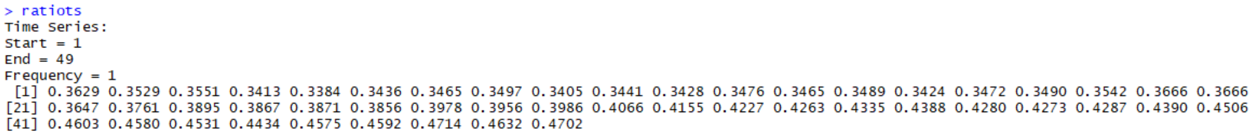
Dev Gupta

2017B3A71082

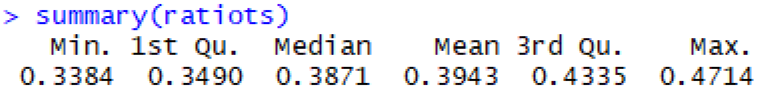
The Data

The data for this assignment has been taken from the World Inequality Database. The data is the income share of the Top 10% in the USA for the years 1966 – 2014. Initially, the dataset chosen had data from 1913 but the values for 1963 and 1965 Ire missing. Hence only the data post 1965 is being used.

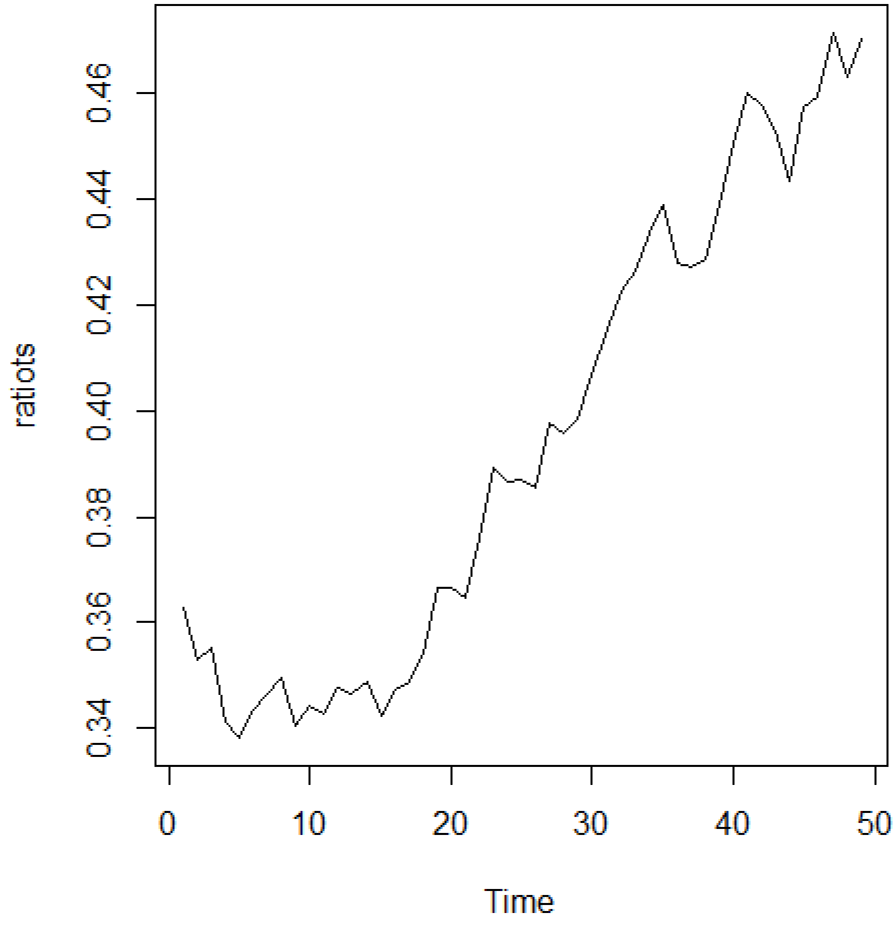
Some exploratory analysis of the data,



```
> ratiots
Time Series:
Start = 1
End = 49
Frequency = 1
 [1] 0.3629 0.3529 0.3551 0.3413 0.3384 0.3436 0.3465 0.3497 0.3405 0.3441 0.3428 0.3476 0.3465 0.3489 0.3424 0.3472 0.3490 0.3542 0.3666 0.3666
[21] 0.3647 0.3761 0.3895 0.3867 0.3871 0.3856 0.3978 0.3956 0.3986 0.4066 0.4155 0.4227 0.4263 0.4335 0.4388 0.4280 0.4273 0.4287 0.4390 0.4506
[41] 0.4603 0.4580 0.4531 0.4434 0.4575 0.4592 0.4714 0.4632 0.4702
```

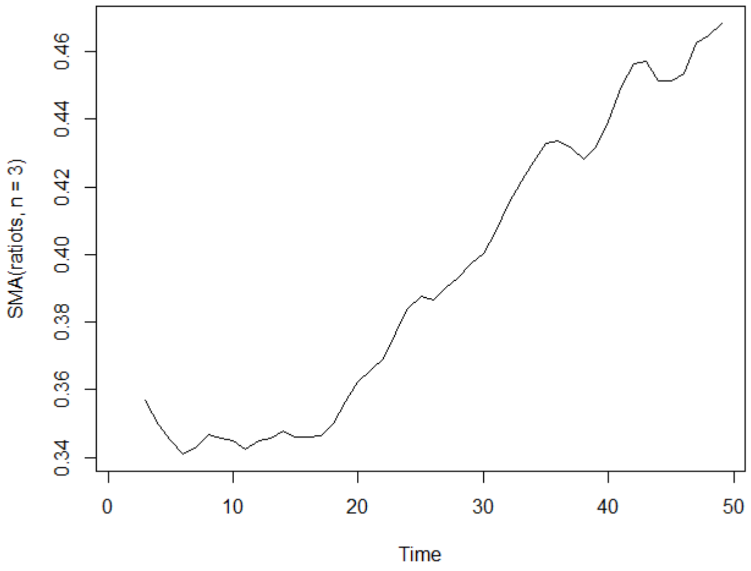


```
> summary(ratiots)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.3384  0.3490  0.3871  0.3943  0.4335  0.4714 
```

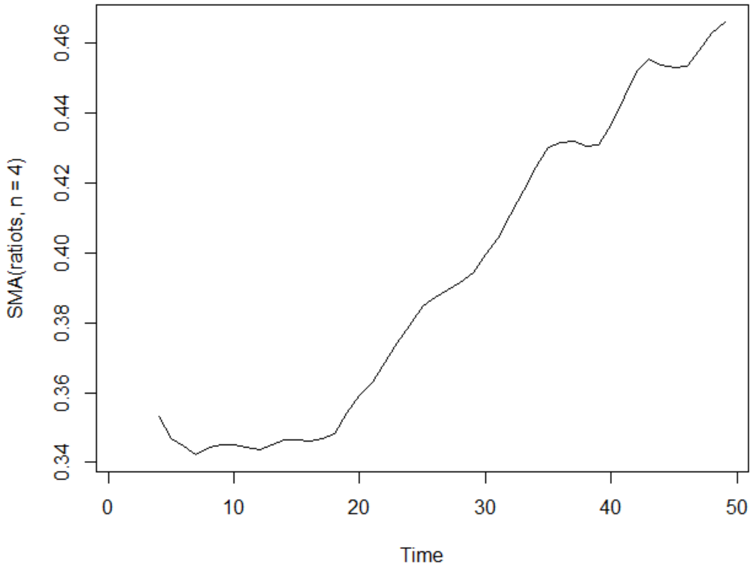


Q1. Decomposing the series

This data is collected on a yearly basis thus there is no season component. So, there is only trend and irregular component. Since, the graph doesn't suggest a multiplicative model, I will assume it to be additive and proceed. To get the trend component I will use a moving average. First, I try a moving average of 3,

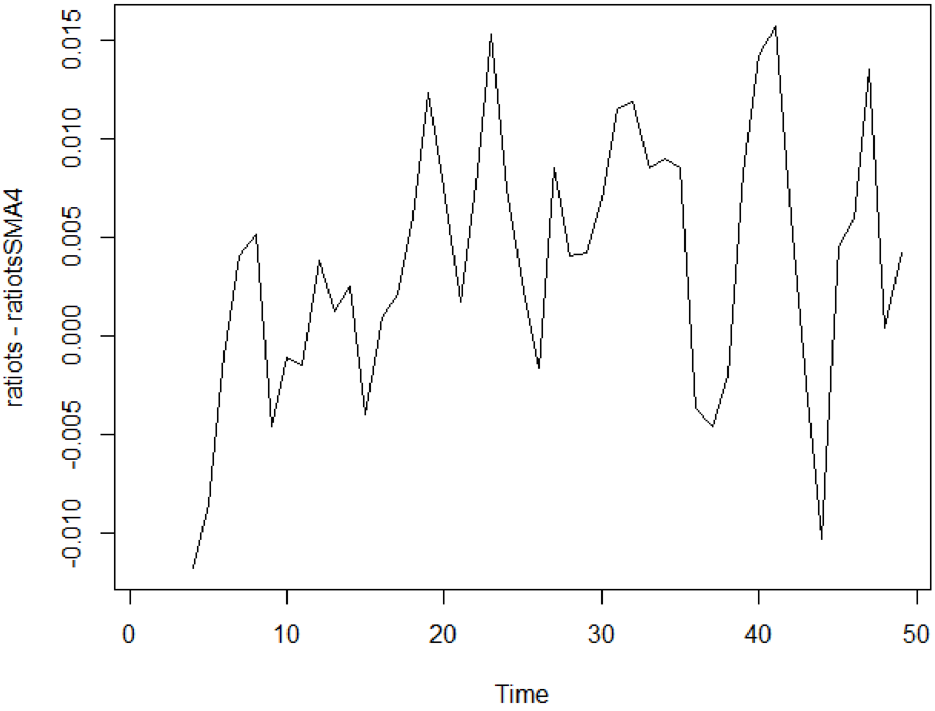


There seem to be some irregularities still present so I will try a moving average of 4,



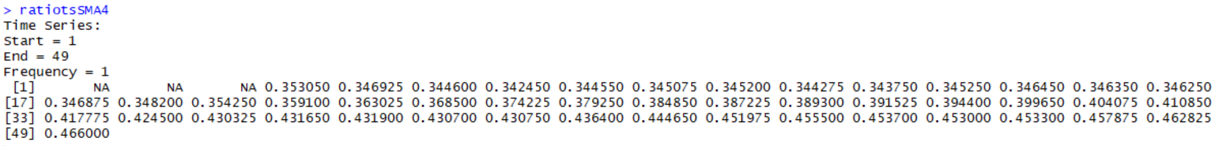
This plot looks sufficiently smooth

Now to check our assumption of additive series, I plot the irregular component obtained by subtracting the smooth series (i.e. the trend) from the original series,



There is no indication that the irregular components depend on the time period, thus, I can reject the multiplicative model.

From the above, it can be concluded that the trend is relatively stable at about 0.35 for the first 20 or so observations and then rapidly rises to end at 0.46.



```
> ratiotsSMA4
Time Series:
Start = 1
End = 49
Frequency = 1
 [1]       NA       NA       NA 0.353050 0.346925 0.344600 0.342450 0.344550 0.345075 0.345200 0.344275 0.343750 0.345250 0.346450 0.346350 0.346250
[17] 0.346875 0.348200 0.354250 0.359100 0.363025 0.368500 0.374225 0.379250 0.384850 0.387225 0.389300 0.391525 0.394400 0.399650 0.404075 0.410850
[33] 0.417775 0.424500 0.430325 0.431650 0.431900 0.430700 0.430750 0.436400 0.444650 0.451975 0.455500 0.453700 0.453000 0.453300 0.457875 0.462825
[49] 0.466000
```

Q2. ARIMA Models

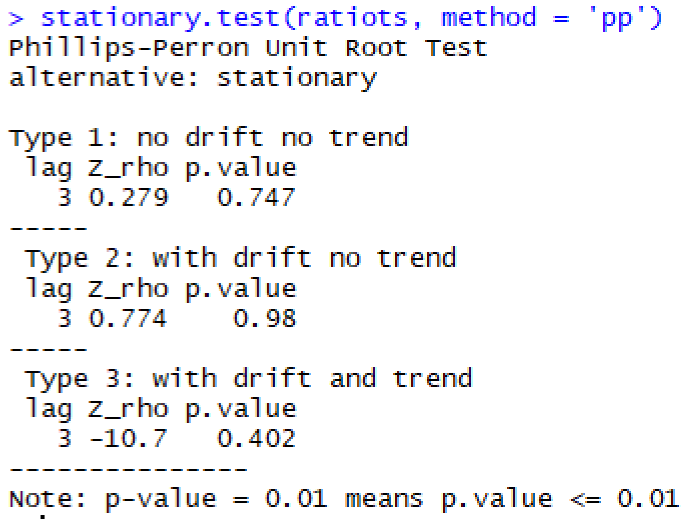
A) Check for stationarity

Graphically,

I see previously the series appears to be not stationary as the values increase rapidly after the first 20 years.

By tests,

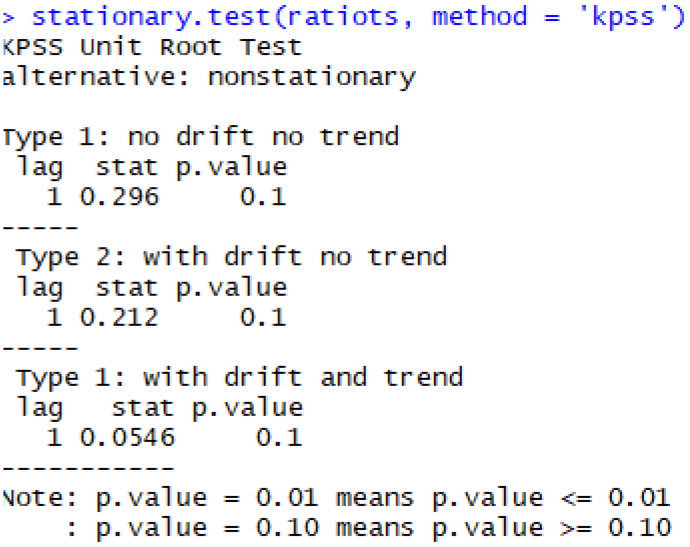
Phillips-Perron test,



```
> stationary.test(ratiots, method = 'pp')
Phillips-Perron Unit Root Test
alternative: stationary

Type 1: no drift no trend
 lag Z_rho p.value
   3 0.279   0.747
-----
 Type 2: with drift no trend
 lag Z_rho p.value
   3 0.774    0.98
-----
 Type 3: with drift and trend
 lag Z_rho p.value
   3 -10.7   0.402
---------------
Note: p-value = 0.01 means p.value <= 0.01
```

I fail to reject the null, that series has a unit root, because of the high p-value

KPSS test,

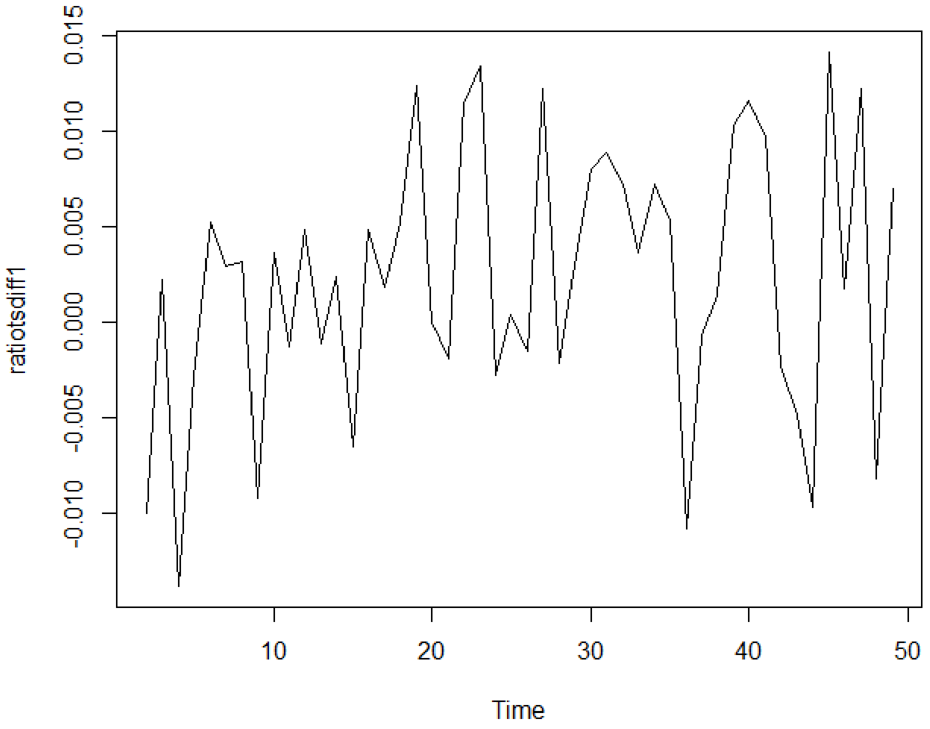


```
> stationary.test(ratiots, method = 'kpss')
KPSS Unit Root Test
alternative: nonstationary

Type 1: no drift no trend
 lag  stat p.value
   1 0.296     0.1
-----
 Type 2: with drift no trend
 lag  stat p.value
   1 0.212     0.1
-----
 Type 1: with drift and trend
 lag   stat p.value
   1 0.0546     0.1
-----------
Note: p.value = 0.01 means p.value <= 0.01
    : p.value = 0.10 means p.value >= 0.10
```

Although, the p-value is too high to reject null at 95% or 99% confidence-level, I can reject null at the 90%

Thus, I can conclude that the series is not stationary. To rectify this, I will use the first difference of the series.

Graphically,



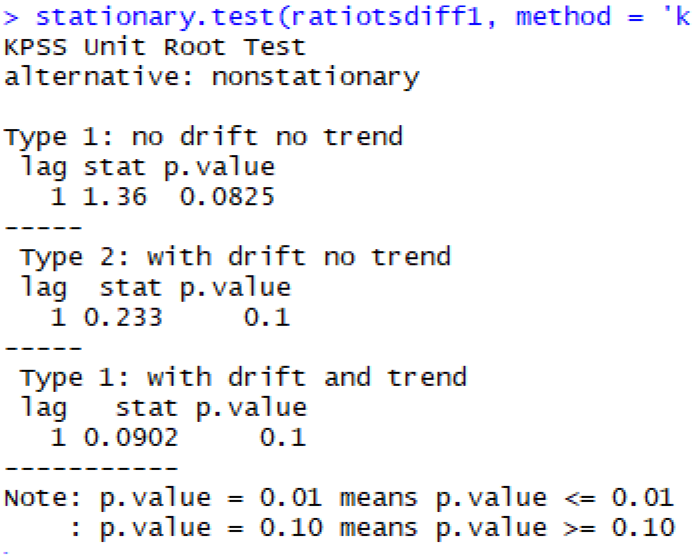
The series looks fairly stationary.

By tests,

Phillips-Perron test,

```
> ratiotsdiff1 = diff(ratiots, differences = 1)
> stationary.test(ratiotsdiff1, method = 'pp')
Phillips-Perron Unit Root Test
alternative: stationary

Type 1: no drift no trend
 lag Z_rho p.value
   3 -43.2    0.01
-----
 Type 2: with drift no trend
 lag Z_rho p.value
   3   -43    0.01
-----
 Type 3: with drift and trend
 lag Z_rho p.value
   3 -43.9    0.01
---------------
Note: p-value = 0.01 means p.value <= 0.01
>
```

KPSS test,



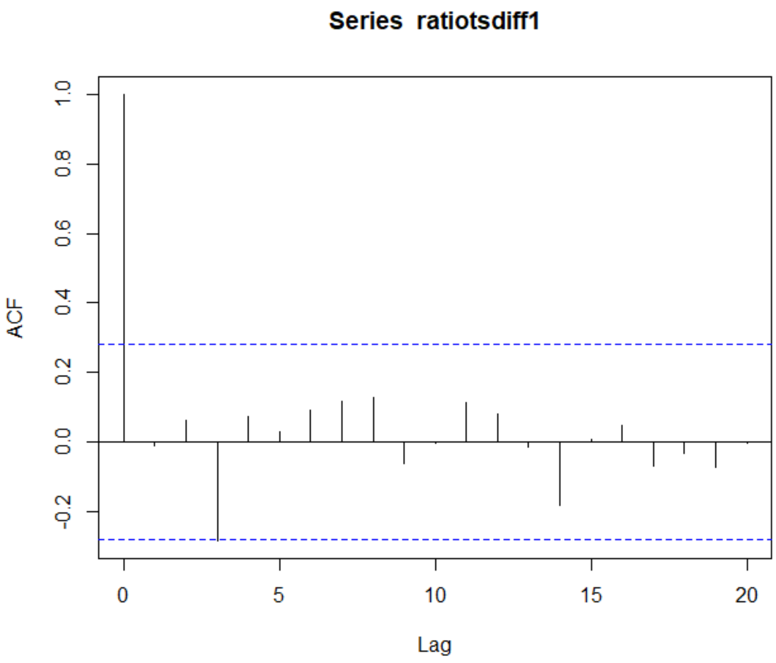
```
> stationary.test(ratiotsdiff1, method = 'kpss')
KPSS Unit Root Test
alternative: nonstationary

Type 1: no drift no trend
 lag stat p.value
   1 1.36  0.0825
-----
 Type 2: with drift no trend
 lag  stat p.value
   1 0.233     0.1
-----
 Type 1: with drift and trend
 lag   stat p.value
   1 0.0902     0.1
-----------
Note: p.value = 0.01 means p.value <= 0.01
    : p.value = 0.10 means p.value >= 0.10
```

Because of the very low p-value of the PP test and a sufficiently high p-value in the KPSS test, I conclude that the first difference in stationary.

B) Proper ARIMA Model

Using ACF and PCF graphs,

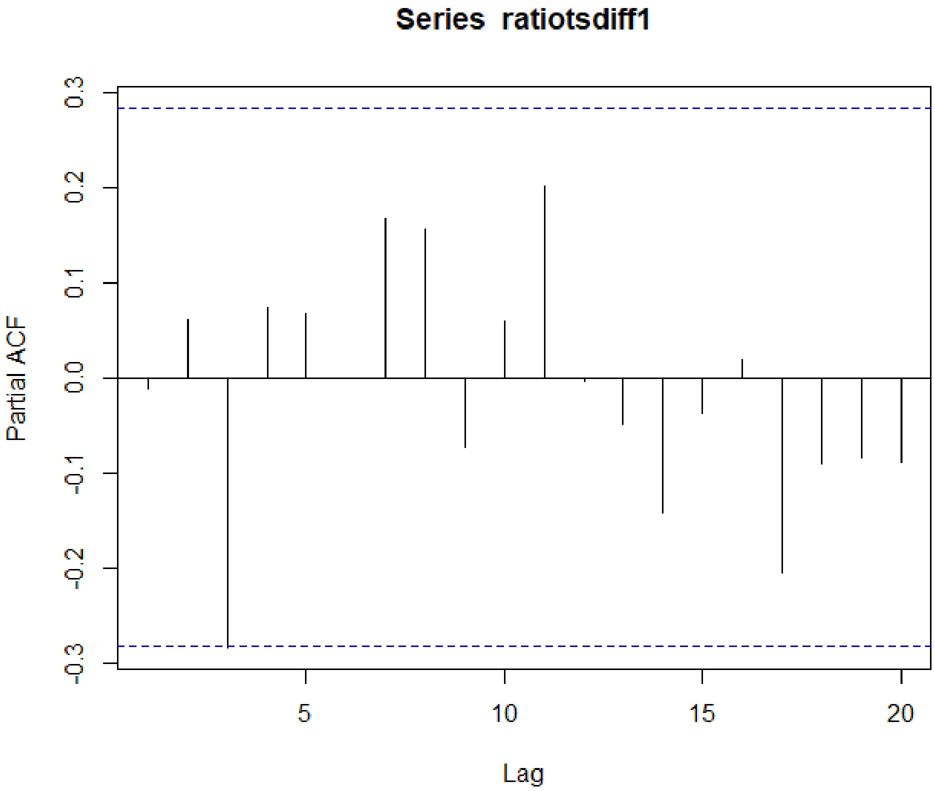
ACF,



s

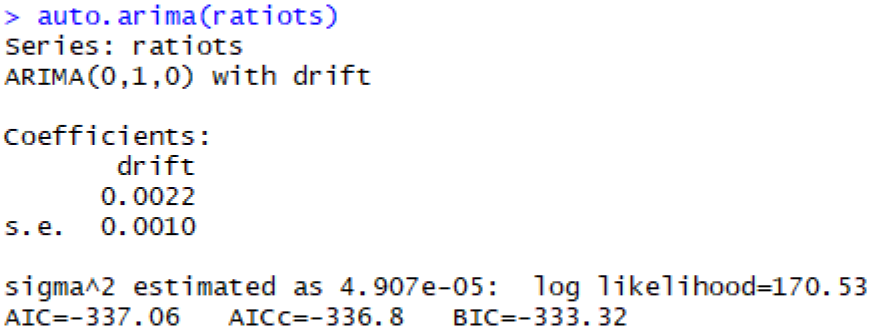
I don’t find any significant peaks starting with Lag1

PACF,



Again, only the third lag exceeds the significance threshold. Thus, I can't use an AR or MA model. This suggests that this is a white noise model

Using auto.arima,



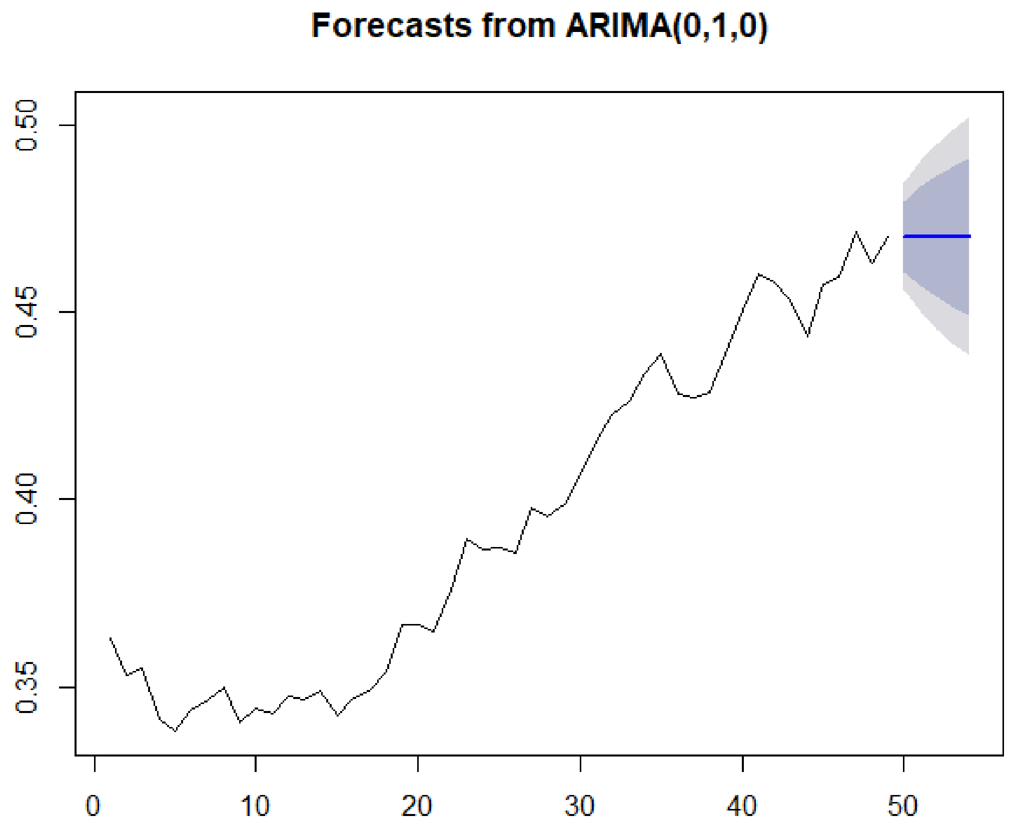
```
> auto.arima(ratiots)
Series: ratiots
ARIMA(0,1,0) with drift

Coefficients:
       drift
      0.0022
s.e.  0.0010

sigma^2 estimated as 4.907e-05:  log likelihood=170.53
AIC=-337.06   AICc=-336.8   BIC=-333.32
```

Here also, I get the same result.

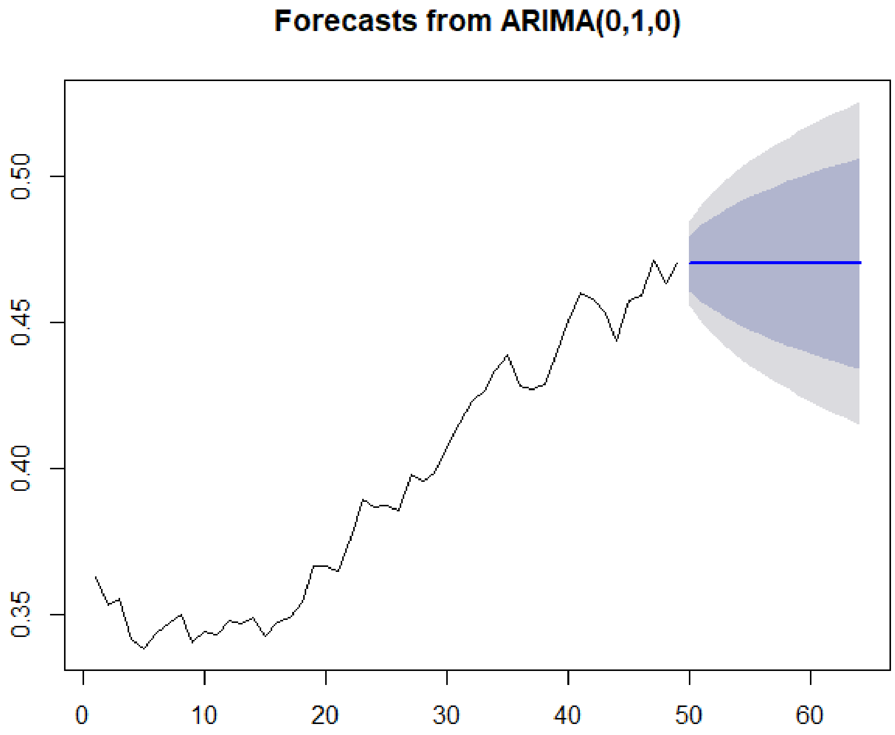
Thus, a ARIMA(0.1.0) on the original or ARIMA(0,0,0) on the first difference series is appropriate.

C) Forecast

```
> ratiotsarima = arima(ratiots, order = c(0,1,0))
> ratiotsarima

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


sigma^2 estimated as 5.304e-05:  log likelihood = 168.16,  aic = -334.31
> ratiotsforecast = forecast(ratiotsarima, h = 5)
> ratiotsforecast
   Point Forecast     Lo 80     Hi 80     Lo 95     Hi 95
50         0.4702 0.4608665 0.4795335 0.4559256 0.4844744
51         0.4702 0.4570004 0.4833996 0.4500130 0.4903870
52         0.4702 0.4540339 0.4863661 0.4454760 0.4949240
53         0.4702 0.4515330 0.4888670 0.4416512 0.4987488
54         0.4702 0.4493296 0.4910704 0.4382815 0.5021185
> plot(ratiotsforecast)
```



```
> ratiotsforecast = forecast(ratiotsarima, h = 15)
> plot(ratiotsforecast)
```



Towards the end of observations in the original series I noticed a trend of rapid increase. Holver, the forecast doesn't allow for that rapid an increase for the next 15 years even at the 95% level. Also, I notice that the point forecast for all the years in same.

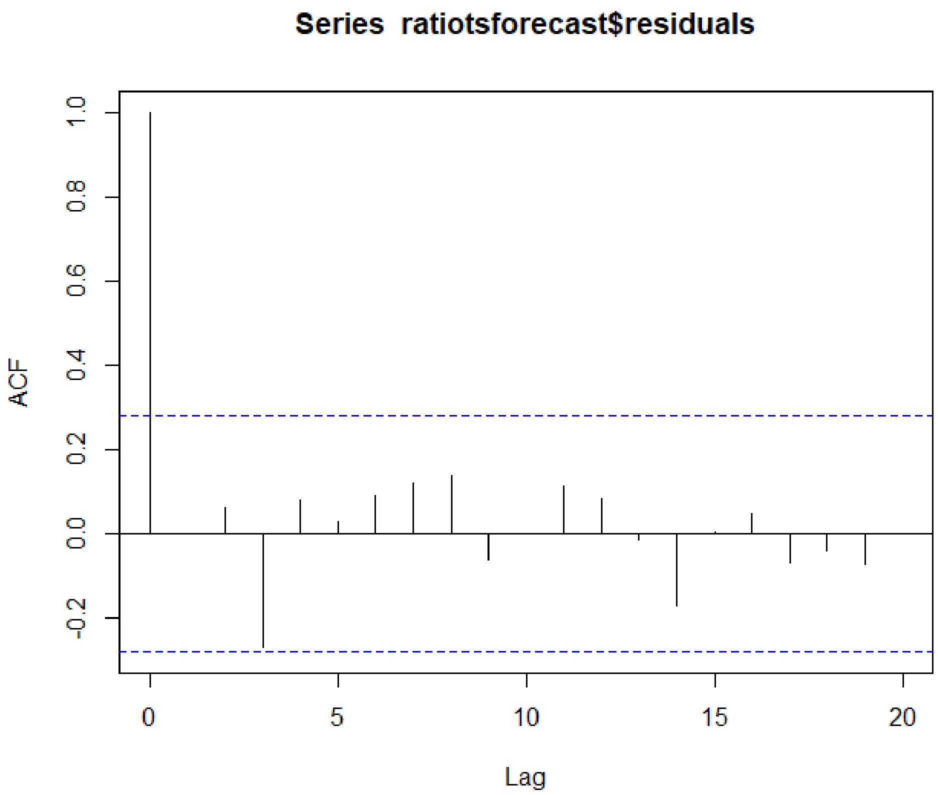
D) Testing for forecast assumptions

I test two things

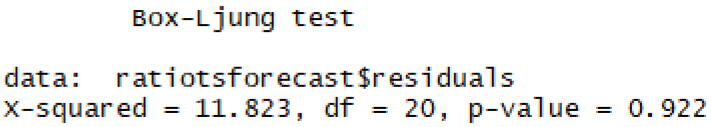
1) Correlation of successive forecast errors

2) If forecasts errors are normally distributed with zero mean and constant variance.

To test the correlation I use the ACF and the Ljung-Box test



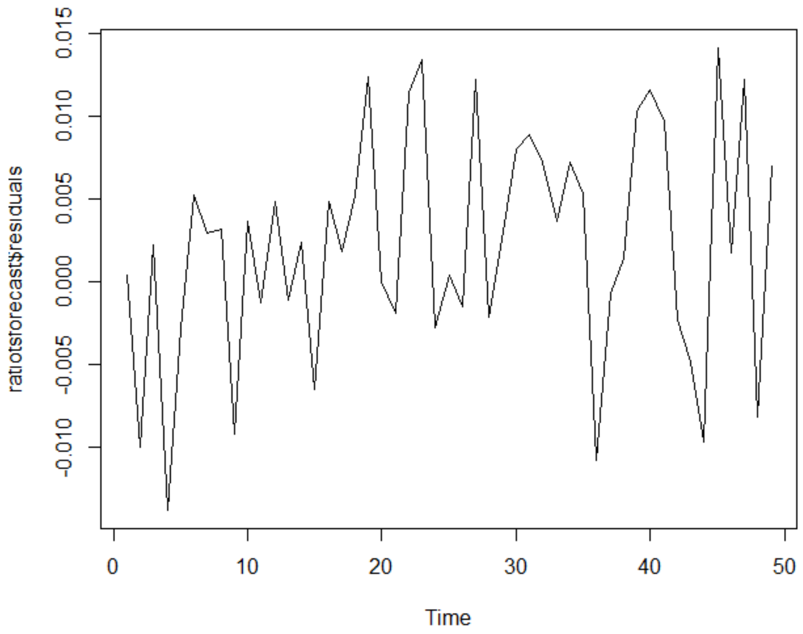
None of the autocorrelations exceed the significance level

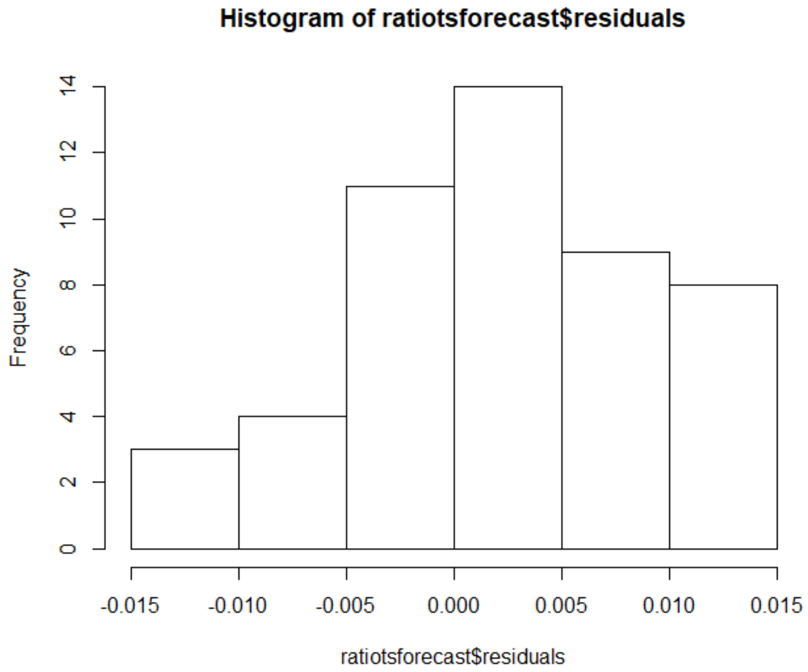


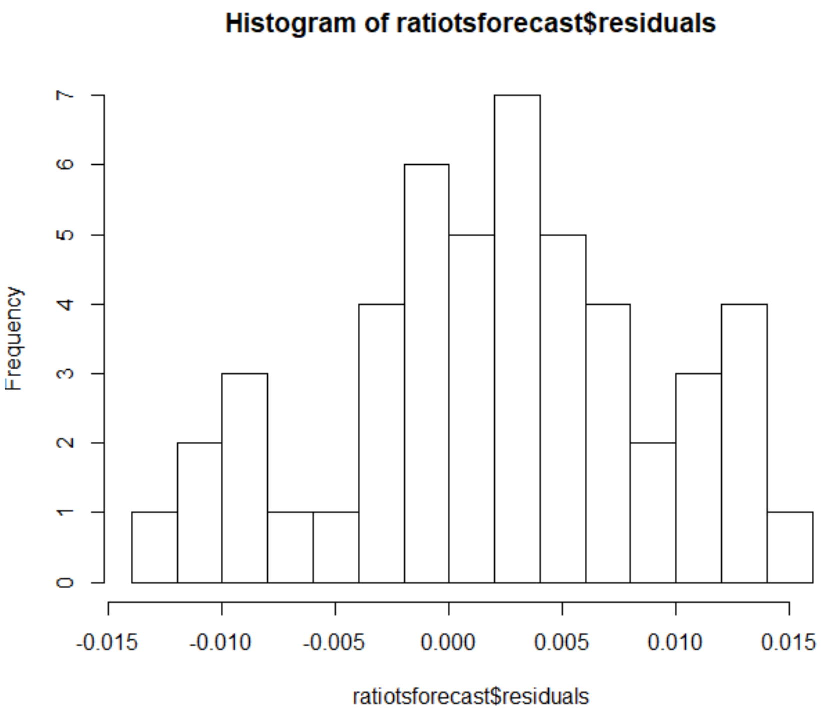
```
        Box-Ljung test

data:  ratiotsforecast$residuals
X-squared = 11.823, df = 20, p-value = 0.922
```

The p-value and the ACF suggests there is very little evidence for non-zero autocorrelations in forecast error terms

To test for the distribution, mean and standard deviation I use the plot of the residuals and a histogram,

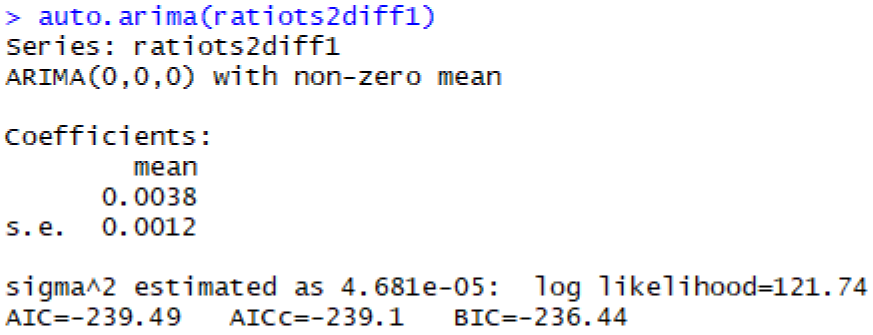

ratiotsforecast$residuals
-0.010
-0.005
0.000
0.005
0.010
0.015
0
10
20
30
40
50
Time


Histogram of ratiotsforecast$residuals
Frequency
0
2
4
6
8
10
12
14
-0.015
-0.010
-0.005
0.000
0.005
0.010
0.015
ratiotsforecast$residuals



While the 0 mean and constant standard deviation are acceptable conclusions but the normal distribution is somewhat questionable due to heavy tails in the histogram.

Since, I found a very significant trend of upward rise during the later observations I tried to fit an arima model on a time series of value only after the year 1980. However even here I found a white noise model.



```
> auto.arima(ratiots2diff1)
Series: ratiots2diff1
ARIMA(0,0,0) with non-zero mean

Coefficients:
        mean
      0.0038
s.e.  0.0012

sigma^2 estimated as 4.681e-05:  log likelihood=121.74
AIC=-239.49   AICc=-239.1   BIC=-236.44
```