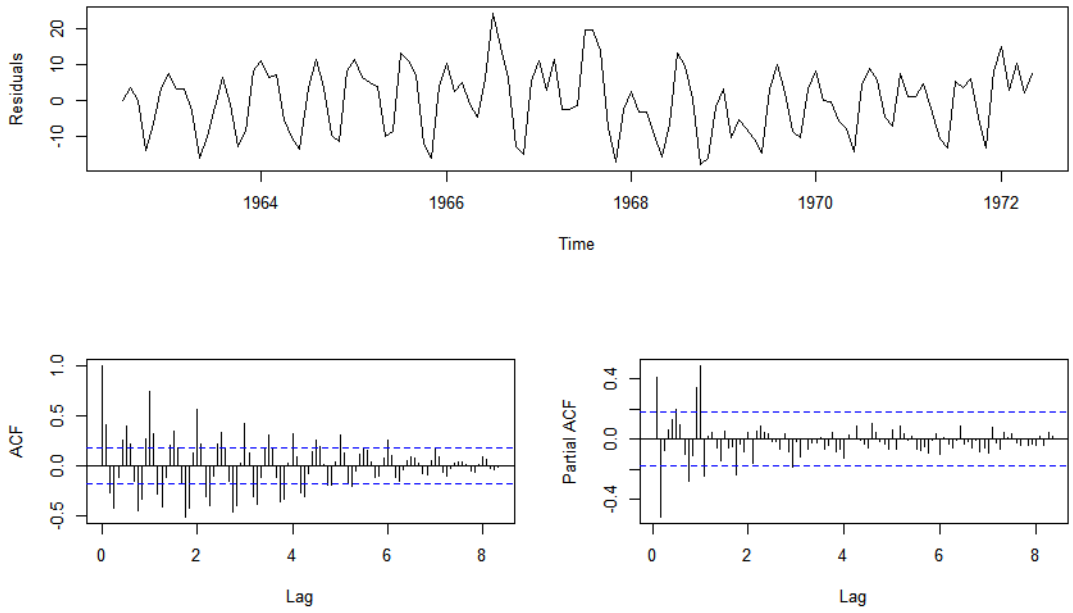
Given the R output, we can write our first estimated model.

$$y_t = -2480000 + 2517t - 0.6385t^2 + s_t + x_t \tag{2.1}$$

$$s_t = -1.974 \sin(2\pi t) - 40.98 \cos(2\pi t) \tag{2.2}$$

We check if x_t still has the seasonality. According to the following ACF and PACF plots of x_t, we can clearly see that the seasonality is not completely removed by the trigonometric functions. We then turn to simulate the seasonality by monthly dummy variables in hope of better results.

```
# residuals of Method 1
resA = passenger - fitted(modelA)
plot(resA, ylab = "Residuals")
acf(resA, lag.max = 100, main = "")
pacf(resA, lag.max = 100, main = "")
```




```
# define monthly dummy variables
M = factor(rep(1:12, 11))
levels(M) = month.abb
M = M[-c(1:6, 126:132)]


# Method 2: quadratic for the trend and dummy variables for seasonality
modelB = lm(passenger ~ t1 + t2 + M)
summary(modelB)



Call:
lm(formula = passenger ~ t1 + t2 + M)

Residuals:
     Min       1Q   Median       3Q      Max
-11.8614  -4.1433   0.0219   3.1384  15.8793
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.559e+06  2.672e+05  -9.576 5.49e-16 ***
t1           2.597e+03  2.716e+02   9.560 5.95e-16 ***
t2          -6.589e-01  6.903e-02  -9.544 6.46e-16 ***
MFeb        -2.546e+00  2.464e+00  -1.033    0.304
MMar         1.441e+01  2.464e+00   5.848 5.69e-08 ***
MApr         2.772e+01  2.464e+00  11.250  < 2e-16 ***
MMay         4.345e+01  2.464e+00  17.632  < 2e-16 ***
MJun         5.889e+01  2.534e+00  23.242  < 2e-16 ***
MJul         8.238e+01  2.465e+00  33.414  < 2e-16 ***
MAug         7.928e+01  2.465e+00  32.163  < 2e-16 ***
MSep         5.947e+01  2.464e+00  24.134  < 2e-16 ***
MOct         2.430e+01  2.464e+00   9.862  < 2e-16 ***
MNov         1.839e+00  2.464e+00   0.746    0.457
MDec         2.725e+00  2.464e+00   1.106    0.271
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.509 on 105 degrees of freedom
Multiple R-squared:  0.9749,Adjusted R-squared:  0.9718
F-statistic: 313.7 on 13 and 105 DF,  p-value: < 2.2e-16
```

Given the R output, we can write our second estimated model.

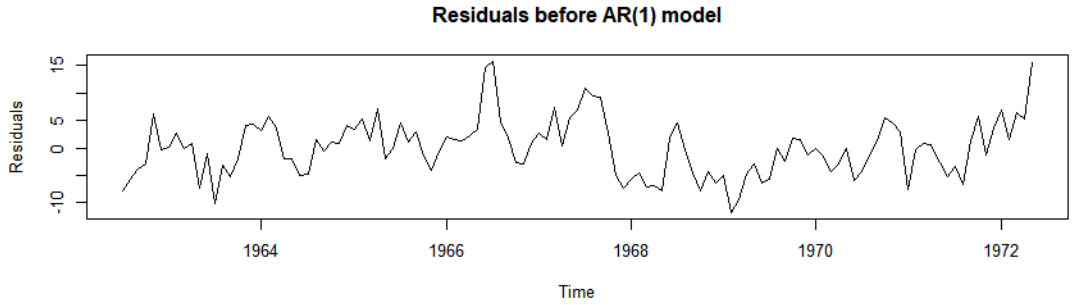
$$y_t = -2559000 + 2597t - 65.89t^2 + s_t + x_t \tag{2.3}$$

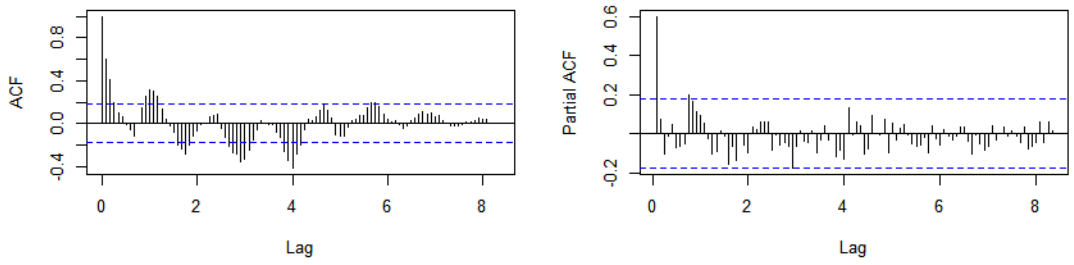
$$\begin{aligned} s_t = &-2.546\text{Feb} + 14.41\text{Mar} + 27.72\text{Apr} + 43.45\text{May} + 58.89\text{Jun} \\ &+ 82.38\text{Jul} + 79.28\text{Aug} + 59.47\text{Sep} + 24.3\text{Oct} + 1.839\text{Nov} + 2.725\text{Dec} \end{aligned} \tag{2.4}$$

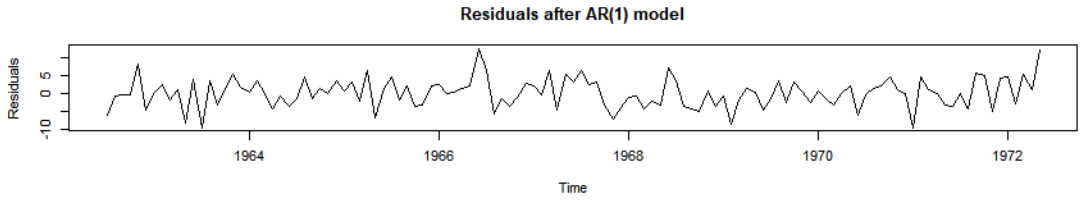
We check if x_t still has the seasonality. According to the following plots of x_t, we can see a significant improvement since the residuals distribute more randomly with equal variance, even though the seasonality is still not completely removed. Due to the cut-off pattern after lag 1 in the PACF plot and the tails-off pattern in the ACF plot, we can try to simulate x_t by AR(1) model. The Q-Q plot demonstrates the normality of residuals and the Ljung-Box test indicates that the residuals are white noise at the 5% significance level.

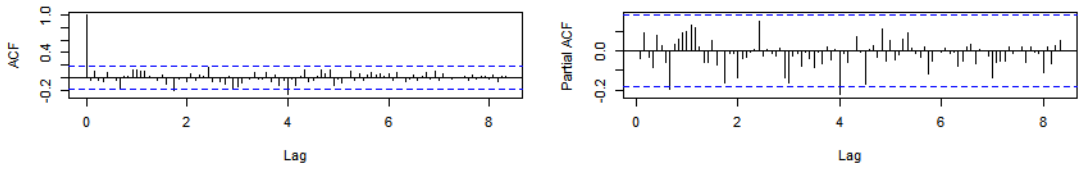
```
# residuals of Method 2
# before AR(1) model
resB1 = passenger - fitted(modelB)
plot(resB1, ylab = "Residuals", main = "Residuals before AR(1) model")
acf(resB1, lag.max = 100, main = "")
pacf(resB1, lag.max = 100, main = "")
# after AR(1) model
pass.ar1 = arima(resB1, order = c(1,0,0))
```

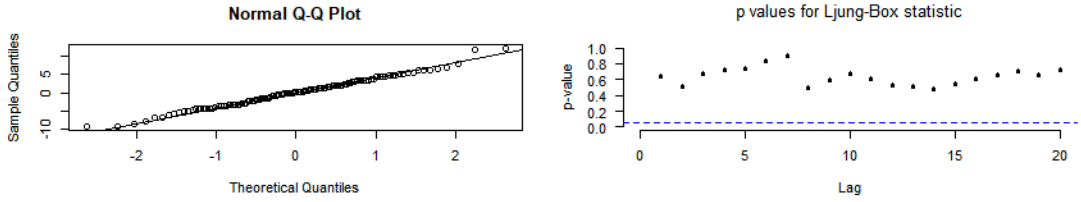
```
resB2 = pass.ar1$residuals
plot(resB2, ylab = "Residuals", main = "Residuals after AR(1) model")
acf(resB2, lag.max = 100, main = "")
pacf(resB2, lag.max = 100, main = "")
qqnorm(resB2)
qqline(resB2)
Box.Ljung.Test(resB2, lag = 20)
```











```
Call:
arima(x = resB1, order = c(1, 0, 0))

Coefficients:
         ar1  intercept
      0.6565     0.1208
s.e.  0.0738     1.0554

sigma^2 estimated as 16.12:  log likelihood = -334.55,  aic = 675.09
```

Given the R output, we write x_t as follows.

$$x_t = 0.1208 + 0.6565x_{t-1} + w_t, \qquad w_t \sim \mathcal{N}(0, 16.12) \tag{2.5}$$

We also compare the result with other possible ARMA models to explain why AR(1) fits the best.

ARMA Model Comparison

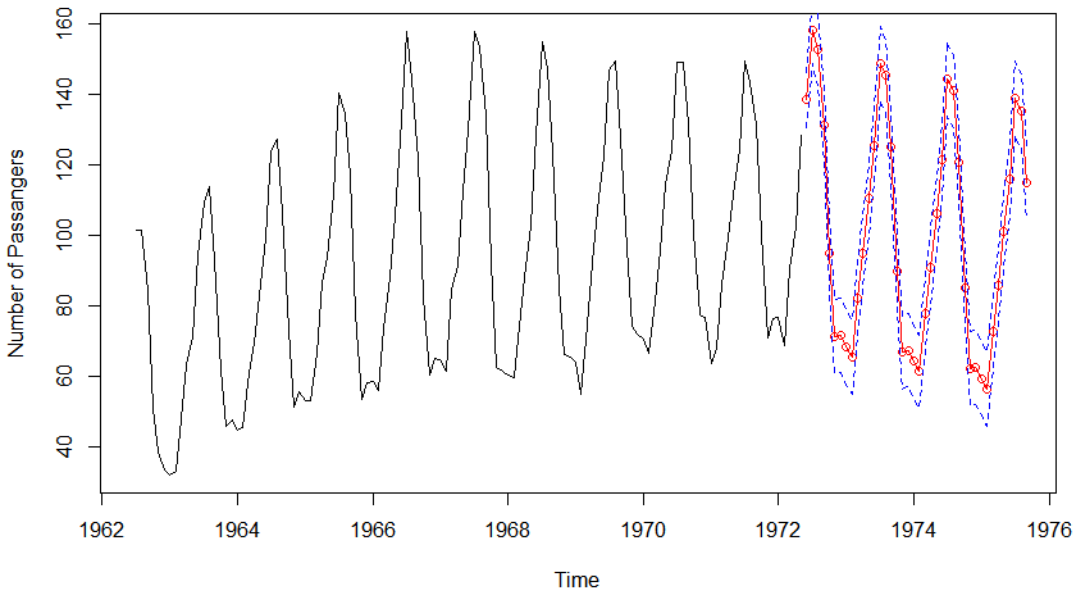
Models	$\hat{\phi}_1$	**SE**$(\hat{\phi}_1)$	$\hat{\phi}_2$	**SE**$(\hat{\phi}_2)$	$\hat{\theta}_1$	**SE**$(\hat{\theta}_1)$	AIC
AR(1)	0.6565	0.0738	-	-	-	-	675.09
AR(2)	0.6188	0.0949	0.0602	0.0957	-	-	676.70
MA(1)	-	-	-	-	0.4912	0.0688	696.37
ARMA(1, 1)	0.6969	0.0994	-	-	-0.0681	0.1250	676.80

- AR(1): It has the lowest AIC and the 95% confidence interval of $\hat{\phi}_1$ includes neither 0 nor 1, which means this coefficient is significant and the process is causal.
- AR(2): The 95% confidence interval of $\hat{\phi}_2$ includes 0 which indicates that this coefficient is not significant.
- MA(1): Compared with other models, AIC of this model is significantly higher. Besides, after applying this model, the residuals does not seem to be stationary.
- ARMA(1, 1): The 95% confidence interval of $\hat{\theta}_1$ includes 0 which indicates that this coefficient is not significant.

Given the second estimated model, we predict 40 months ahead and demonstrate results in the following plot. The red solid line shows the future predictions while the blue dashed lines indicate 95% confidence range of predictions.

```
pass.ar1.pr = predict(pass.ar1, n.ahead = 40)

newTime = seq(from=1972+4/12, to=1975+7/12, by=1/12)
newMonths = factor(rep(1:12, 4))
levels(newMonths) = month.abb
```

```
newMonths = newMonths[-c(1:5, 46:48)]
trend = predict(modelB, newdata=data.frame(t1=newTime, t2=newTime^2, M=newMonths))

plot(passenger, ylab = "Number of Passangers", xlim = c(1962+6/12, 1975+7/12), main="")
lines(trend + pass.ar1.pr$pred, type="o", col="red")
lines(trend + pass.ar1.pr$pred + 1.96*pass.ar1.pr$se, lty=2, col="blue")
lines(trend + pass.ar1.pr$pred - 1.96*pass.ar1.pr$se, lty=2, col="blue")
```

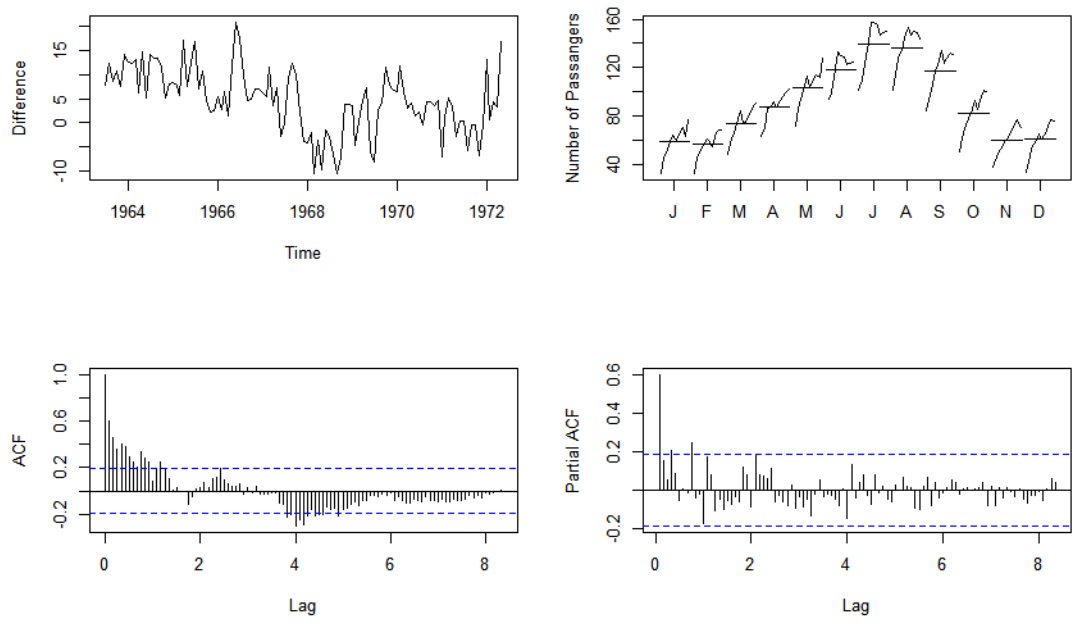


3 SARIMA Model

We apply the differening with the order 1 and the lag 12. After removing the seasonality, we observe the tails-off pattern in ACF plot and the cut-off pattern in PACF plot after lag 2, which indicate that the AR(2) model may be appropriate for the remainder.

```
pass.diff = diff(passenger, lag = 12, differences = 1)
plot(pass.diff, ylab = "Difference")
acf(pass.diff, lag.max = 100, main = "")
pacf(pass.diff, lag.max = 100, main = "")
```

Also, within the seasonality, we also observe the tails-off pattern in the ACF plot and the cut-off pattern in the PACF plot after lag 1, which indicates that AR(1) model may be appropriate for the seasonal component. Thus, we decide to adopt ARIMA$(2,0,0) \times (1,1,0)_{12}$ model.



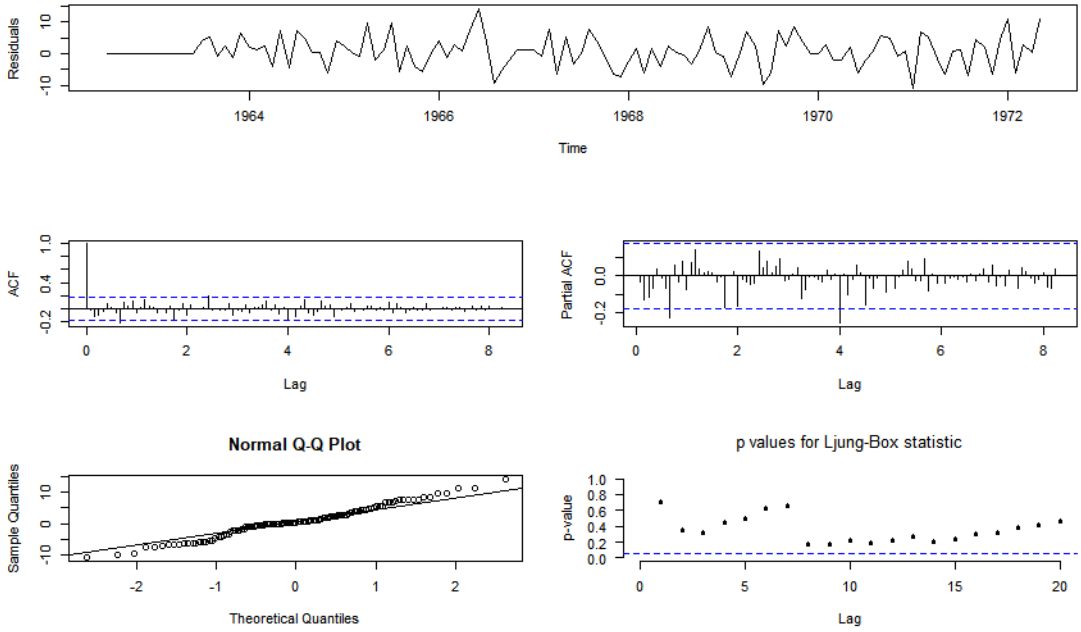

```
# Results of other SARIMA models for comparisons

    (1,0,0) * (0,1,1)             (1,0,0) * (1,1,1)
         ar1     sma1          ar1    sar1     sma1
      0.9304  -0.5806       0.9436  0.1378  -0.7103
s.e.  0.0577   0.1568 s.e.  0.0565  0.1657   0.1853

    (0,0,1) * (1,1,0)             (1,0,1) * (1,1,0)
         ma1     sar1          ar1      ma1     sar1
      0.4959   0.1695       0.950  -0.3947  -0.3486
s.e.  0.0775   0.1168 s.e.  0.038   0.1313   0.1013

# For models above, the 95% confidence intervals of
# some coefficents include 0 or 1.

pass.arima = arima(passenger, order=c(2,0,0), seasonal=list(order=c(1,1,0), period=12))

Call:
arima(x = passenger, order = c(2, 0, 0),
      seasonal = list(order = c(1, 1, 0), period = 12))

Coefficients:
         ar1     ar2     sar1
      0.6387  0.2550  -0.3717
s.e.  0.0957  0.0966   0.0994

sigma^2 estimated as 24.96:  log likelihood = -325.45,  aic = 658.89
```

Given the R output, we write the model as follows.

$$(1 + 0.6387B + 0.255B^2)(1 - B^{12})x_t = (1 - 0.3717B^{12})w_t, \qquad w_t \sim (0, 24.96) \tag{3.1}$$

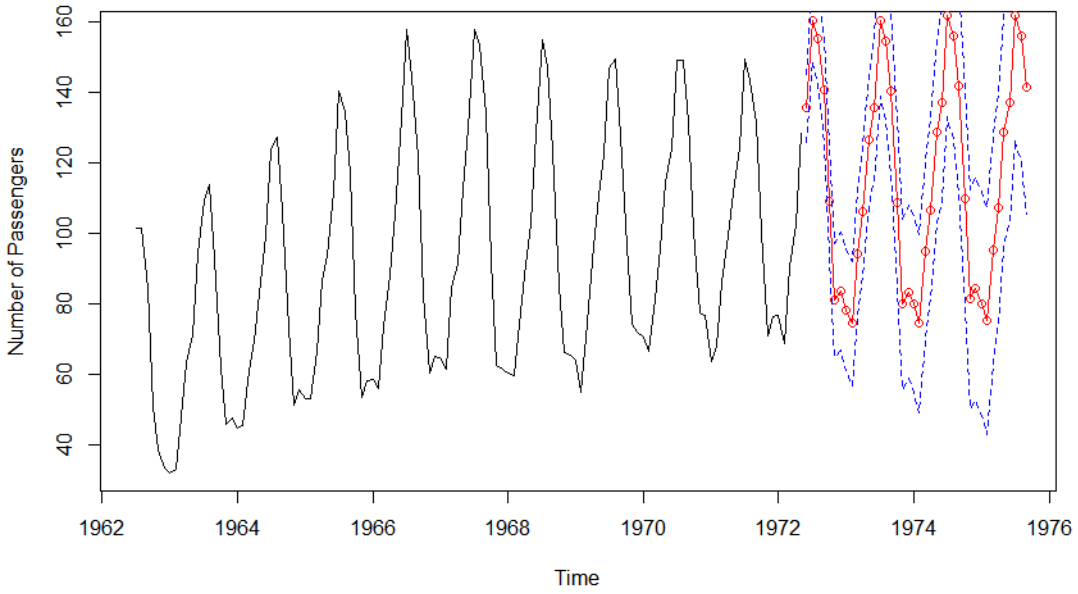
We also plot the residuals to see if it is stationary. The residuals seem to be random and stationary according to the ACF and the PACF plots. The Q-Q plot demonstrates the normality of residuals and the Ljung-Box test indicates that the residuals are white noise at the 5% significance level.

```
resC = pass.arima$residuals
plot(resC, ylab = "Residuals")
acf(resC, lag.max = 100, main = "")
pacf(resC, lag.max = 100, main = "")
qqnorm(resC)
qqline(resC)
Box.Ljung.Test(resC, lag = 20)
```



Given the estimated model, we predict 40 months ahead and demonstrate results in the following plot. The red solid line shows the future predictions while the blue dashed lines indicate 95% confidence range of predictions.

```
pass.arima.pr = predict(pass.arima, n.ahead = 40)
plot(passenger, ylab="Number of Passengers", xlim = c(1962+6/12, 1975+7/12), main="")
lines(pass.arima.pr$pred, type="o", col="red")
lines(pass.arima.pr$pred + 1.96*pass.arima.pr$se, lty=2, col="blue")
lines(pass.arima.pr$pred - 1.96*pass.arima.pr$se, lty=2, col="blue")
```



4 Model Comparison

Both models simulate the trend and seasonality remarkably and they have few differences of simulating the seasonality. However, the model in Section 2 shows a downward trend due to the quadratic component while the one in Section 3 does not. Given the historical data showing that the trend has become horizontal close to the end, we consider that the SARIMA model is more accurate.

5 Conclusion

Given the analysis above, we prefer the SARIMA model for the future predictions.

References

[1] Passenger miles (Mil) flown domestic U.K. Jul. '62-May '72. Retrieved from https://datamarket.com/data/set/22mb/passenger-miles-mil-flown-domestic-uk-jul-62-may-72