ASSIGNMENT 7 SOLUTIONS

STAT 485/685 E100/G100: Applied Time Series Analysis

Fall 2020

Simon Fraser University

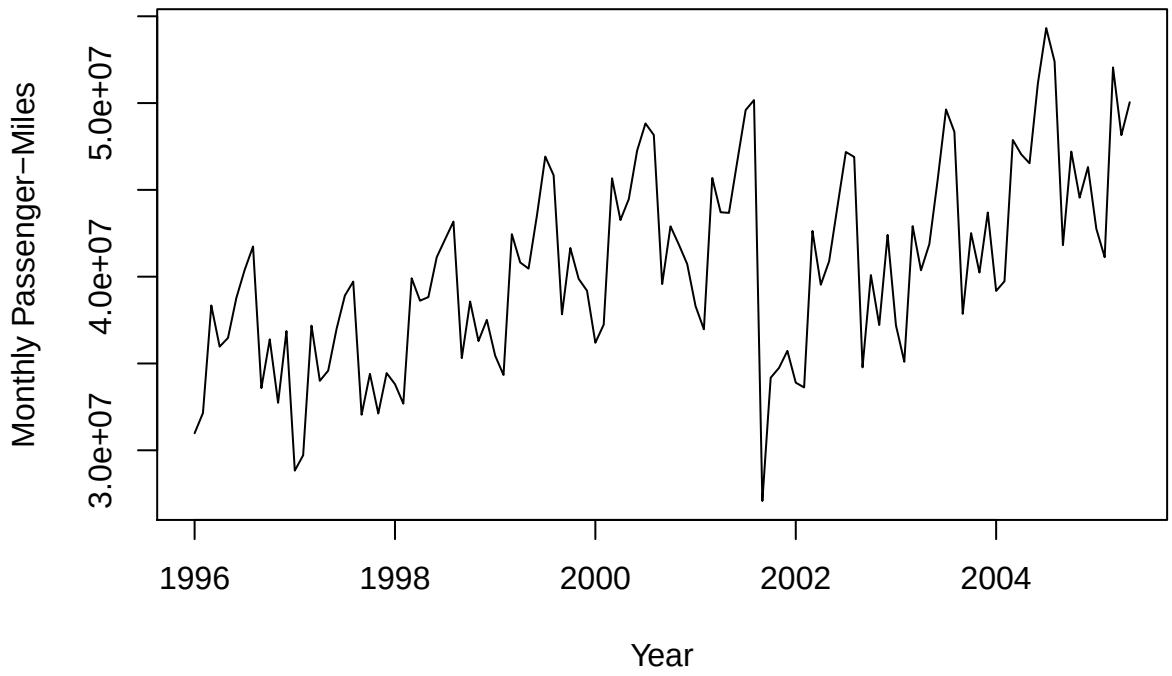
1. The "airmiles" dataset in the TSA package gives the monthly U.S. airline passenger-miles, from 1996 to 2005.

a) Load in and plot the time series dataset. Does this data appear to come from a stationary process? Why or why not?

Solution:

```
library(TSA)
```

```
data(airmiles)
plot(airmiles, xlab='Year', ylab='Monthly Passenger-Miles')
```

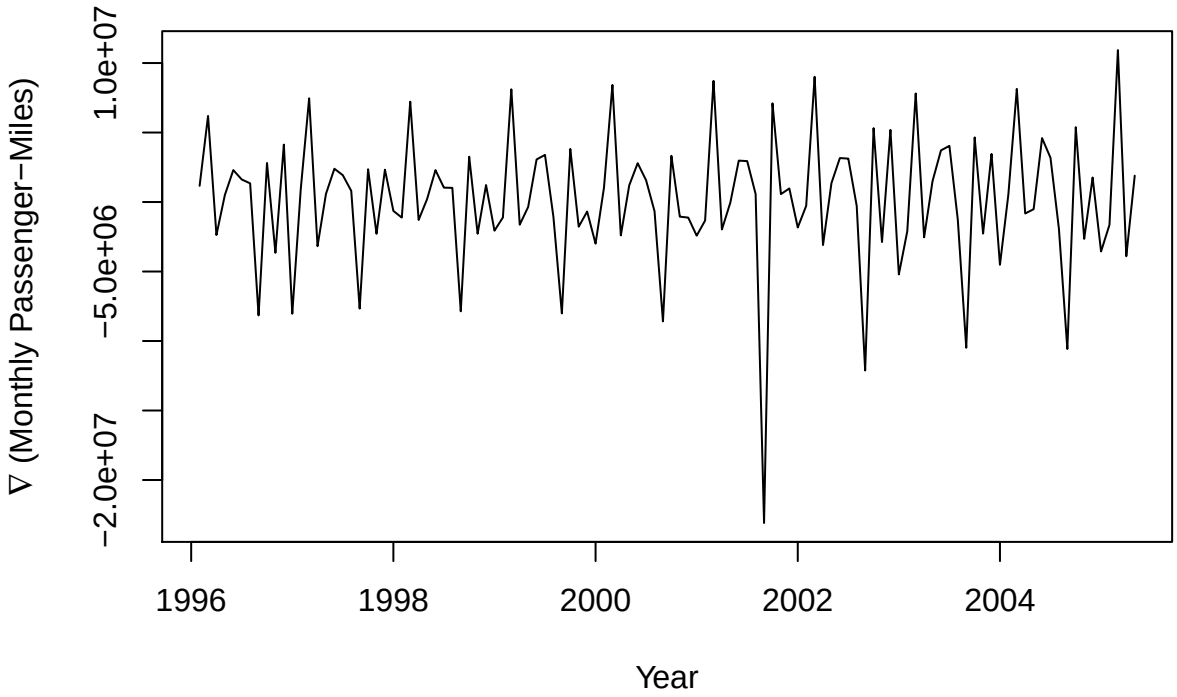


This data does not appear to come from a stationary process. There is a non-constant (possibly linear) mean. There may also be some indication of unequal variances, and perhaps even seasonality.

b) Plot the first difference of the time series. What improvements do you see here? Is there still something in this data that needs to be accounted for?

Solution:

```
plot(diff(airmiles), xlab='Year',
         ylab=expression(paste(nabla,' (Monthly Passenger-Miles)')))
```



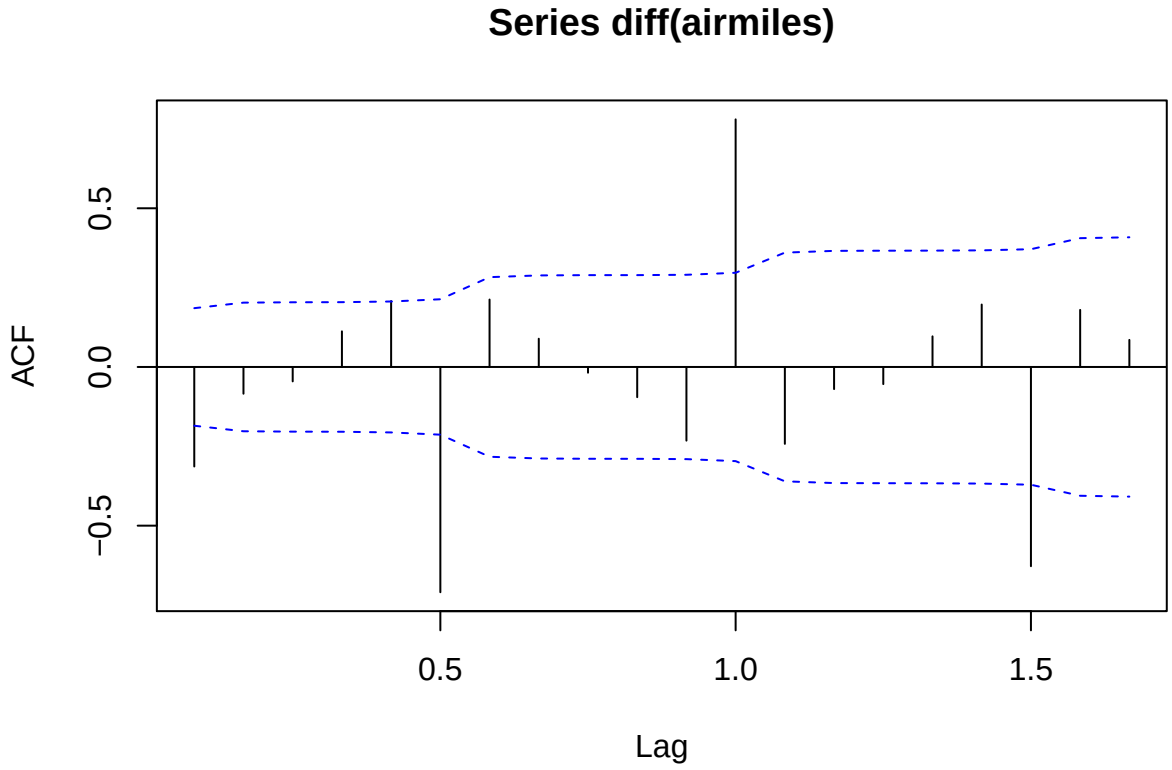
We see that the mean now appears to be constant.

However, we still see seasonality in this data. There is also a potential outlier.

c) Create the sample ACF plot, sample PACF plot and sample EACF table for the *first difference* of the time series. Explain any conclusions you can make from each of these visualizations. If you find that it is difficult to make a single overall conclusion about the underlying model, explain why it is difficult.

Solution:

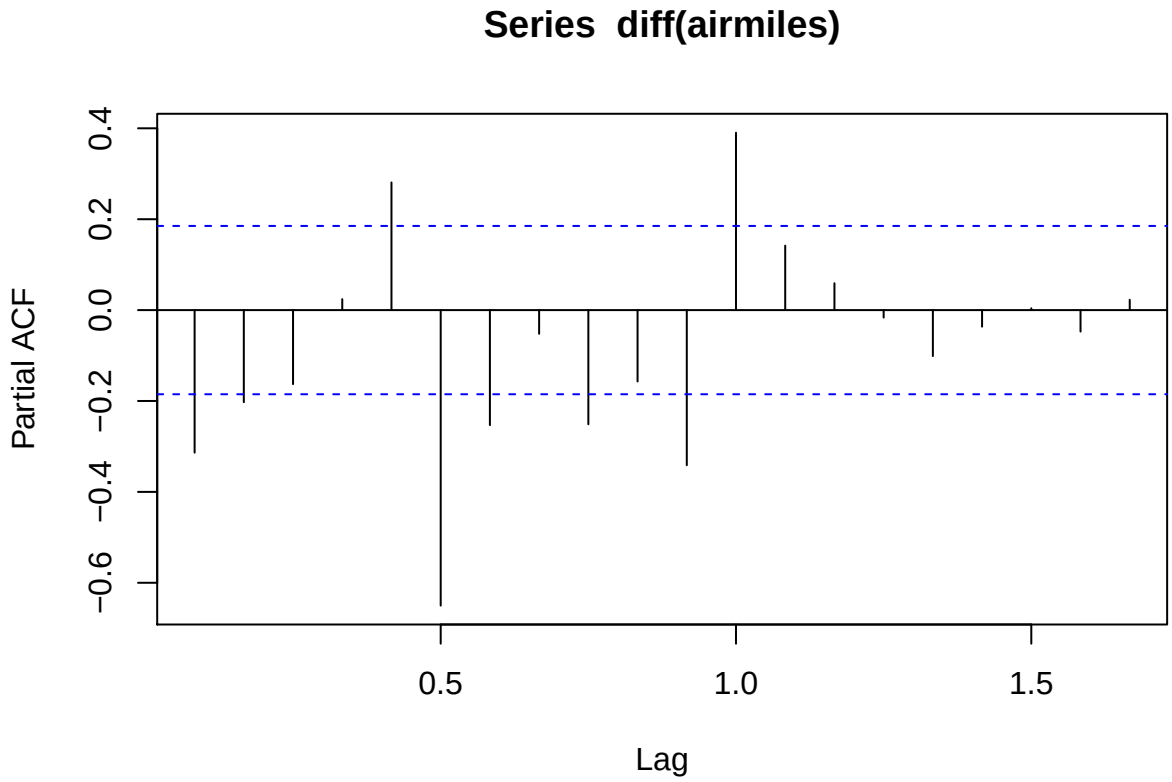
```
acf(diff(airmiles), ci.type='ma')
```



The sample ACF plot of the first differences is not very clear. It appears to indicate that this is perhaps not an MA model, since the values don't "cut off" anywhere.

We also notice that there are very high autocorrelations at "lags" of 0.5, 1.0 and 1.5. Since this dataset is coded such that 1 time unit = 1 year (i.e., 12 months), these results are telling us that there are high correlations between first differences that are 6, 12 or 18 months apart.

```
pacf(diff(airmiles))
```



The sample PACF plot of the first differences is also not very clear. It appears to indicate that this is perhaps not an AR model, since the values don't "cut off" anywhere.

We again see high partial autocorrelations at "lags" of 0.5 and 1.0, indicating to us that (controlling for observations in between) there are high correlations between first differences that are 6 or 12 months apart.

```
eacf(diff(airmiles))
```

```
## AR/MA
##   0 1 2 3 4 5 6 7 8 9 10 11 12 13
## 0 x o o o x x x o o o x  x  x  o
## 1 x o o o o x x o o o o  x  x  o
## 2 x o o o o x x o o o o  x  x  o
## 3 o x o o o x x o o o o  x  o  o
## 4 o x o o o x o o o o o  x  o  o
## 5 x o o x x x x o o o x  x  x  o
## 6 x o x o x x o o x o o  x  o  o
## 7 o o x o o x o o x o o  x  o  o
```

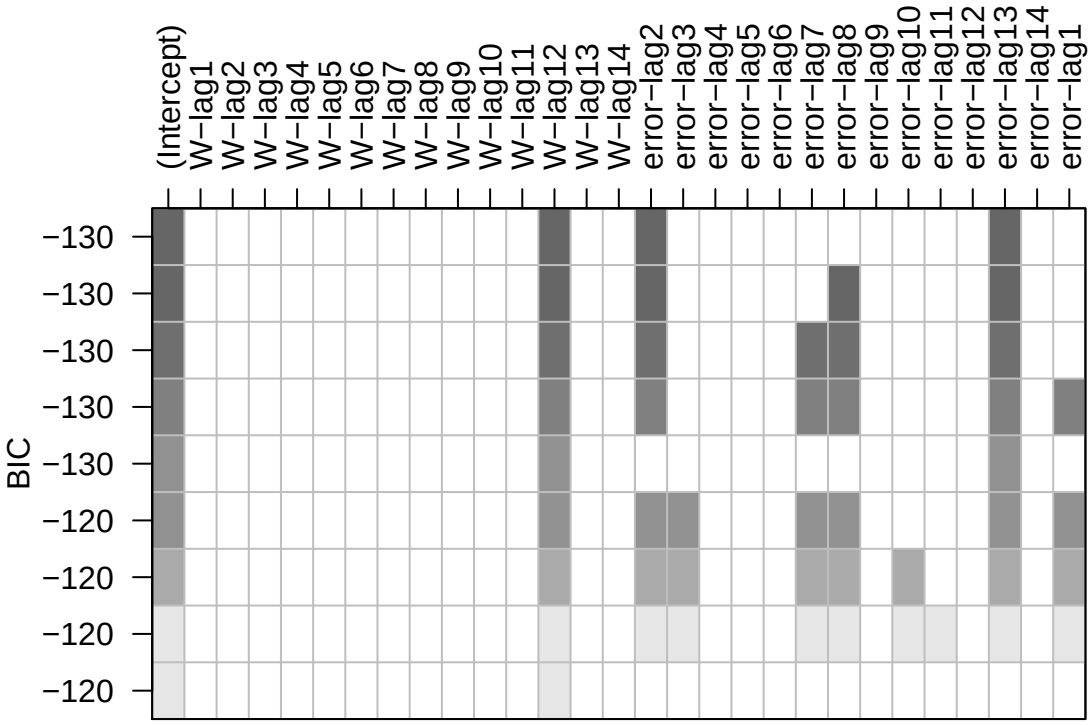
The EACF table is also difficult to read. There may be some evidence of an MA(1), MA(7) or AR(3) model, but the evidence is quite poor.

Overall, it is difficult to make an overall conclusion, since none of the visualizations show clear results.

d) Sometimes this type of data is seasonal. So, perhaps the difficulty in choosing an appropriate ARMA model is due to the fact that this is a subset-ARMA model, i.e. some of the coefficients are zero. Create the "best subset-ARMA selection" plot.
(Hint: An example of how to create this plot is shown in Video 28. You can ignore the Warning message that R gives you.)

Solution:

```
armasubsets.select <- armasubsets(y=diff(airmiles), nar=14, nma=14, y.name='W')
plot(armasubsets.select)
```



e) Interpret the best subset-ARMA selection plot you created in part (d). What does it generally tell you about the "most important" terms in the model?

Solution:

We see that, for the "best" model (top row) and for almost all of the other "very good" models, the most important terms appear to be: Intercept, W_{t-12}, e_{t-2} and e_{t-13}.

(Note: In our notation here, $W_t = \nabla Y_t$.)

f) Write out the top choice of model based on the best subset-ARMA selection plot in part (d). *(Hint: Remember that this ARMA model is for the first difference of the series, $\{W_t\}$, not for the original series $\{Y_t\}$.)*

Solution:

Based on the above results, it appears that the best choice of model is:

$$W_t = \theta_0 + \phi_{12} W_{t-12} + e_t - \theta_2 e_{t-2} - \theta_{13} e_{t-13}$$

g) Does this choice of model make sense to you? Why do you think these lags were chosen to be the most important?

Solution:

This model generally does make sense, due to the fact that it includes the term W_{t-12}. In our plot of the first difference series in part (b), we saw that there was clear seasonality in the series. We also saw this again in the sample ACF and sample PACF plots of part (c).

The reason for the inclusion of e_{t-2} and e_{t-13} is less obvious from the plots.

h) Since this ARMA model was constructed for the *first difference* of the time series, write out the equation for the original series, $\{Y_t\}$.

Solution:

From our equation for the model for $\{W_t\}$ in part (f), and the fact that $W_t = Y_t - Y_{t-1}$ for any t, we obtain the following:

$$Y_t - Y_{t-1} = \theta_0 + \phi_{12}(Y_{t-12} - Y_{t-13}) + e_t - \theta_2 e_{t-2} - \theta_{13} e_{t-13}$$

$$Y_t = \theta_0 + Y_{t-1} + \phi_{12} Y_{t-12} - \phi_{12} Y_{t-13} + e_t - \theta_2 e_{t-2} - \theta_{13} e_{t-13}$$

2. Describe the importance of the "$2k$" term in Akaike's Information Criterion (AIC). What does it achieve? What would happen if we were to not include this term in the criterion?
(This answer only needs to be about 2 sentences long.)

Solution:

The $2k$ term in AIC penalizes models that have too many terms. Since the goal is to have a low value of AIC, a model with a large number of parameters k will struggle to have a low AIC unless the terms are all deemed to be very "useful".

If we were to not include this term in the AIC criterion, we would be inclined to choosing models that are too large (i.e. have too many parameters). Large models will often appear to fit our data very well, but they may be simply "chasing the errors" in our data rather than effectively describing the true underlying behaviour. Therefore, it is desirable to prevent this from happening, by including the "$2k$" penalty.