ECON 144: Project 1

Amos Tong, Shreesh Agarwal, Kevin Tian, and Matthew Craig

Contents

I. Introduction

The dataset selected for this project records the monthly **Air Revenue Passenger Miles** (ARPM) reported by the U.S. Bureau of Transportation Statistics. The data contains 240 observations of ARPM and the dates on which it was recorded, from January 2000 to December 2019 (not including 2020 and beyond). Data is reported in thousands, which we convert to millions for ease of display. ARPM is a statistic that measures the volume of air passengers transported. One ARPM is equal to one paying passenger carried one mile via air transport. Therefore, the analysis of this dataset will provide insight into the trend, seasonality, and overall behavior of the volume of air passengers transported.

Source

II. Results

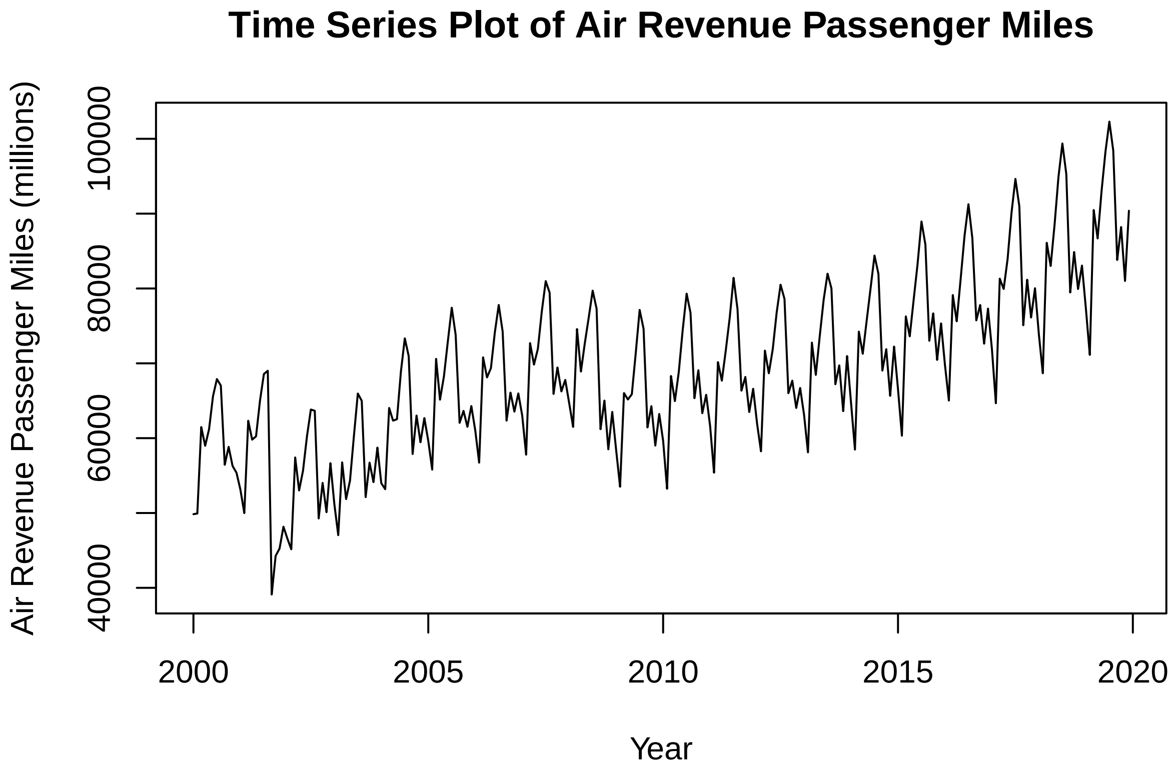
1. Modeling and Forecasting Trend

(a) Show a time-series plot of your data.

```r
# Read data in from csv downloaded from FRED
air_rpm <- read.csv("AIRRPMTSI.csv", header=TRUE)
head(air_rpm)
```

```
##         DATE AIRRPMTSI
## 1 2000-01-01  49843099
## 2 2000-02-01  49931931
## 3 2000-03-01  61478163
## 4 2000-04-01  58981617
## 5 2000-05-01  61223861
## 6 2000-06-01  65601574
```

```
# Clear out date column (the ts object will assign a date for us)
air_rpm$DATE <- NULL
# Create a ts object (converting to millions instead of thousands) and plot
air_rpm <- ts(air_rpm[,1] / 1000, start=2000, freq=12)
options(scipen=5)
plot(air_rpm, xlab="Year", ylab="Air Revenue Passenger Miles (millions)",
     main = "Time Series Plot of Air Revenue Passenger Miles")
```

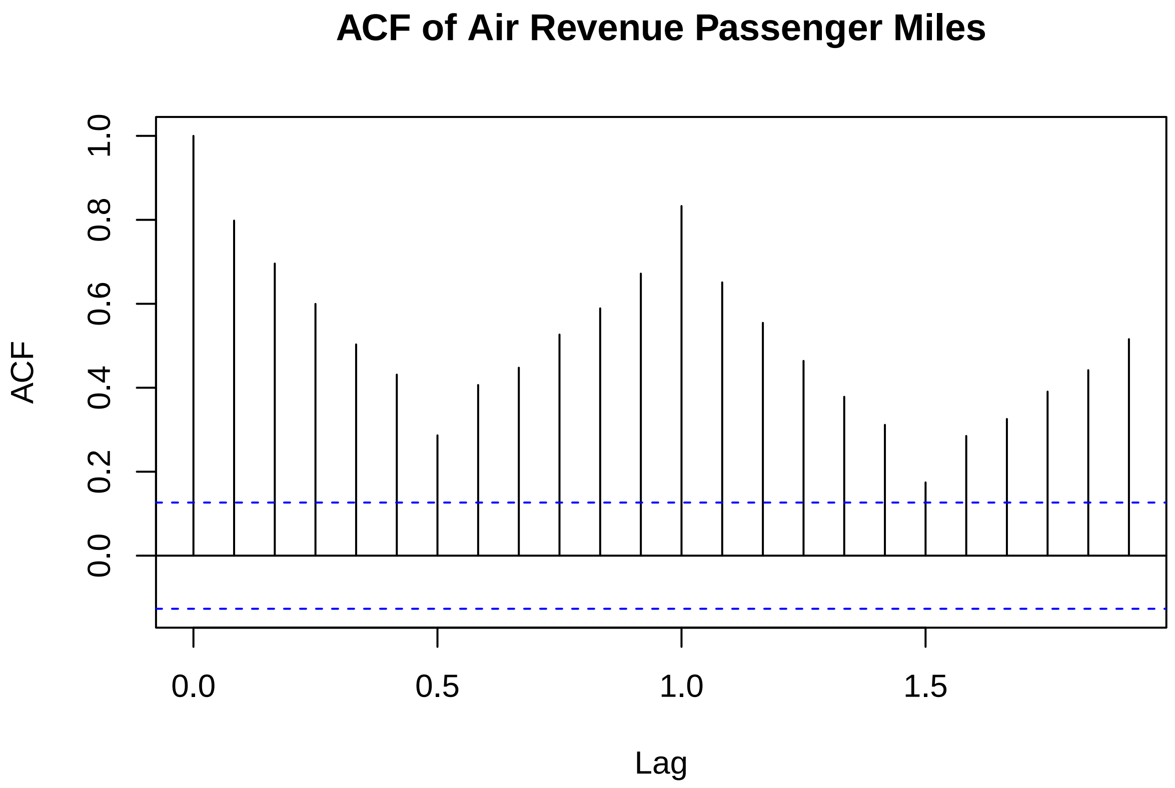


(b) Does your plot in (a) suggest that the data are covariance stationary? Explain your answer.

The plot suggests that there is a long-term upward trend in the time-series; therefore, it is unlikely that the means of the underlying random variables are identical, which means that the time-series is not covariance stationary of any order.

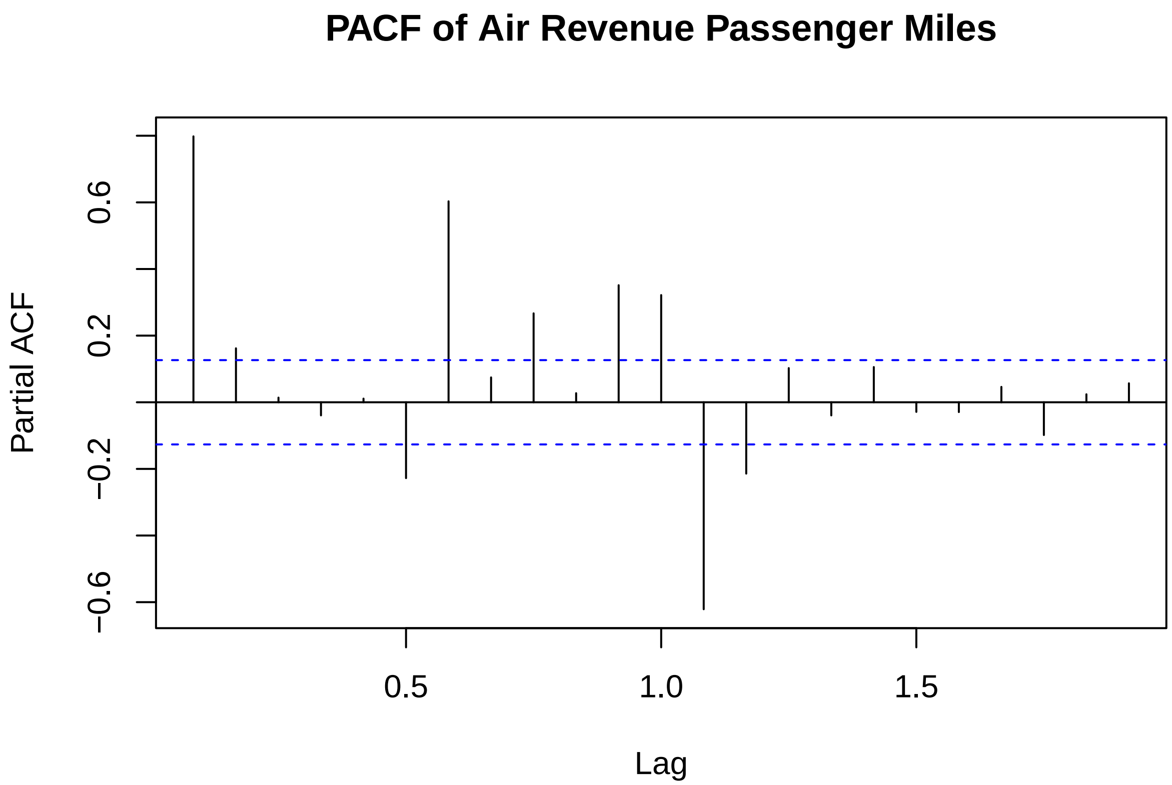
(c) Plot and discuss the ACF and PACF of your data.

```
acf(air_rpm, main = "ACF of Air Revenue Passenger Miles")
```



According to the ACF plot, there is statistically significant positive correlation between passenger miles and almost all lag intervals. The upward-downward pattern in the ACF plot may also suggest seasonality, as ACF decreases until a half-year (6 month) lag, then increases to a 1 year (12 month) lag, before decreasing again.

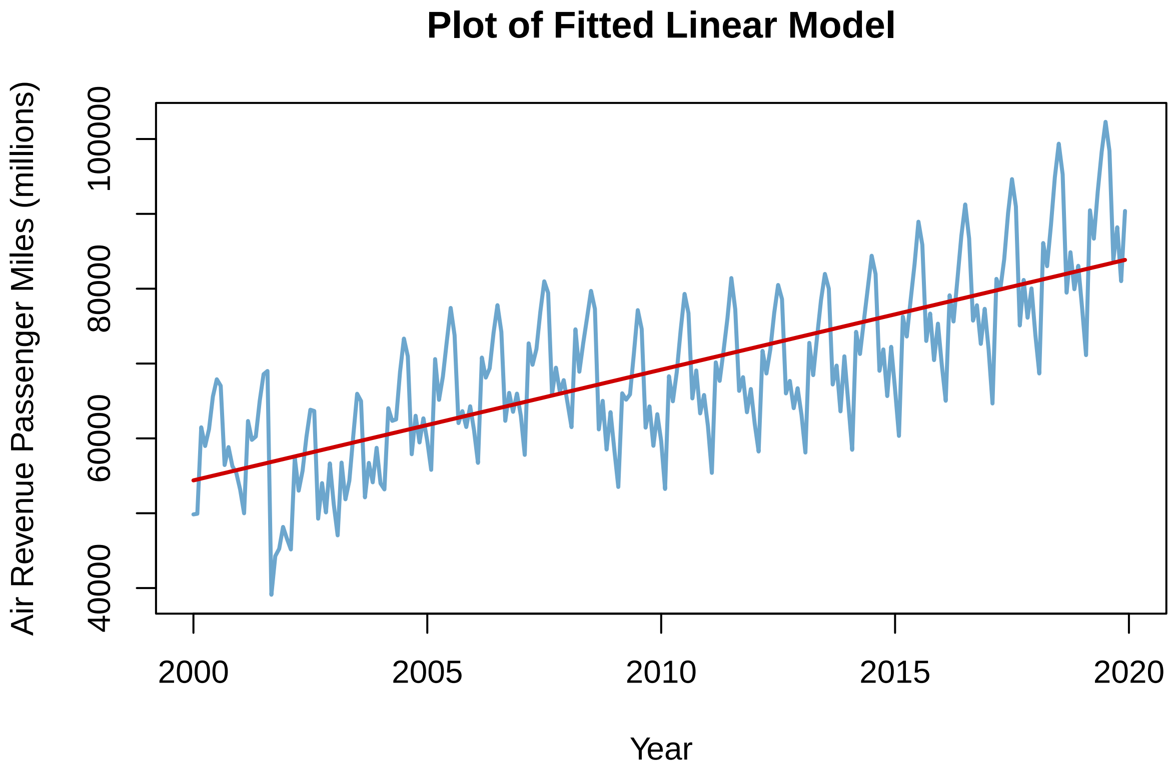
```
pacf(air_rpm, main = "PACF of Air Revenue Passenger Miles")
```



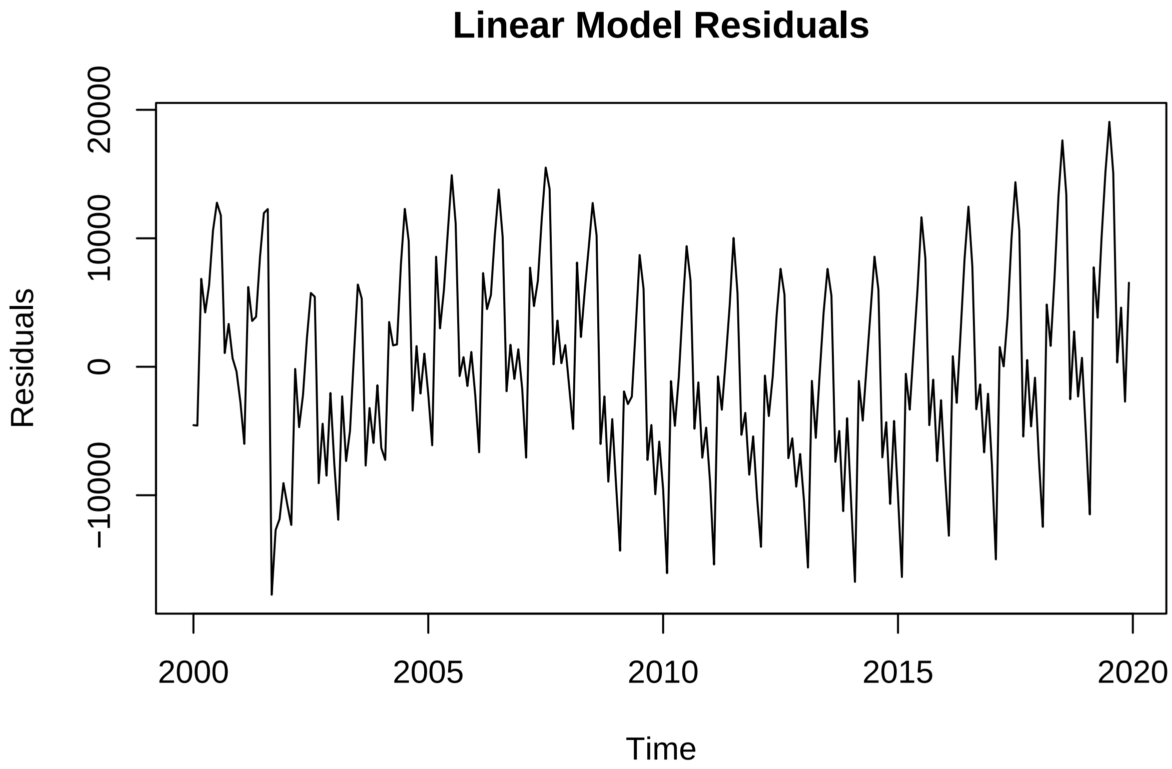
According to the PACF plot, there is a statistically significant time dependence due to multiple significant spikes.

(d) Fit a linear and nonlinear (e.g., polynomial, exponential, quadratic + periodic, etc.) model to your series. In one window, show both figures of the original times series plot with the respective fit.

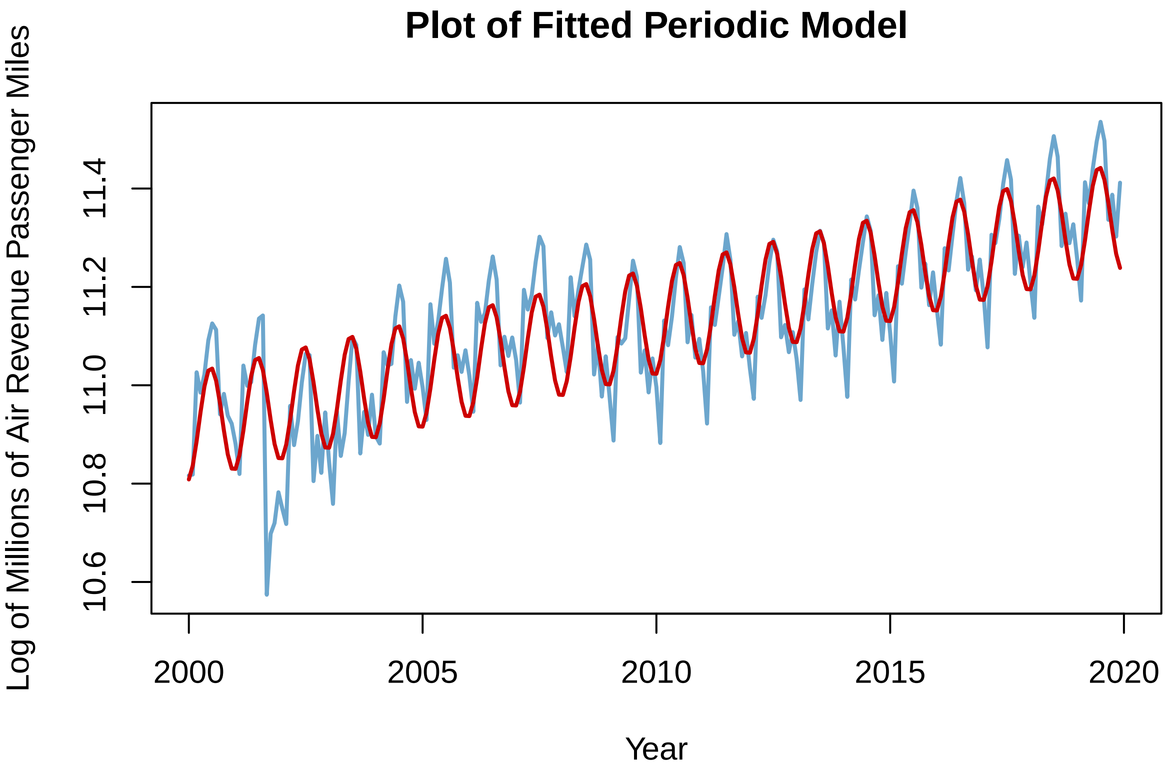
```r
# Linear model
t <- seq(from = 2000, by = 1/12, length = length(air_rpm))
m1 <- lm(air_rpm ~ t)
# Plot model
plot(air_rpm, ylab="Air Revenue Passenger Miles (millions)", xlab="Year", lwd=2,
     col='skyblue3', xlim=c(2000,2020), main="Plot of Fitted Linear Model")
lines(t, m1$fit, col="red3", lwd=2)
```



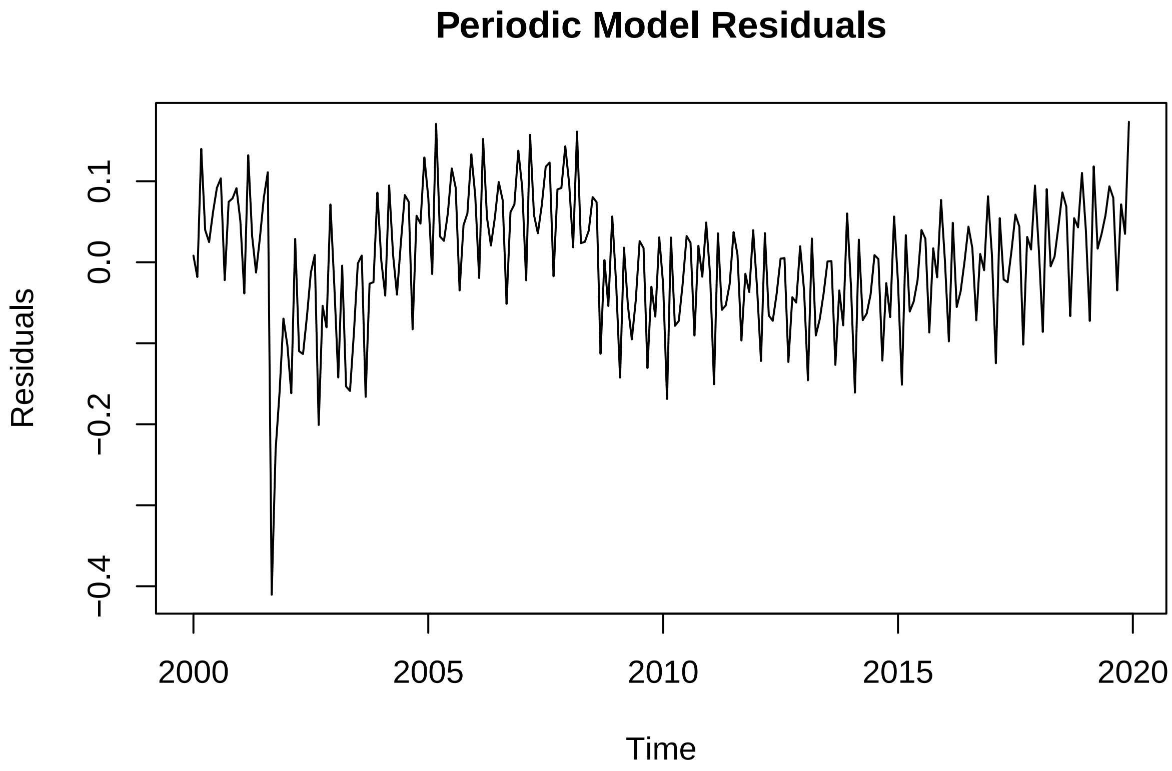
```
# Plot residuals
plot(t, m1$res, ylab="Residuals", type = 'l', xlab="Time", main="Linear Model Residuals")
```



```
# Log-periodic model
sin.t<-sin(2*pi*t)
cos.t<-cos(2*pi*t)

m2 <- lm(log(air_rpm) ~ t  + sin.t + cos.t)
# Plot Model
plot(log(air_rpm), ylab="Log of Millions of Air Revenue Passenger Miles", xlab="Year",
     lwd = 2, col='skyblue3', xlim=c(2000,2020), main = "Plot of Fitted Periodic Model")
lines(t,m2$fit,col="red3",lwd=2)
```

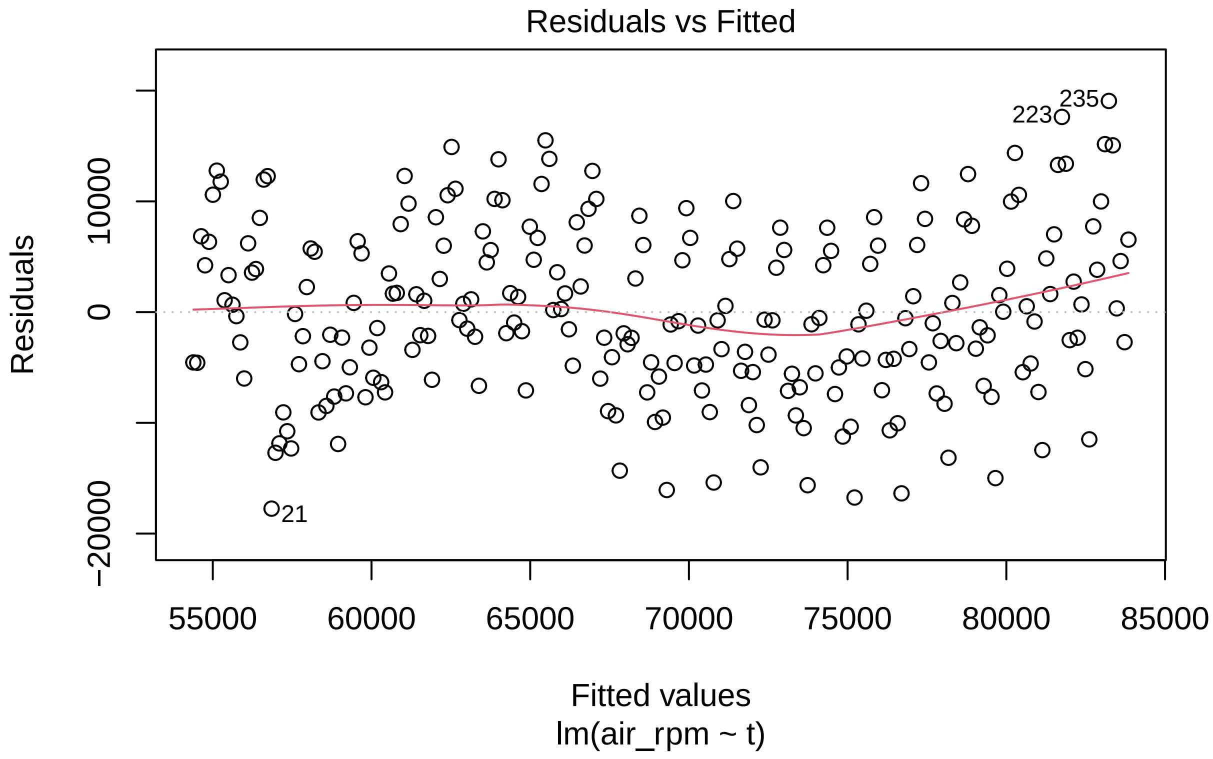


```
# Plot residuals
plot(t,m2$res, ylab="Residuals", type='l', xlab="Time", main = "Periodic Model Residuals")
```



(e) For each model, plot the respective residuals vs. fitted values and discuss your observations.

```
plot(lm(air_rpm ~ t), which = c(1,1))
```

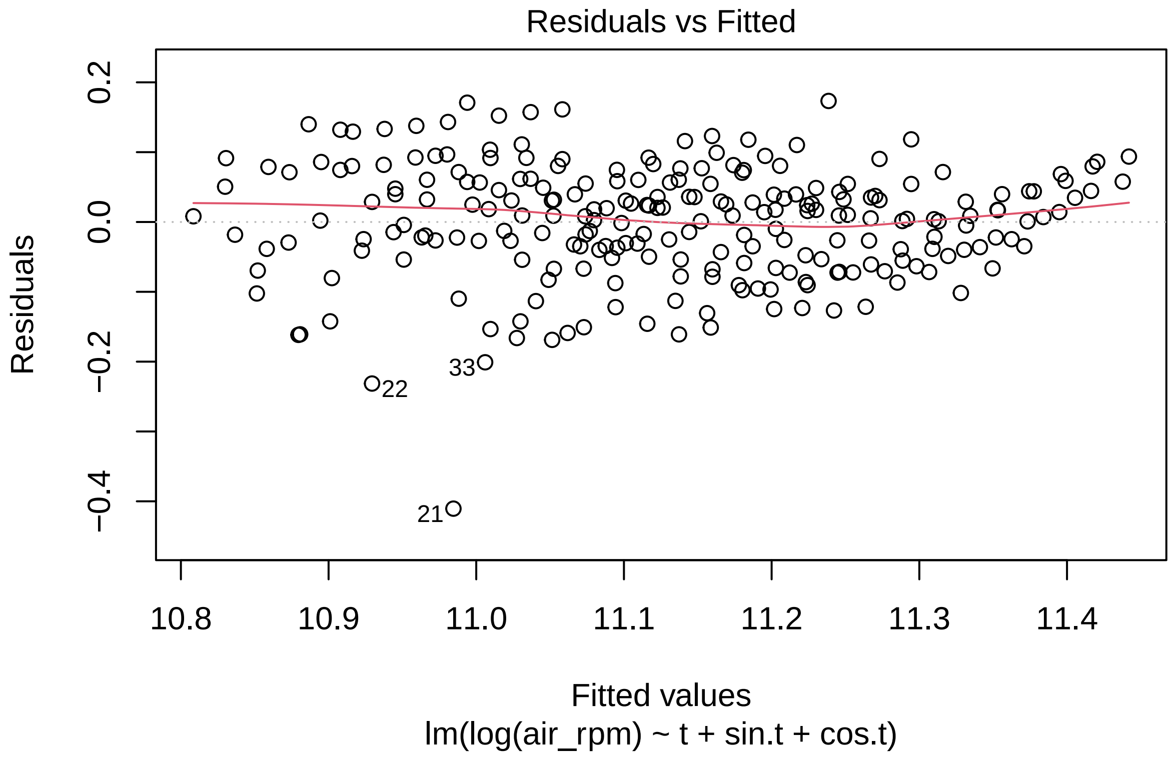


Observations

- The residuals are randomly distributed about the y=0 line which suggests that the residuals are normally distributed.
- The residuals do not form a "horizontal band" around the y=0 line. Instead, a non-linear deviation can be seen in the plot. Moreover, the magnitude of the residuals gets larger over time. This suggests that the variance of the residuals might not be constant.
- Magnitude or spread of residuals is very large.
- Potential outliers are present in the residuals, suggesting extraordinary events.

The observations above suggest that a linear model might not be the best fit for the time-series due to a violation of the assumption of constant variance of residuals, and observed non-linear effects in the residuals.

```
plot(lm(log(air_rpm) ~ t + sin.t + cos.t, data = air_rpm), which = c(1,1))
```



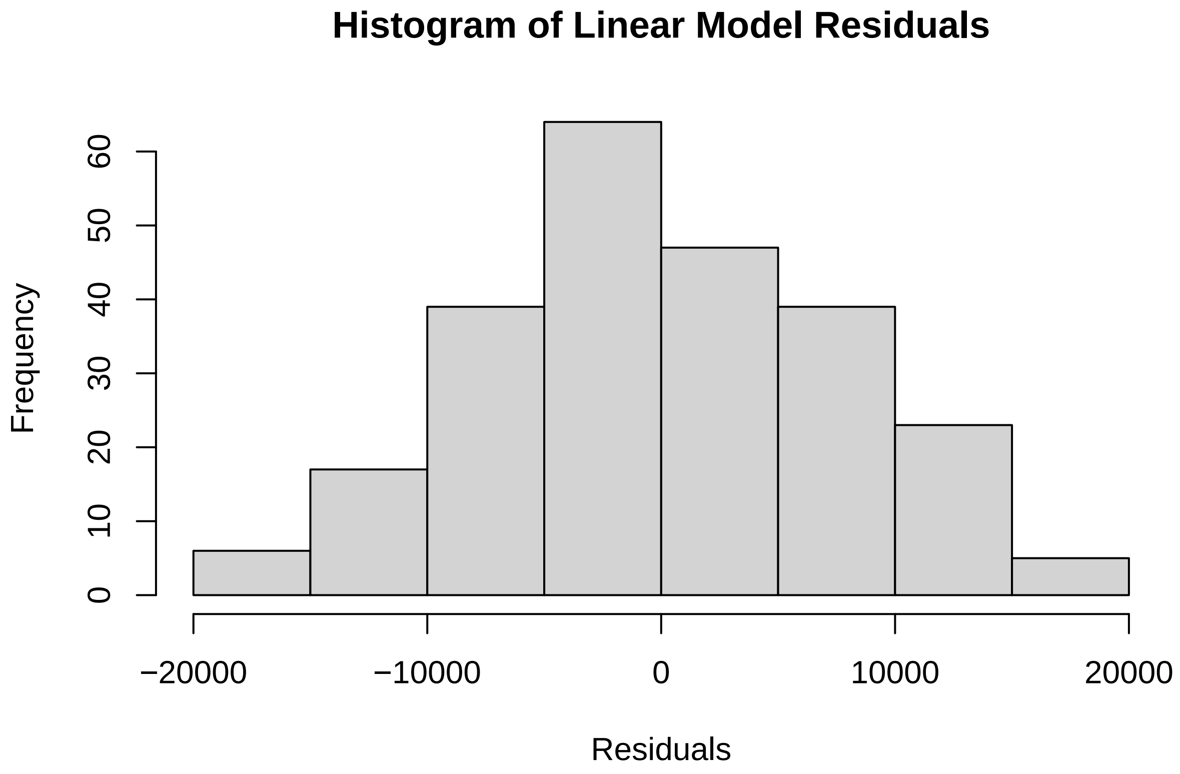
Observations

- The residuals are evenly distributed about the y=0 line. However, the spread of the residuals is changing over time, which suggests non-constant variance of residuals.
- Magnitude or spread of residuals is very small (on a log scale).
- Potential outliers are present, suggesting potential extraordinary events.

The observations above indicate that a log-periodic model fits the data better than a linear model.

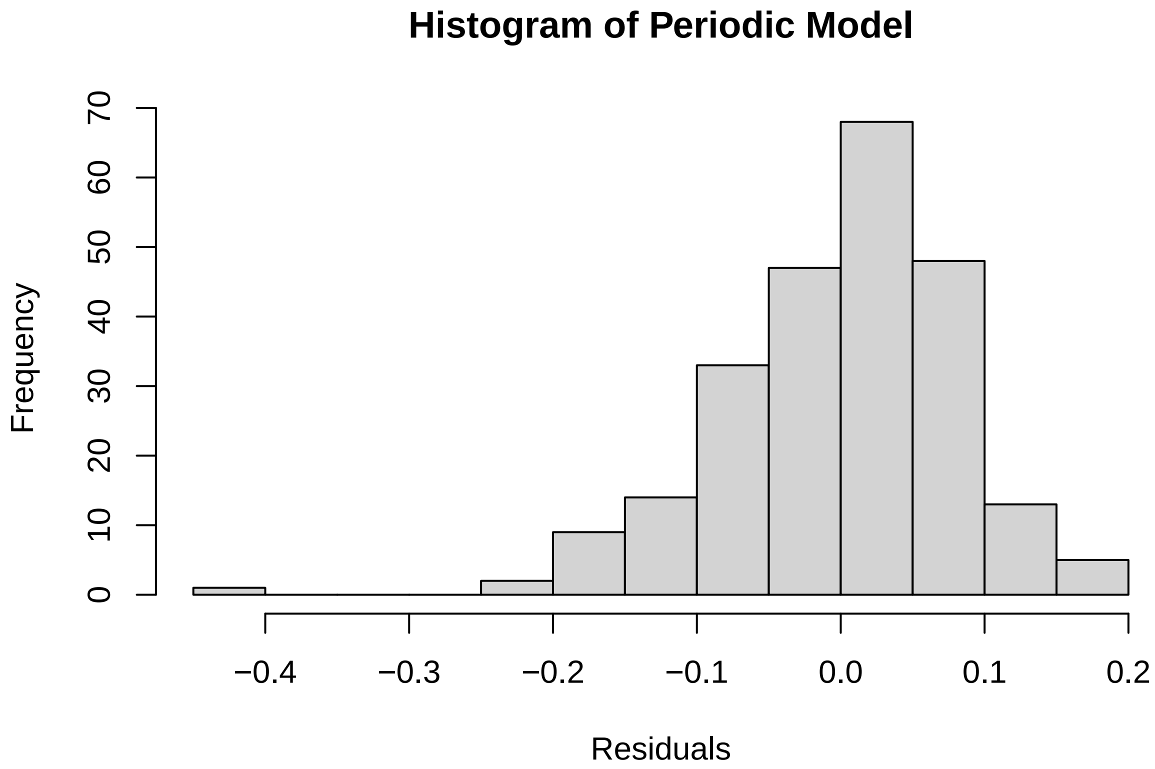
(f) For each model, plot a histogram of the residuals and discuss your observations.

```
hist(m1$residuals, xlab="Residuals", main="Histogram of Linear Model Residuals")
```



The distribution of residuals is slightly positively skewed and is roughly symmetrically distributed about 0 (especially compared to m2). However, the magnitude of residuals is very large compared to the scale of the data.

```
hist(m2$residuals, xlab="Residuals", main="Histogram of Periodic Model")
```



The distribution of residuals has a long negative tail, with some potential outliers. The magnitude and spread of residuals compared to the log scale of the data is very small.

(g) For each model, discuss the associated diagnostic statistics (R^2, t-distribution, F-distribution, etc.)

```
summary(m1)
```

```
##
## Call:
## lm(formula = air_rpm ~ t)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17745.6  -5328.8   -732.6   5805.7  19064.0
##
## Coefficients:
##                Estimate  Std. Error t value Pr(>|t|)
## (Intercept) -2903961.51   172407.50  -16.84   <2e-16 ***
## t               1479.17       85.78   17.25   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7672 on 238 degrees of freedom
## Multiple R-squared:  0.5555, Adjusted R-squared:  0.5536
## F-statistic: 297.4 on 1 and 238 DF,  p-value: < 2.2e-16
```

- The adjusted R^2 is 0.55, hence the model explains roughly half of the variation in the data.
- The t values are statistically significant. Hence, we are confident that the intercept and coefficient of t are statistically not 0 and that passenger miles is time-dependent to some extent.
- The F-statistic is statistically significant, indicating that at least one of the coefficients is statistically significant; m1 is a better model than a model with no independent variables.

```
summary(m2)
```

```
##
## Call:
## lm(formula = log(air_rpm) ~ t + sin.t + cos.t)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41049 -0.04893  0.00907  0.05647  0.17331
##
## Coefficients:
##                Estimate  Std. Error t value Pr(>|t|)
## (Intercept) -32.0325146   1.8458088 -17.354  < 2e-16 ***
## t             0.0214742   0.0009183  23.384  < 2e-16 ***
## sin.t         0.0239786   0.0074977   3.198  0.00157 **
## cos.t        -0.1073529   0.0074926 -14.328  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08207 on 236 degrees of freedom
## Multiple R-squared:  0.7641, Adjusted R-squared:  0.7611
## F-statistic: 254.8 on 3 and 236 DF,  p-value: < 2.2e-16
```

- The adjusted R^2 is 0.76, hence the model explains more of the variation in the data than m1.
- The t values are statistically significant only for all coefficients. Hence, we are confident that the coefficient of these variables are statistically not 0. This means that the added trigonometric transformations of time help to explain more variation in the data, especially the variation caused by non-linear components like seasonality.
- The F-statistic is statistically significant, indicating that at least one of the coefficients is statistically significant; m2 is a better model than a model with no independent variables.

(h) Select a trend model using AIC and one using BIC (show the values obtained from each criterion). Do the selected models agree?

```
AIC(m1, m2)
```

```
##    df       AIC
## m1  3 4978.8417
## m2  5 -513.0125
```

The AIC of m2 is far lower than m1. Hence, m2 is the better model according to AIC.

```
BIC(m1, m2)
```

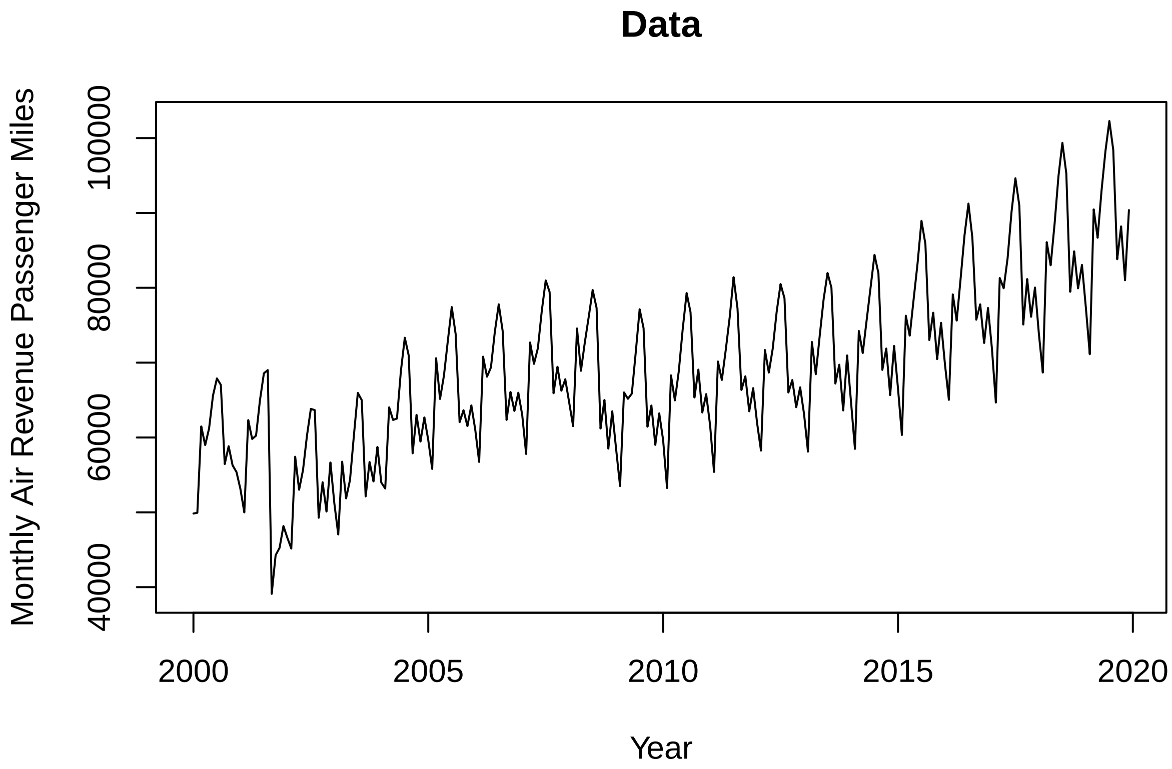
```
##    df       BIC
## m1  3 4989.2836
## m2  5 -495.6093
```

The BIC of m2 is far lower than m1. Hence, m2 is the better model according to BIC.

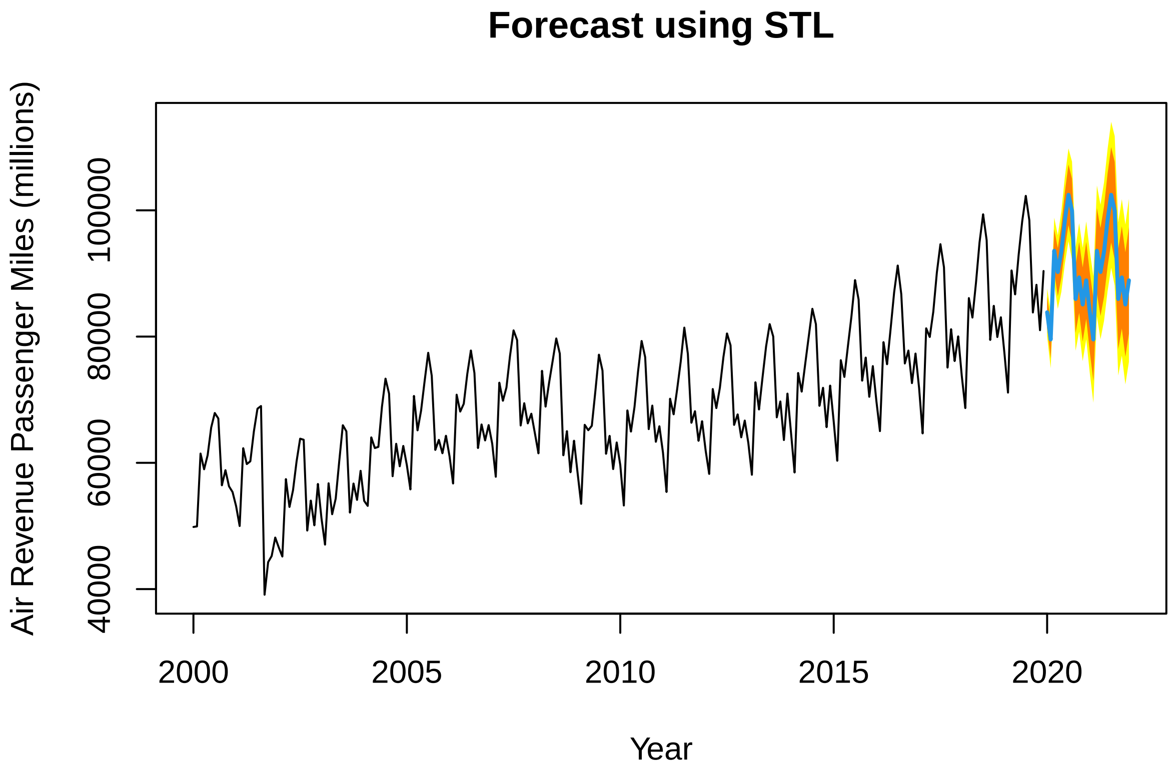
Both AIC and BIC agree than m2 is a better model than m1.

(i) Use your preferred model to forecast h-steps (at least 16) ahead. Your forecast should include the respective uncertainty prediction interval. Depending on your data, h will be in days, months, years, etc.

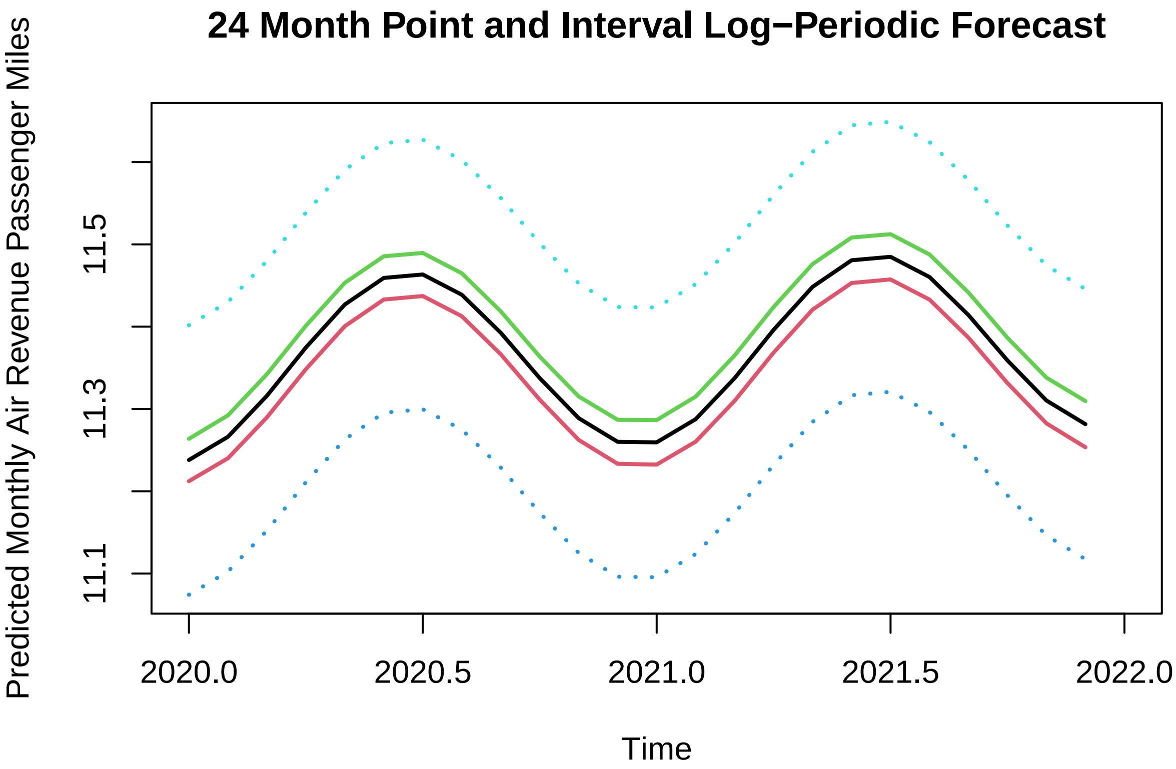
```
plot(air_rpm, main="Data", xlab="Year", ylab="Monthly Air Revenue Passenger Miles")
```



```
plot(forecast(air_rpm, h = 24), main = "Forecast using STL", xlab = "Year",
     ylab = "Air Revenue Passenger Miles (millions)", shadecols = "oldstyle")
```



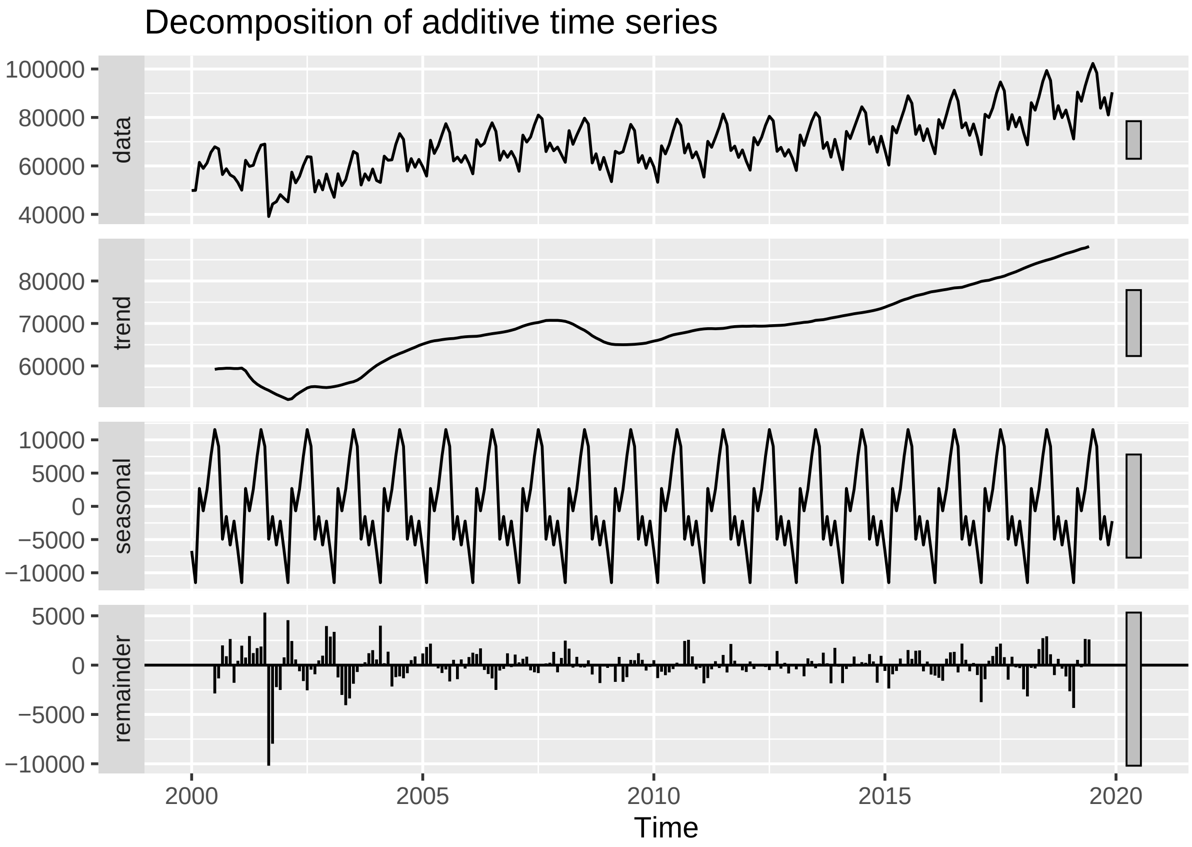
```
t = seq(from = 2020, by = 1/12, length = 24)
tn = data.frame(t = t, sin.t = sin(2*pi*t), cos.t = cos(2*pi*t))
pred = predict(m2, tn, se.fit = TRUE)
pred.plim = predict(m2, tn, level=0.95, interval="prediction")
pred.clim = predict(m2, tn, level=0.95, interval="confidence")
matplot(tn$t, cbind(pred.clim, pred.plim[,-1]), lty=c(1,1,1,3,3), type="l",
        xlim=c(2020, 2022), lwd=2,ylab="Predicted Monthly Air Revenue Passenger Miles",
        xlab="Time", main="24 Month Point and Interval Log-Periodic Forecast")
```



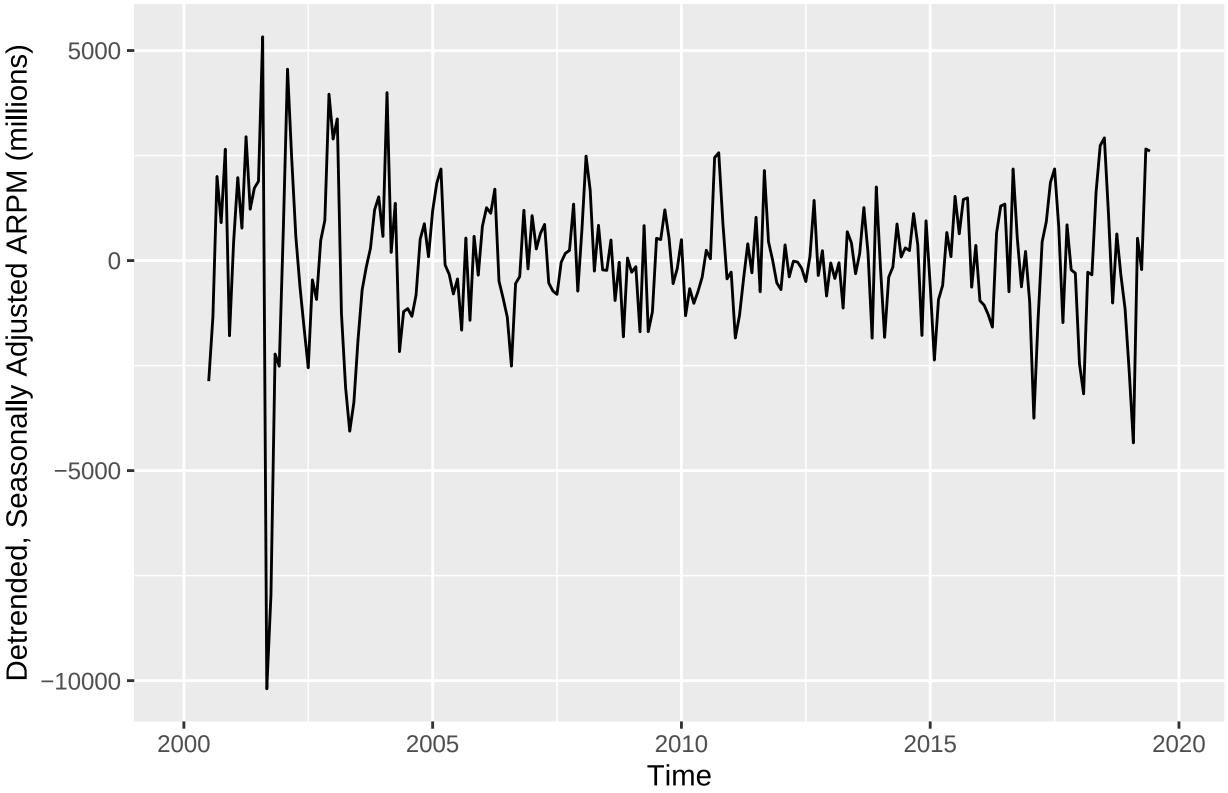
2. Trend and Seasonal Adjustmennts

(a) Perform an additive decomposition of your series. Remove the trend and seasonality, and comment on the ACF and PACF of the residuals (i.e., what is left after detrending and seasonally adjusting the series). Comment on the results.

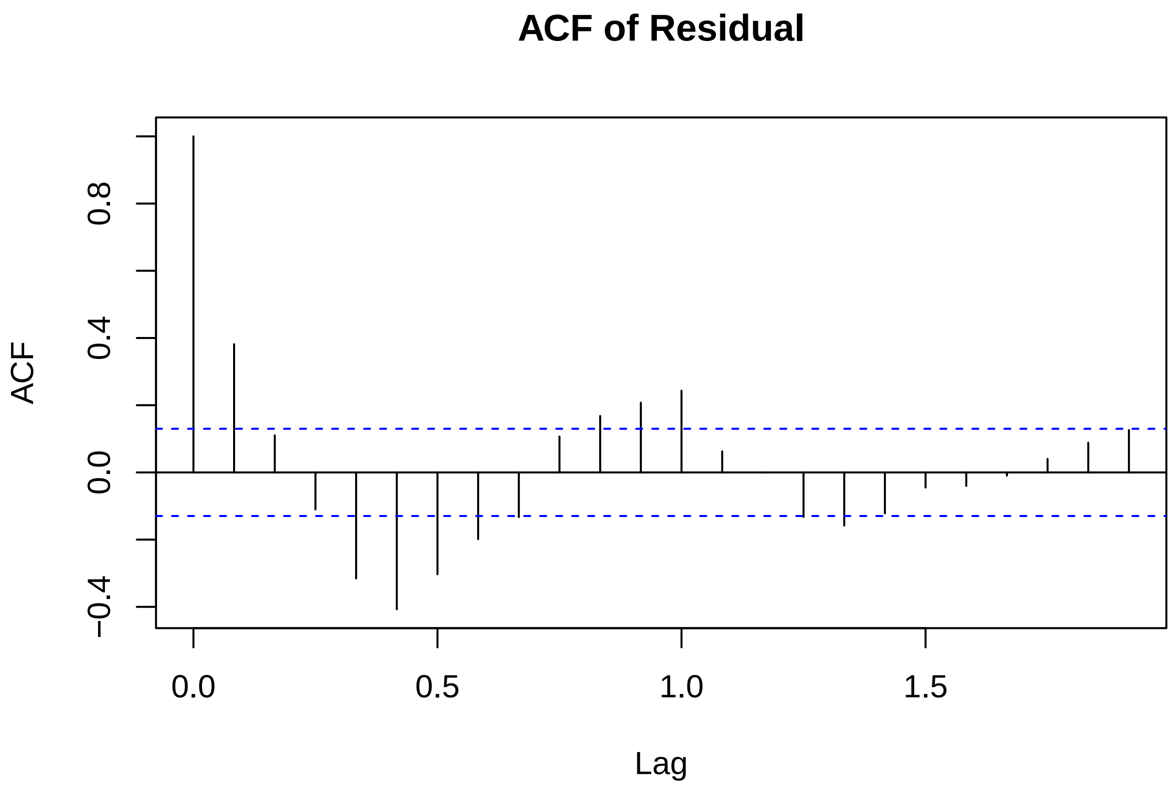
```
# Decompose the series (extract the components)
dcmp_arpm <- decompose(air_rpm, "additive")
autoplot(dcmp_arpm)
```



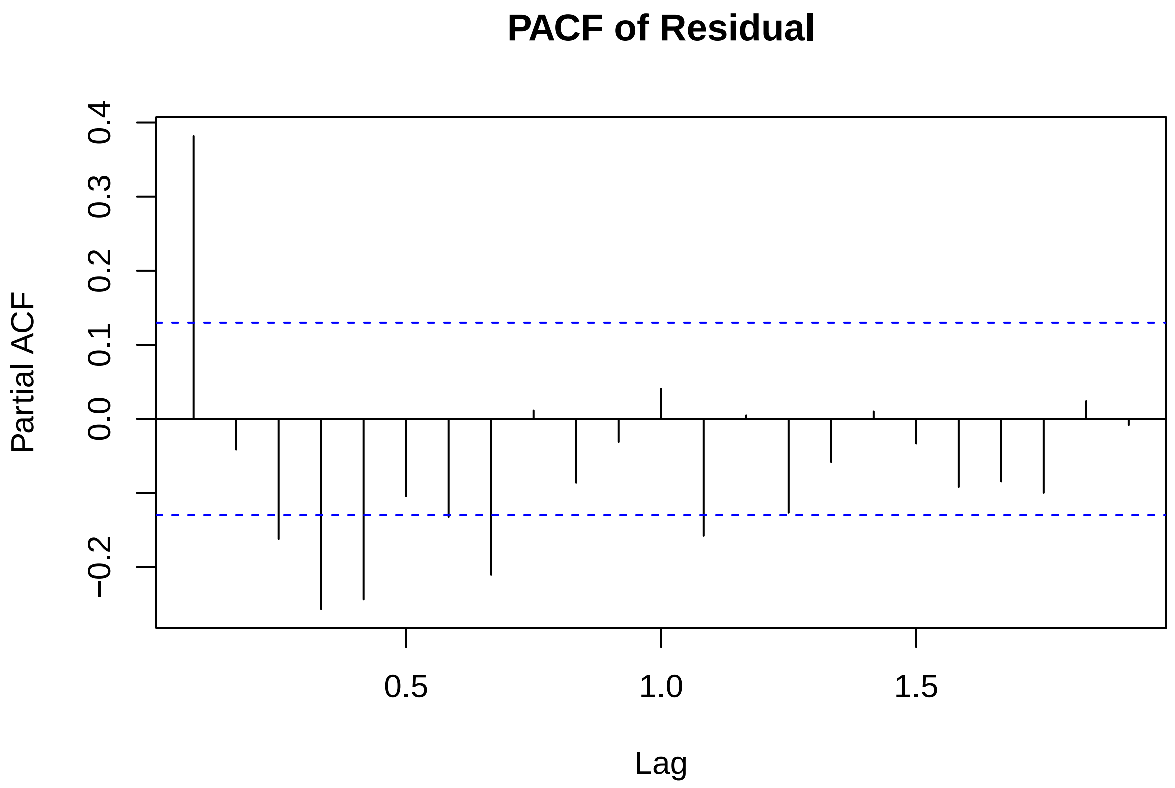
```
# Store the inidividual components
trend_arpm <- dcmp_arpm$trend
seasonal_arpm <- dcmp_arpm$seasonal
random_arpm <- dcmp_arpm$random

# Remove trend and seasonality
detrend_seas_adj_arpm <- air_rpm - trend_arpm - seasonal_arpm
autoplot(detrend_seas_adj_arpm, ylab="Detrended, Seasonally Adjusted ARPM (millions)")
```



```r
# ACF(detrend_seas_adj_arpm)
acf(detrend_seas_adj_arpm,na.action = na.omit, main = "ACF of Residual")
```



```
# PACF(detrend_seas_adj_arpm)
pacf(detrend_seas_adj_arpm,na.action = na.omit, main = "PACF of Residual")
```

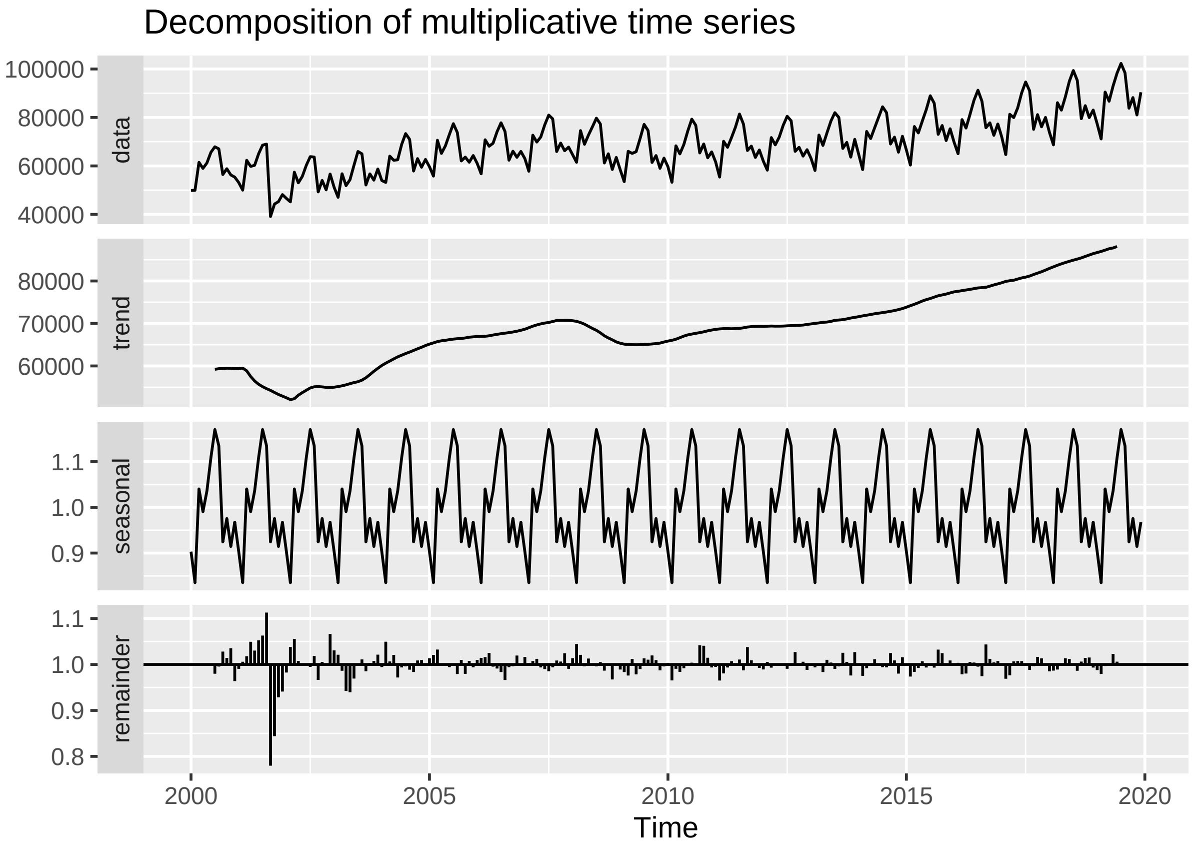


Trend and seasonality is removed. However, the variance of the detrended and seasonally adjusted time series is not constant.

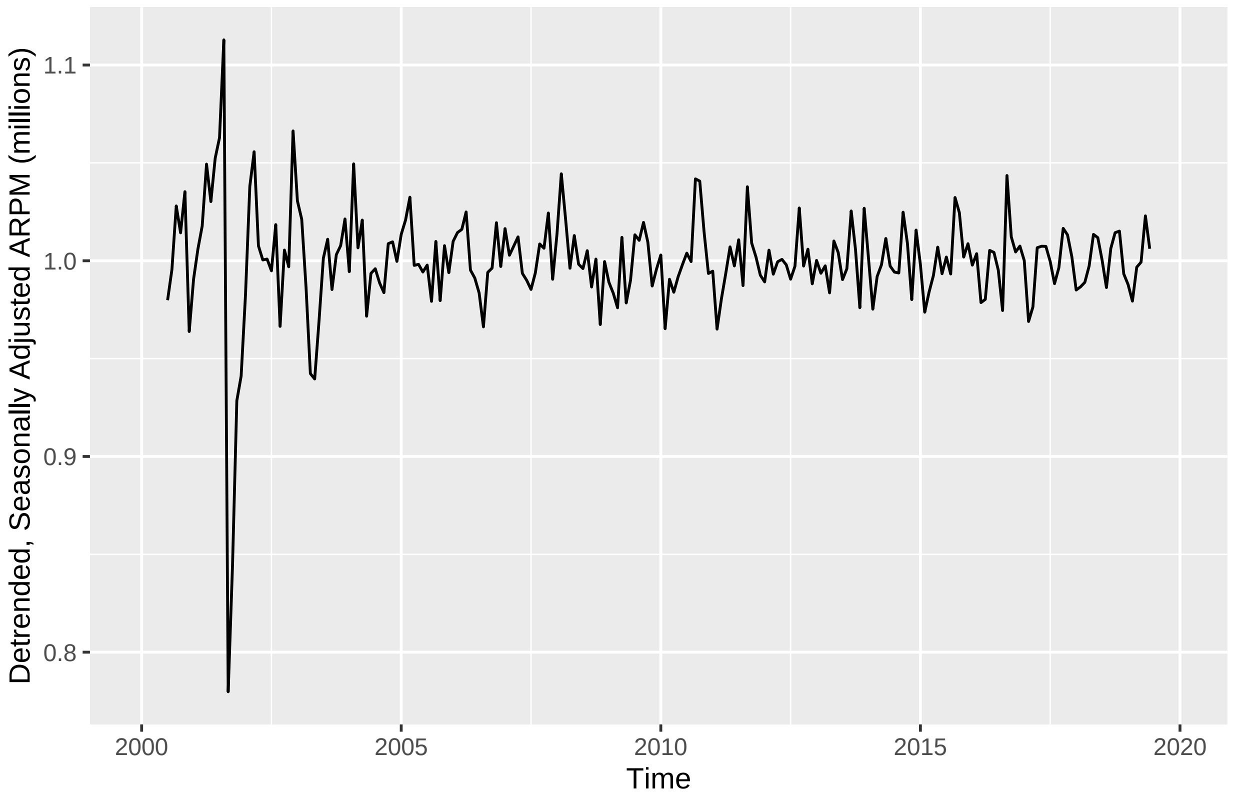
Although the ACF and PACF plot of the residuals shows significantly less autocorrelation compared to the original time series, we can still observe that there is some amount of autocorrelation at a lag of one month, as demonstrated by the significant spikes in the ACF and PACF plot. There is also a significant negative correlation at a 4 or 5 month lag.

(b) Perform a multipliative decomposition of your series. Remove the trend and seasonality, and comment on the ACF and PACF of the residuals (i.e., what is left after detrending and seasonally adjusting the series). Comment on the results.

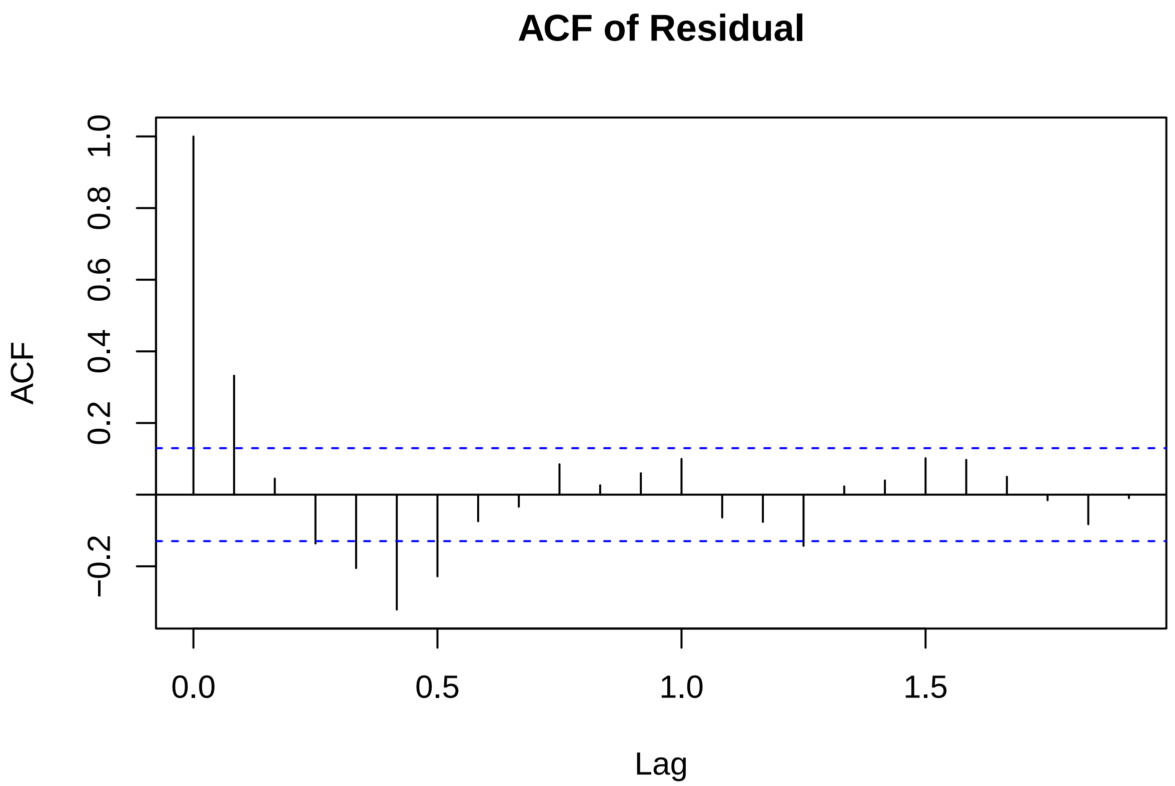
```
# Decompose the series (extract the components)
Mdcmp_arpm <- decompose(air_rpm, "multiplicative")
autoplot(Mdcmp_arpm)
```



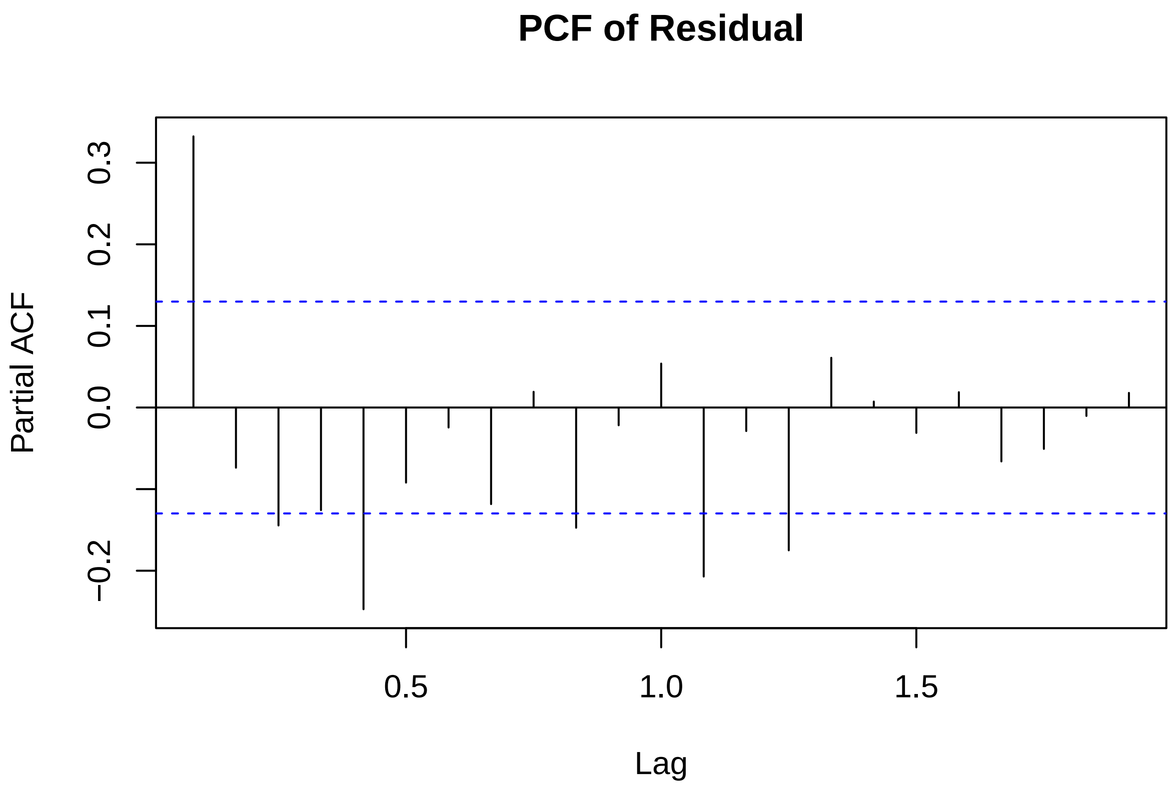
```
# Store the inidividual components
Mtrend_arpm = Mdcmp_arpm$trend
Mseasonal_arpm = Mdcmp_arpm$seasonal
Mrandom_arpm = Mdcmp_arpm$random

# Remove trend and seasonality
Mdetrend_seas_adj_arpm = (air_rpm/Mtrend_arpm)/Mseasonal_arpm
autoplot(Mdetrend_seas_adj_arpm, ylab="Detrended, Seasonally Adjusted ARPM (millions)")
```



```
# ACF of residual
acf(Mdetrend_seas_adj_arpm,na.action = na.omit, main = "ACF of Residual")
```



```
# PACF of residual
pacf(Mdetrend_seas_adj_arpm,na.action = na.omit, main = "PCF of Residual")
```



Trend and seasonality is removed. The variance of the detrended and seasonally adjusted time series is roughly constant except in 2001 (probably due to 9/11).

Similar to the additive method, the ACF and PACF plot shows much less autocorrelation than the original data. However, there are still a few significant spikes, around a 5 month lag.

(c) Which decomposition is better, additive or multiplicative? Why?

The multiplicative decomposition is better because the variance of the random component remains roughly constant throughout, and the magnitude of the random component is relatively small.

(d) Based on the two decompositions, and interpretation of the random components, would your models for the cycles be similar (additive vs. multiplicative) or very different? Why?

Both additive and multiplicative methods of decomposition provided similar results in terms of trend, seasonality, and residuals as seen in the graphs. Furthermore, the ACF and PACF of the residuals shows similar behavior with significantly less autocorrelation overall past a 1 year lag. Therefore, we can conclude that models for cycles would be very similar.

III. Conclusions and Future Work

- Our final model is a log-periodic model, which gives a good performance on various parameters such as R^2, BIC, and AIC. It explains most of the variation in the data and provides a good fit for both the trend and the seasonal variation.

- We choose a non-linear design to model the time series in order to account for seasonality which the linear model was unable to capture.

Potential Improvements:

- There might be presence of a cycle in the series, a cycle that resurfaces over a longer period of time than that for which the data is available. We could perhaps improve forecasts by including such cyclical behavior in our model and analysis. However, to do this we would need more years of data.
- Since air traffic generally falls during periods of economic recession, we could incorporate forecasts of economic recession within our model to better understand and predict the random component observed in the data.

IV. References

1. U.S. Bureau of Transportation Statistics, Air Revenue Passenger Miles [AIRRPMTSI], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/AIRRPMTSI, January 19, 2022.

V. R Source Code

Our R source code is included in the document throughout, with comments.