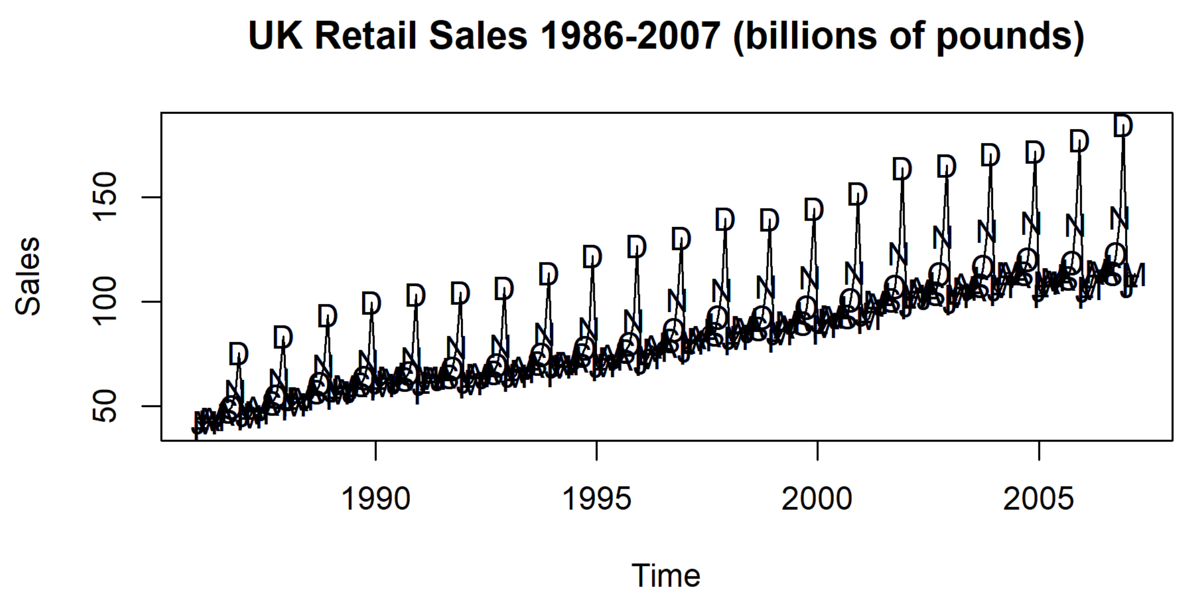
Mason Galvin MTHSTAT 764 Midterm

1) I would expect that over time, retail sales will increase. This is due to economic growth causing people to have more money to spend, inflation causing prices to rise, and population growth increasing demand for retail products. I would also expect there to be seasonality present, due to increased sales during the Christmas season.



As we can see, there is a general upward trend and a very clear seasonal trend, with a massive spike in sales in December, right around the holiday season. This lines up well with my prior assumptions.

2)
I will use a linear regression model to try and capture the trend
The summary of the model is as follows

```
Call:
lm(formula = retail ~ ts)

Residuals:
    Min      1Q  Median      3Q     Max
-19.165  -7.587  -4.675  -0.438  59.937

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

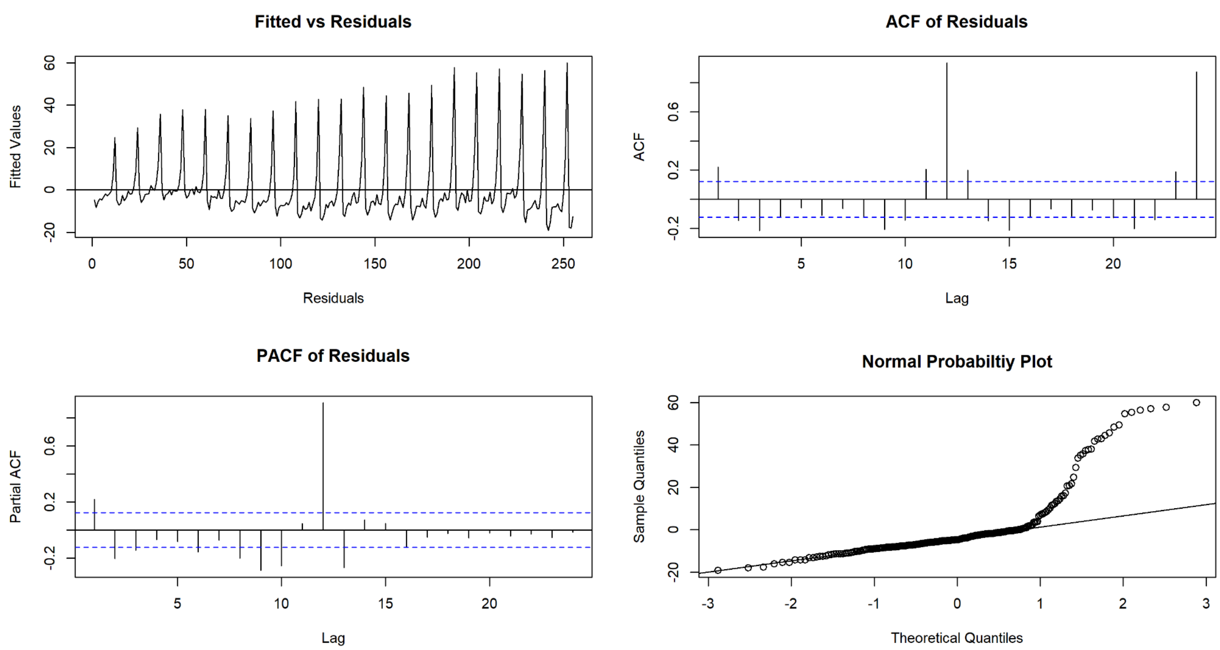
(Intercept) 46.80425   1.89246   24.73   <2e-16 ***
ts   0.30976   0.01282   24.17   <2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.07 on 253 degrees of freedom
Multiple R-squared: 0.6978, Adjusted R-squared: 0.6966
F-statistic: 584.1 on 1 and 253 DF, p-value: < 2.2e-16

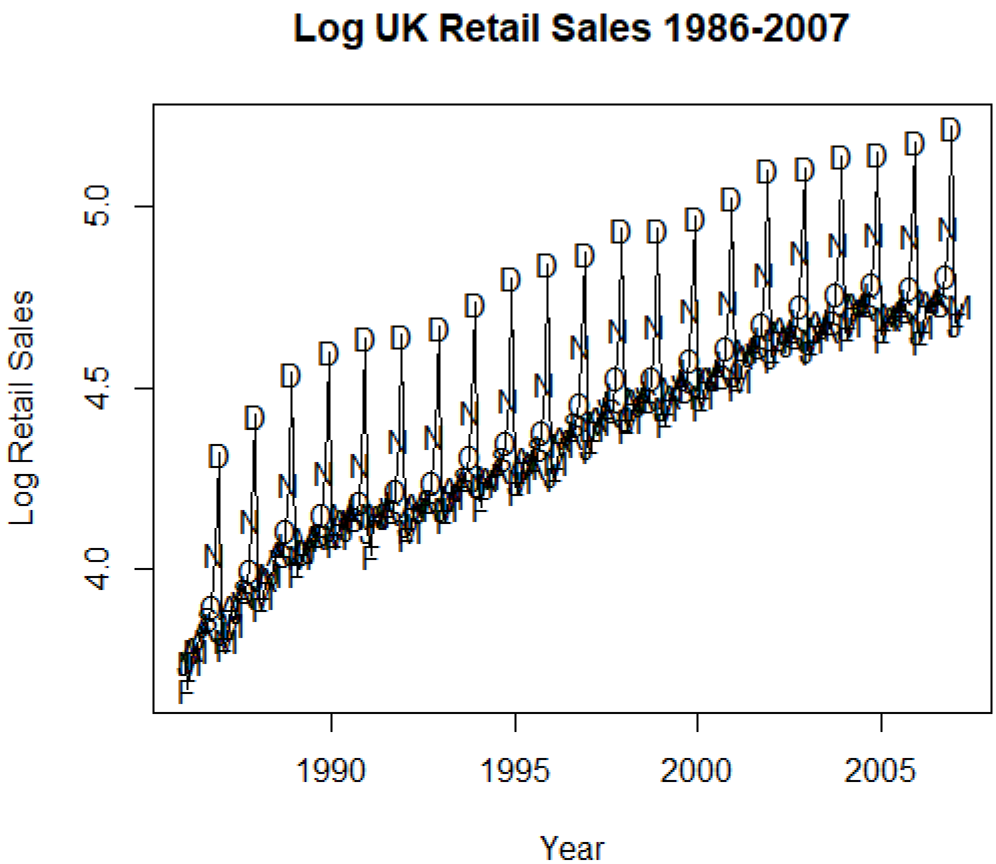
Our coefficient is .30976, and it has a p value of less than 2*10^(-16) so it is significant at all confidence levels. Our model has a F statistic value of 584.1 So it is also significant at all confidence levels. However, the R squared and adjusted R squared value are rather low, at .6978 and .6966 which means there is a significant amount of variation in the data that is not explained by the model. Looking at the residuals will give us more insight into this

3)

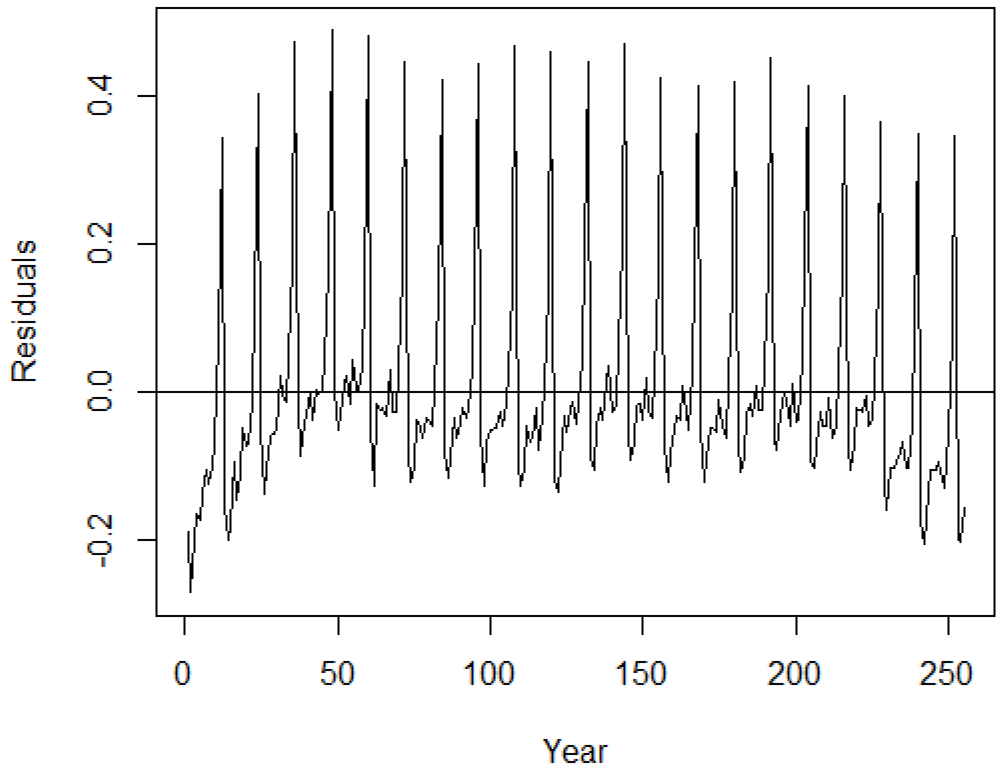


As we can see from the residual plots that the linear model is not a great model for the data. The Fitted vs Residuals plot shows clear patterns, with large, cyclical spikes in the data and very uneven spread. The ACF plot shows significant autocorrelation at lags 1 and 12, showing that the seasonal trend is present in the ACF. Otherwise, the autocorrelation looks negligible. The PACF shows a similar spike at lag 12, again supporting that the seasonal trend is still present in the data. The normal probability plot does not look linear, with a sharp climb and curved upper tail on the right hand side, which challenges the normality assumption. We will need to transform the data in order to find a better model. I will try a seasonal first difference transformation of the logs

3) Plotting the log transformed Retail time series



The variance seems to stabilize over time, however if we try just using a simple logarithmic transformation the residual plot looks like this:



The residuals still show clear patterns and uneven spread. Thus further transforming the model using seasonal differencing produces a better model.

The model for the seasonal first difference is as follows:

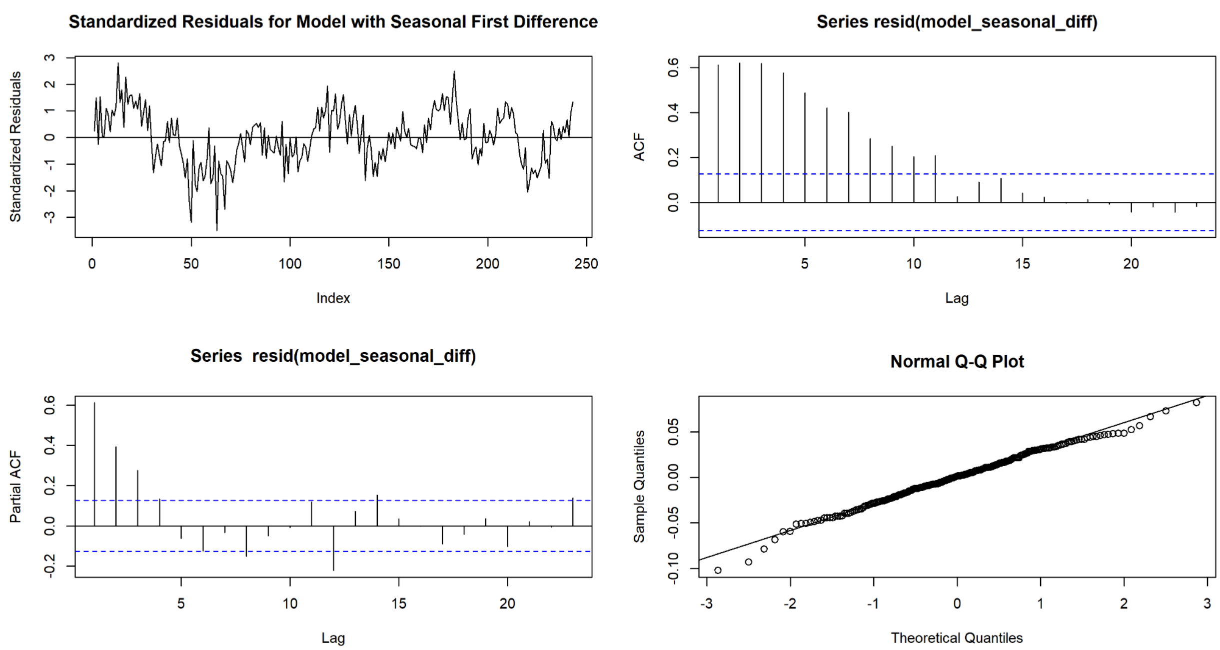
Call:

```
lm(formula = retail_seasonal_diff ~ time(retail_seasonal_diff))

Residuals:
      Min        1Q    Median        3Q       Max
-0.102062 -0.018825  0.001083  0.020998  0.081874

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 5.2175315  0.6408375   8.142 2.09e-14 ***
time(retail_seasonal_diff) -0.0025896  0.0003209  -8.070 3.33e-14 ***
---
```

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

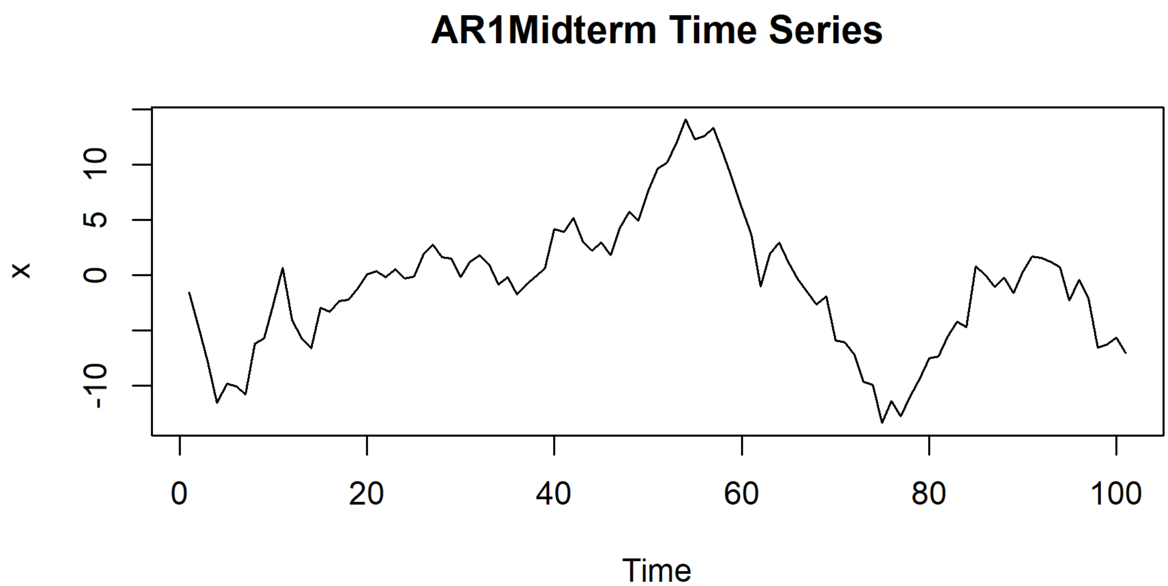
Residual standard error: 0.02924 on 241 degrees of freedom
Multiple R-squared: 0.2127, Adjusted R-squared: 0.2095
F-statistic: 65.13 on 1 and 241 DF, p-value: 3.328e-14

The model has a coefficient of -.0025896 with a p value of 3.33*10^-14, it is significant at all confidence levels. The model has an F statistic value of 65.13 with a p value of 3.32*10^-14, meaning that it is also significant at all levels. It does however have lower R squared and adjusted R squared values than the simple linear model. However, looking at the residuals:

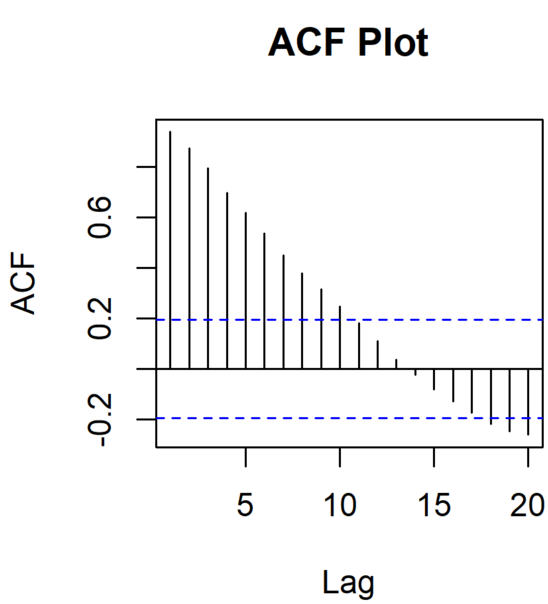


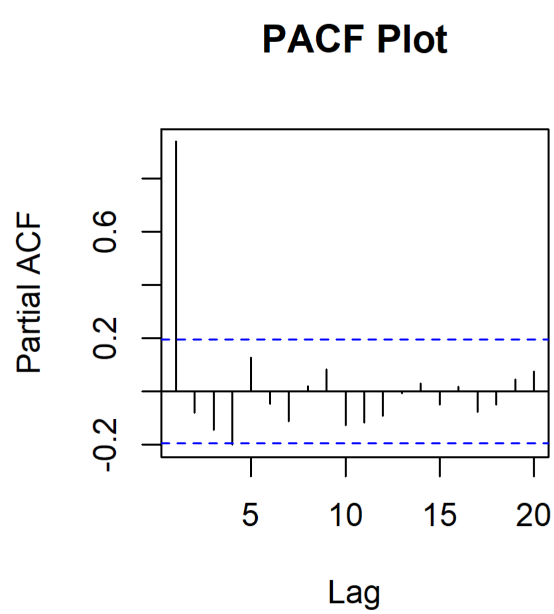
The residual plot shows random movement about 0, with no clear patterns. There does not seem to be many outliers, as all the residuals look to be between -3 and 3. The variance seems to be fairly stable as time goes on. The ACF plot shows autocorrelation decaying over time and centering around 0, indicating that this model has captured the time structure of the data and that autocorrelation does not appear to be present. The PACF graph backs up this assumption by also decaying to around zero and then showing random movement about 0 at each lag. The Normal Probability plot is very linear save for some curling at the tails, which suggests this model follows the normality assumption. From the Normal QQ plot there does appear to be at least two and maybe three outliers. Overall, the log transformed seasonal first difference seems to be a good fit for the retail sales data.

Part 2:

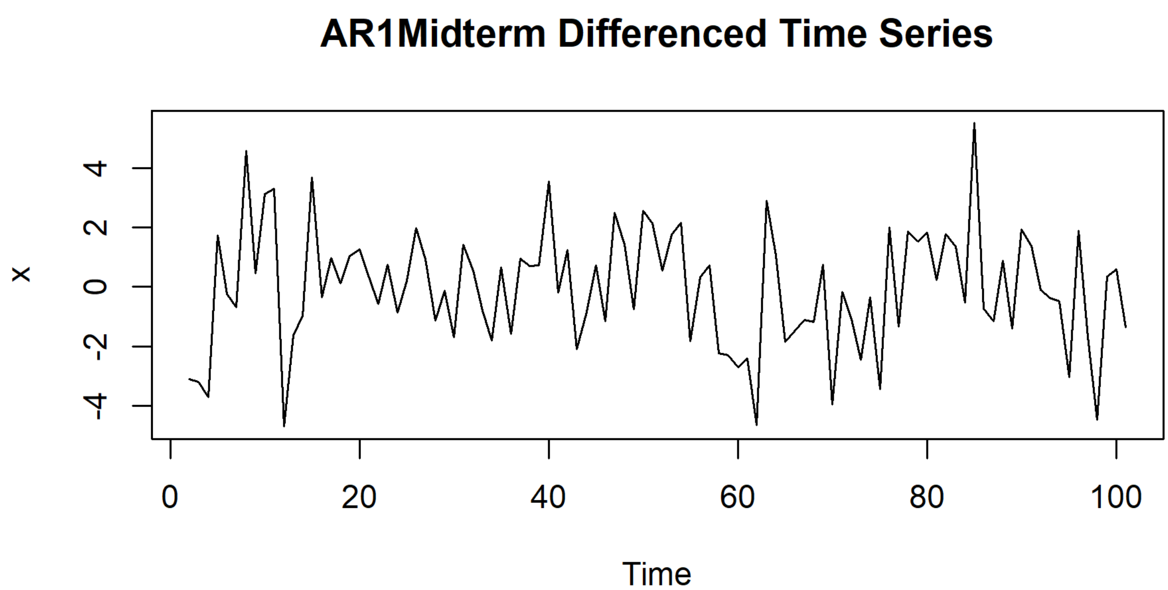


It is hard to determine from this graph alone whether or not the series is stationary. It seems to stay around 0 over time but the variance also seems to increase as well. Looking at the ACF and PACF graphs will give us better insight.

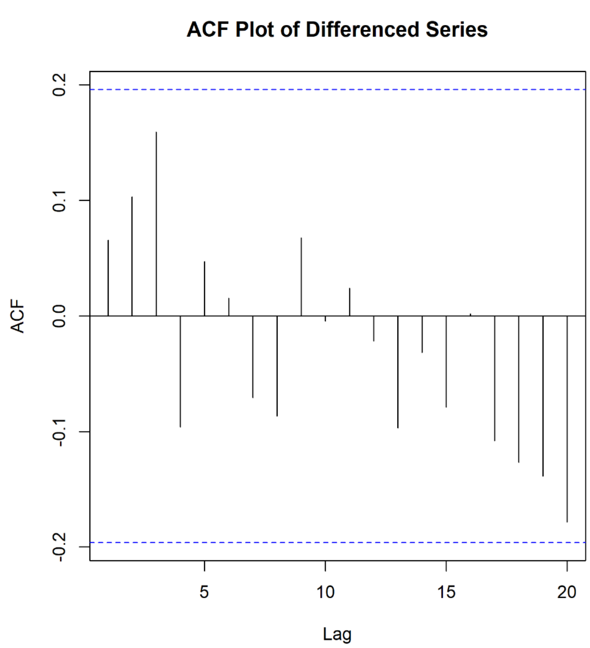


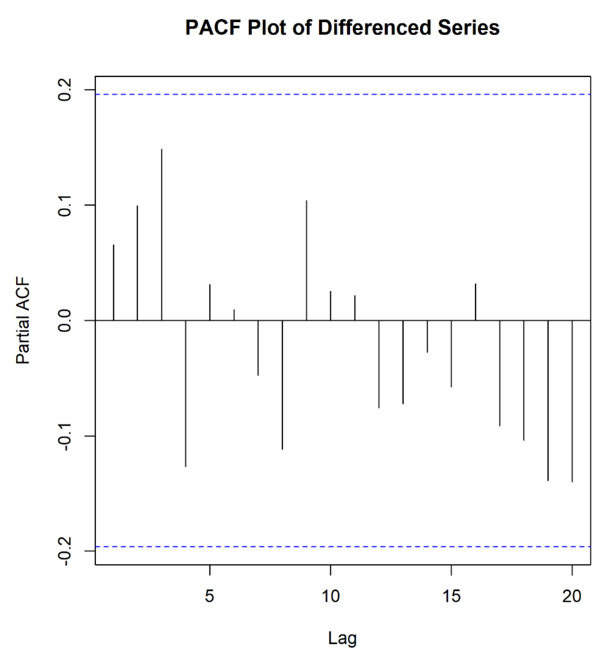


As we can see here it is much more clear: the autocorrelation constantly decreases as time goes on and thus does depend on time, so the series is not stationary. Let us again use differencing to correct it. This time I will use simple first differencing.



The series now seems to be stationary, staying around 0 with a more constant variance. Looking at the ACF and PACF plots:

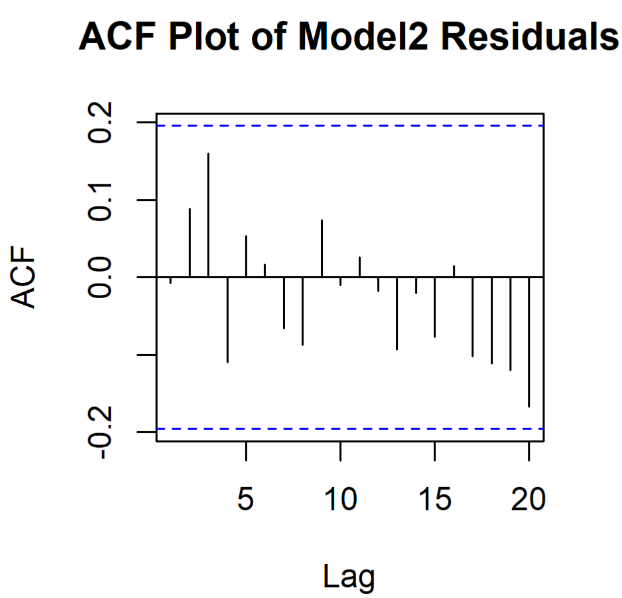


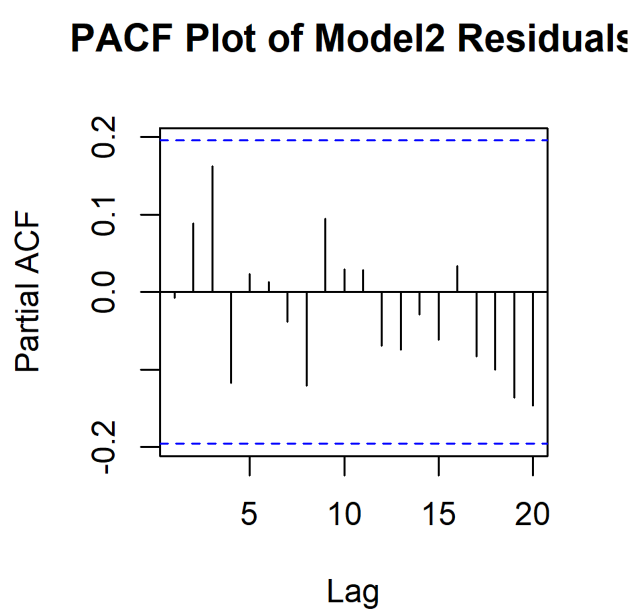


They are both very small (within .2 of 0) as to be negligible and they do not seem to be changing over time significantly. Differencing seems to have corrected non-stationarity. We can fit a differenced model to an AR(1) model: the model is as follows:

```
summary(model2)
          Length Class  Mode
coef        2    -none- numeric
sigma2      1    -none- numeric
var.coef    4    -none- numeric
mask        2    -none- logical
loglik      1    -none- numeric
aic         1    -none- numeric
arma        7    -none- numeric
residuals 100    ts     numeric
call        3    -none- call
series      1    -none- character
code        1    -none- numeric
n.cond      1    -none- numeric
nobs        1    -none- numeric
model      10    -none- list
```

Looking at the ACF and PACF of the residuals:





We can see that there is no significant autocorrelation of this model, meaning that the AR(1) model is a good fit for the data.