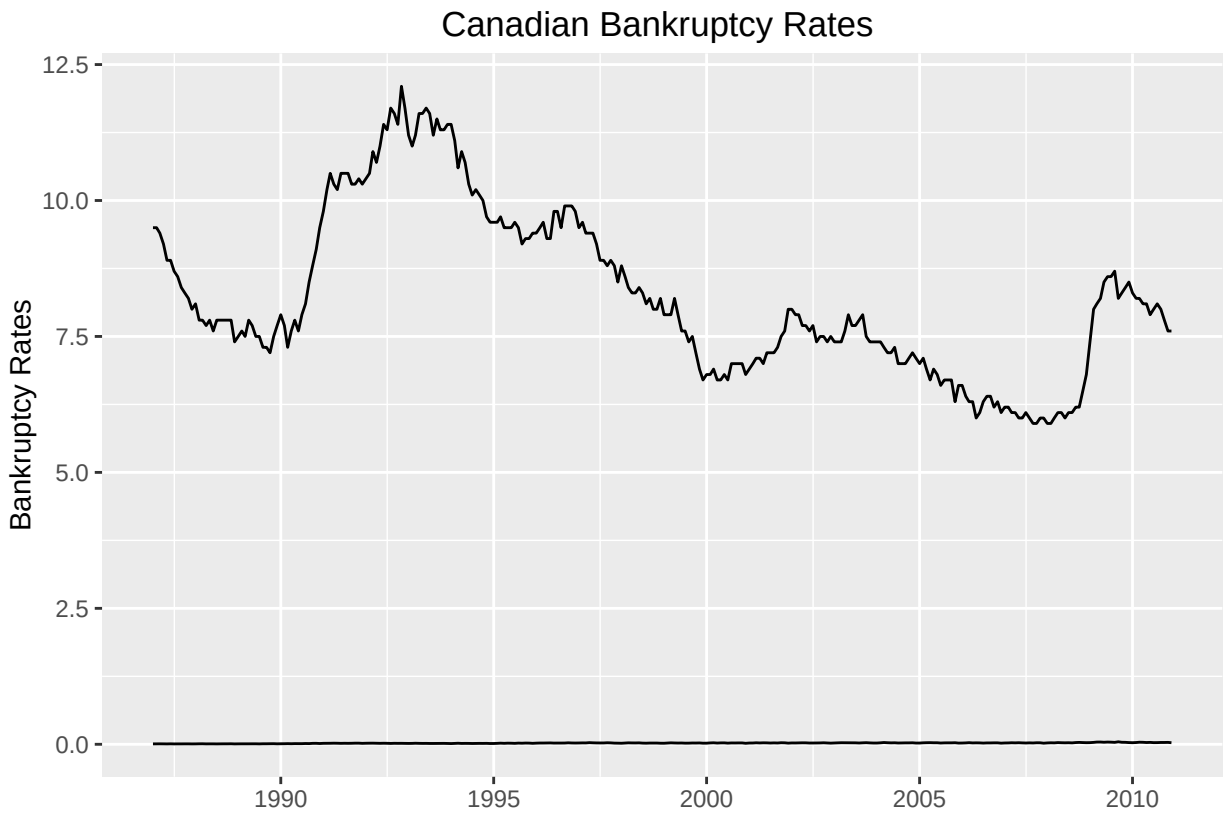
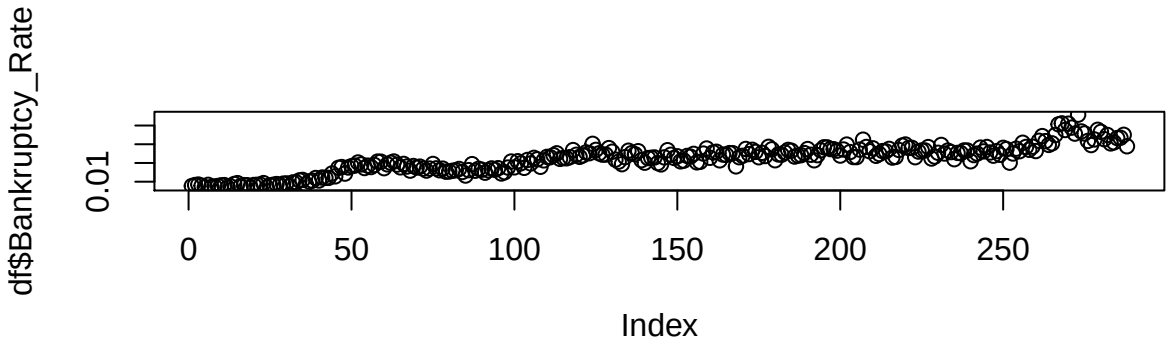
Final Method chosen

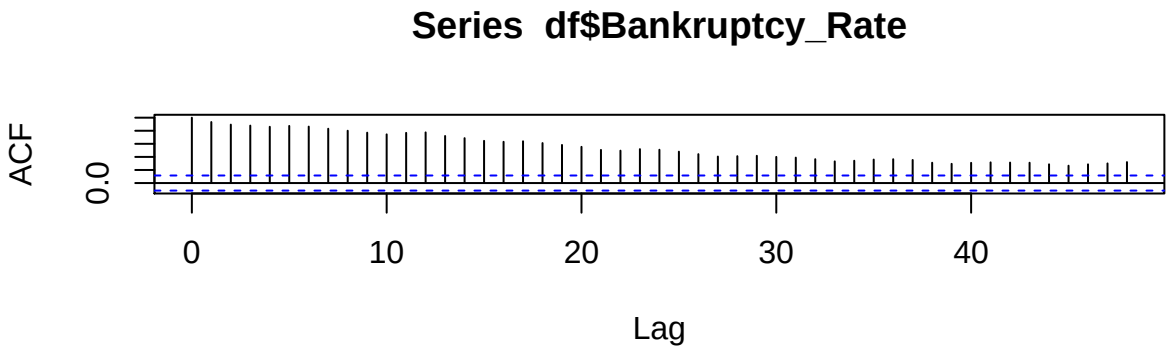
Evelyn Peng

November 27, 2016



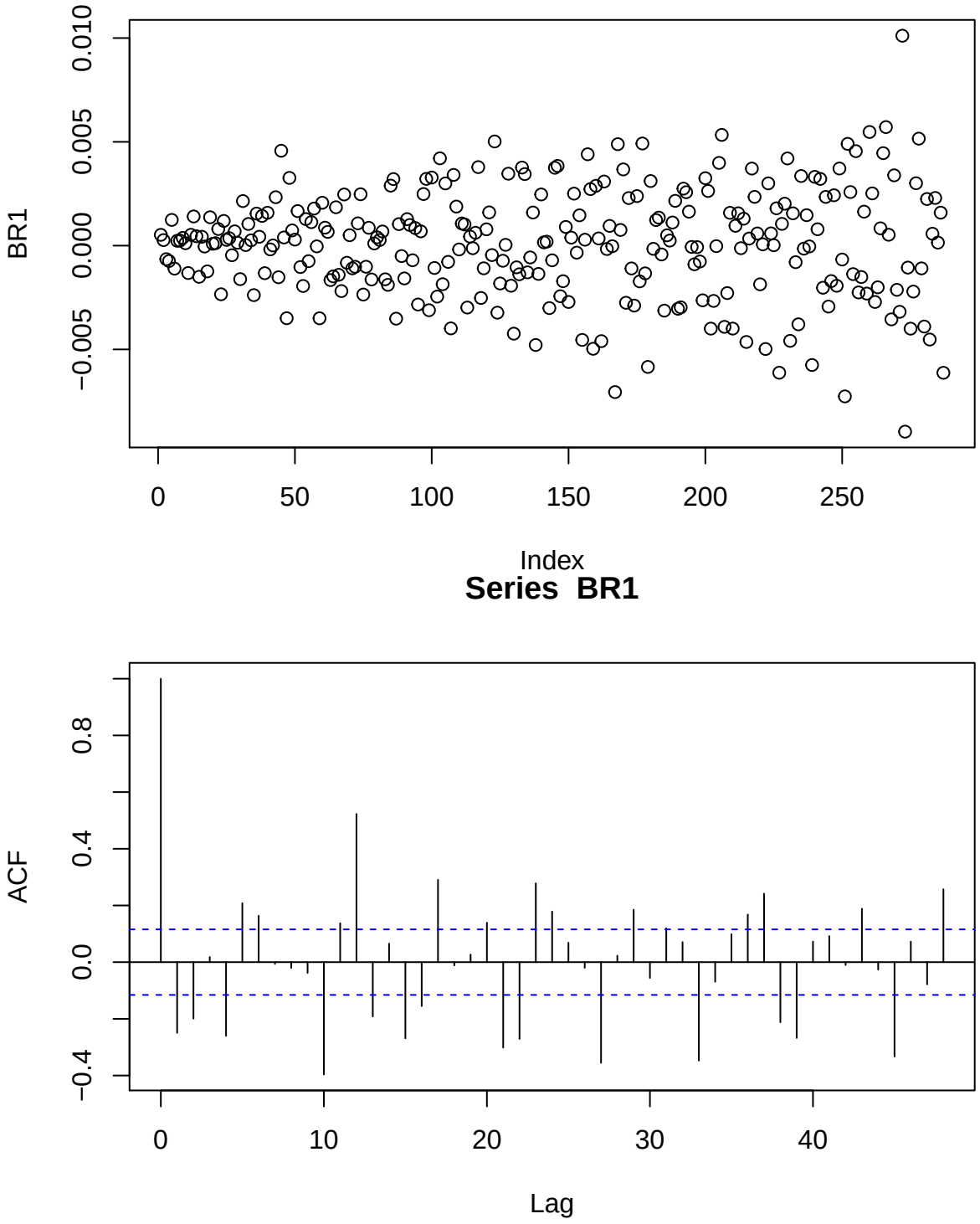
Plot time series data and ACF to see if the data is stationary.






```
##
##  Augmented Dickey-Fuller Test
##
## data:  df$Bankruptcy_Rate
## Dickey-Fuller = -2.4516, Lag order = 6, p-value = 0.3859
## alternative hypothesis: stationary
```

We found the ACF showing undesirable slow decay indicating the data has a trend. And formal test, Augmented Dicky-Fuller, also support this observation with p-value = 0.3859. We should not reject the null hypothesis that the time series data is not stationary. To stable the time series, we begin with ordinary difference to minimize the trend. We then difference the data until the transformed time series looks flat.

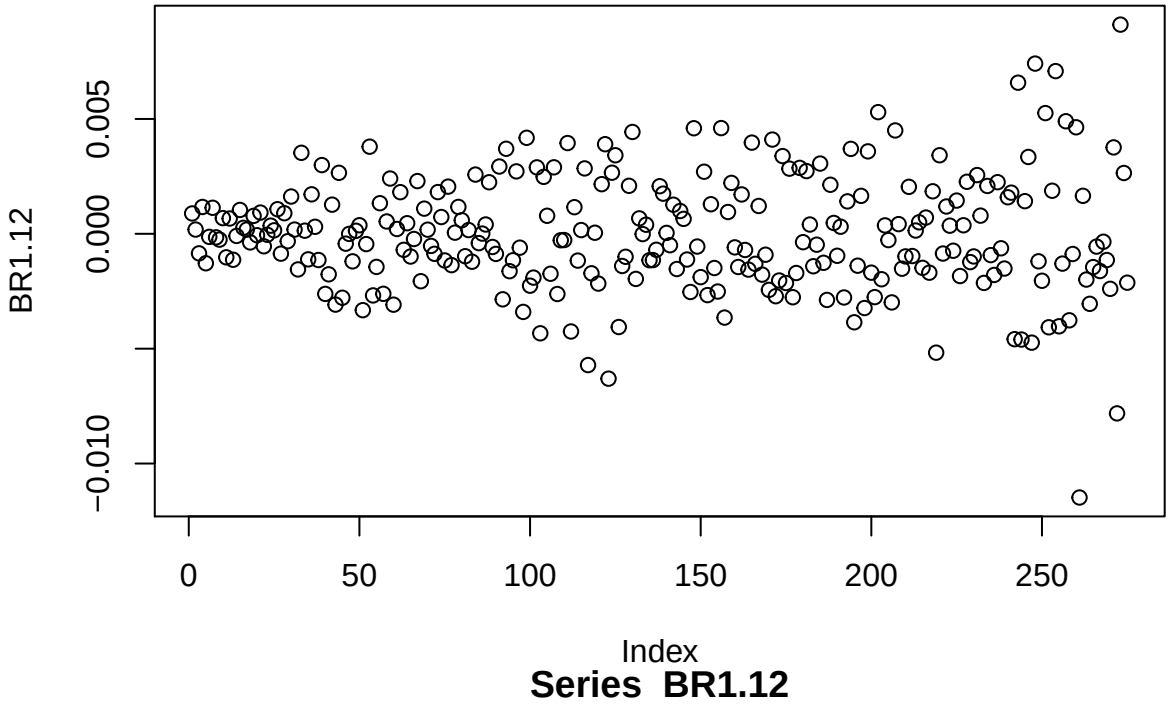



```
## Warning in adf.test(BR1): p-value smaller than printed p-value
```

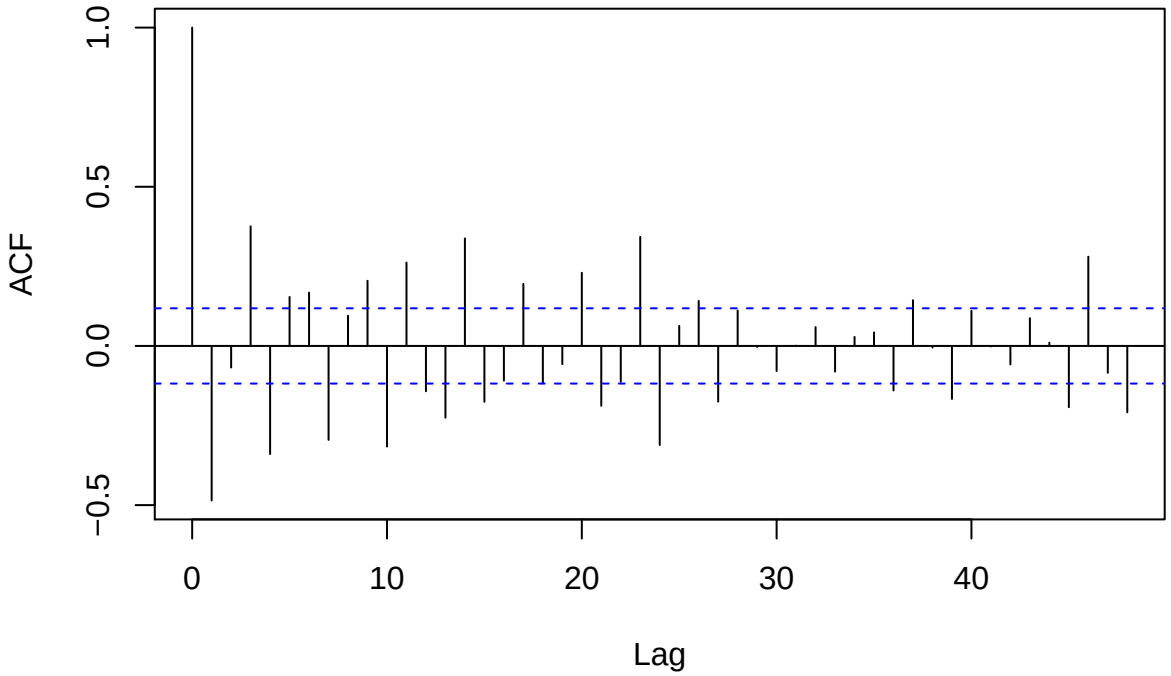
```
##
##  Augmented Dickey-Fuller Test
##
## data:  BR1
## Dickey-Fuller = -7.2702, Lag order = 6, p-value = 0.01
```

```
## alternative hypothesis: stationary
```

The Bankruptcy Rate passed Dickey-Fuller test after difference once. The data is now stationary without trend and we need to eliminate seasonality next.



Series BR1.12



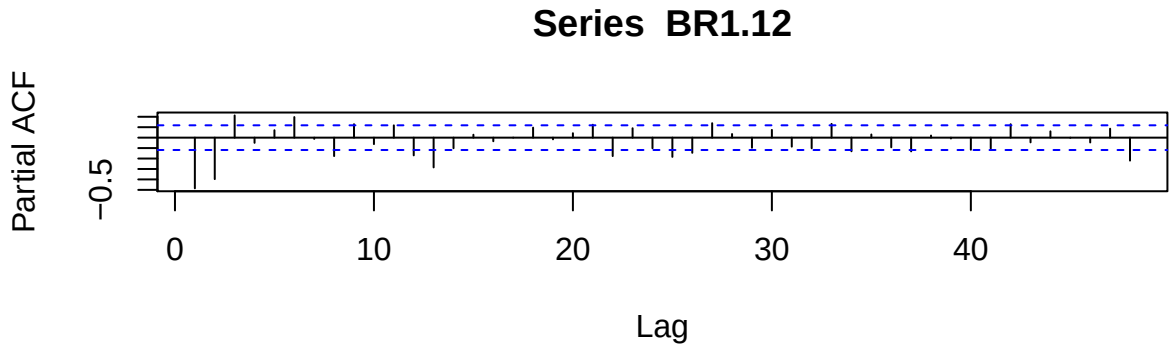
ACF plots looks better after seasonal difference once. The peaks between two cycle decay rapidly. We then use auto.arima to check if using library will give the same result.

```
## Series: df$Bankruptcy_Rate
```

```
## ARIMA(1,1,2)
##
## Coefficients:
##           ar1     ma1      ma2
##       -0.4965  0.0852  -0.5334
## s.e.   0.1028  0.0866   0.0510
##
## sigma^2 estimated as 5.312e-06:  log likelihood=1336.85
## AIC=-2665.71   AICc=-2665.57   BIC=-2651.07
```

```
## [1] 0
```

Even though auto.arima diagnose the data should perform without seasonality, we could still see a clear seasonal cycle in the data. We will process with ordinary difference with once and seasonal difference with twice. From the acf plot of the final different-ed data, the candidate q is between 1 and 7. The possible range of p is 1 and 8. And P, Q both could fall in 0 to 2.



(Using nested for loop) [should not included in the final report], (p,q,P,Q) with value (5,5,2,2) has the lowest AIC value, this servers as our first candidate model. And (5, 6, 2, 1) server as our second candidate model.

```
##
## Call:
## arima(x = df$Bankruptcy_Rate, order = c(5, 1, 5), seasonal = list(order = c(2,
##     1, 2), period = 12), method = "CSS-ML")
##
## Coefficients:
##           ar1      ar2     ar3     ar4     ar5     ma1     ma2      ma3
##       -1.4708  -0.9380  0.2854  0.5521  0.1138  0.9539  0.2668  -0.5025
## s.e.   0.2699   0.5403  0.6167  0.4113  0.1811  0.2599  0.4059   0.3774
```

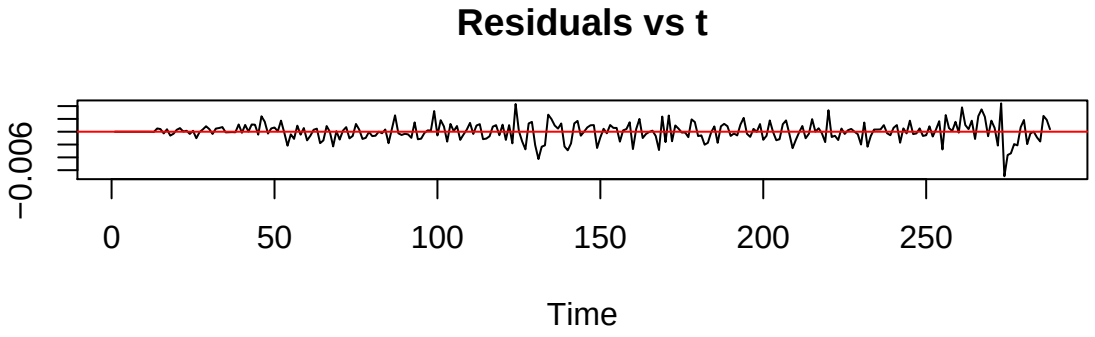
```
##           ma4     ma5    sar1     sar2     sma1    sma2
##       -0.2438  0.2758  0.7263  -0.3098  -1.4290  0.6152
## s.e.   0.2378  0.1572  0.1736   0.0962   0.1797  0.1300
##
## sigma^2 estimated as 2.016e-06:  log likelihood = 1402.02,  aic = -2774.04
```

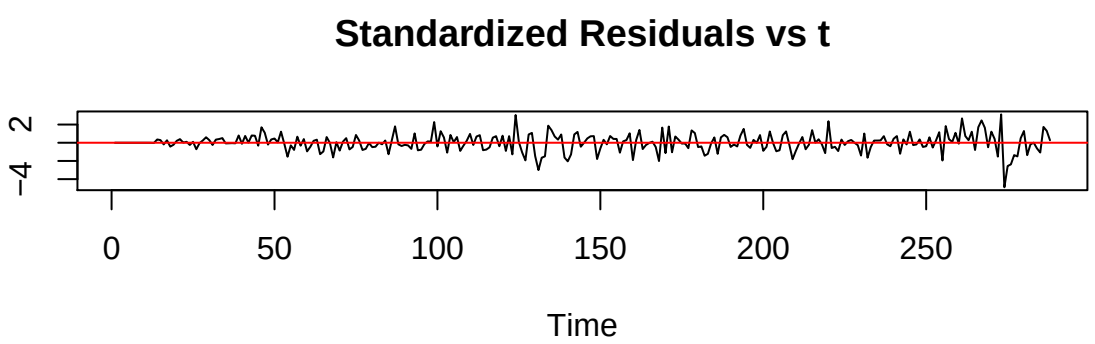
```
##
## Call:
## arima(x = df$Bankruptcy_Rate, order = c(5, 1, 6), seasonal = list(order = c(2,
##     1, 1), period = 12), method = "CSS-ML")
##
## Coefficients:
##           ar1     ar2     ar3      ar4      ar5      ma1      ma2      ma3
##       -0.2956  0.8192  1.0801  -0.0144  -0.7158  -0.2558  -0.9227  -0.5228
## s.e.   0.1065  0.1174  0.0238   0.1122   0.0990   0.1230   0.0925   0.0788
##          ma4     ma5      ma6    sar1     sar2     sma1
##       0.4953  0.7272  -0.5210  0.0892  -0.1326  -0.7058
## s.e.  0.1253  0.0767   0.0731  0.0957   0.0797   0.0743
##
## sigma^2 estimated as 1.998e-06:  log likelihood = 1401.81,  aic = -2773.61
```

	(5,1,5) * (2,1,2)	(5,1,6) * (2,1,1)
AIC	-2774.04	-2773.61
log likelihood	1402.02	1401.81

From the comparison table, we could clearly confirm the first model perform better than the second one. We therefore choose the parameter of arima model (p,q,P,Q) accordingly. With the candidate model arima (5,5,2,2), we start checking if it satisfy the formal and informal residual diagnostics.

i. Zero-Mean

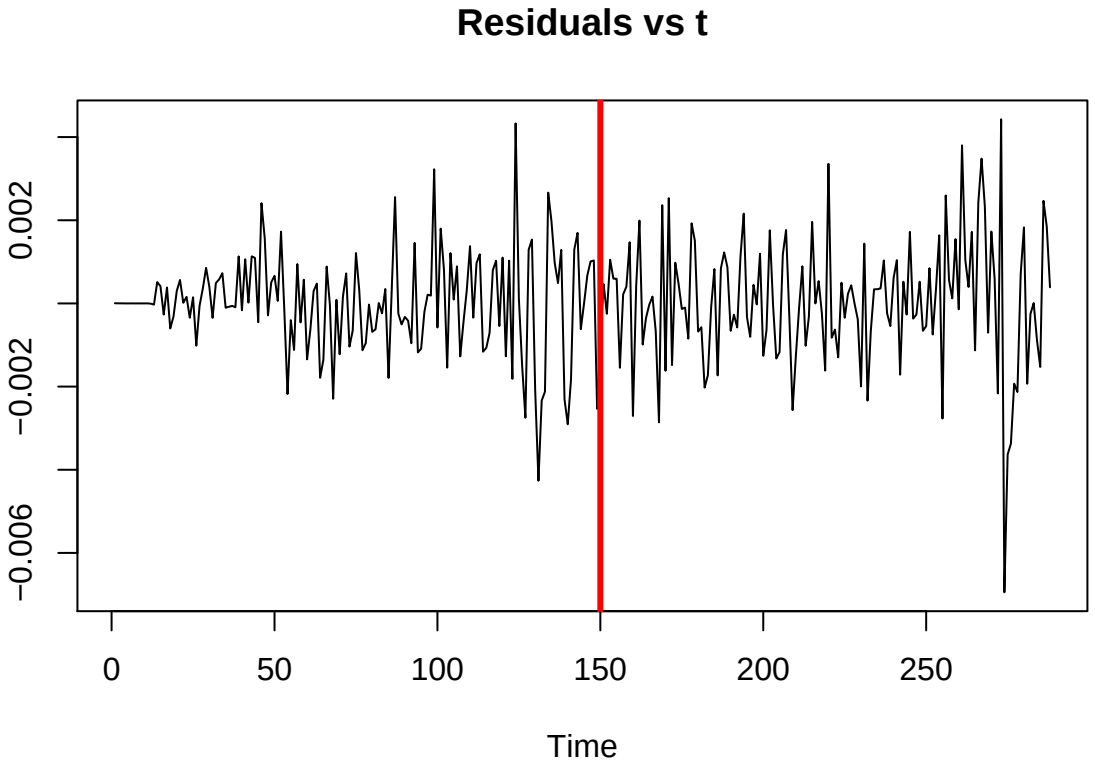





```
##                                   t
## "Test Statistic:"          "0.1317"        "P-value:"          "0.8953"
```

From the plot of standardized residuals and time t, we see no obvious above or below 0. And the one sample t-test gave p-value = 0.8953 indicating we should not reject the null hypotheses. The true mean is equal to 0 with more than 95 percent confidence level.

ii. Homoscedasticity

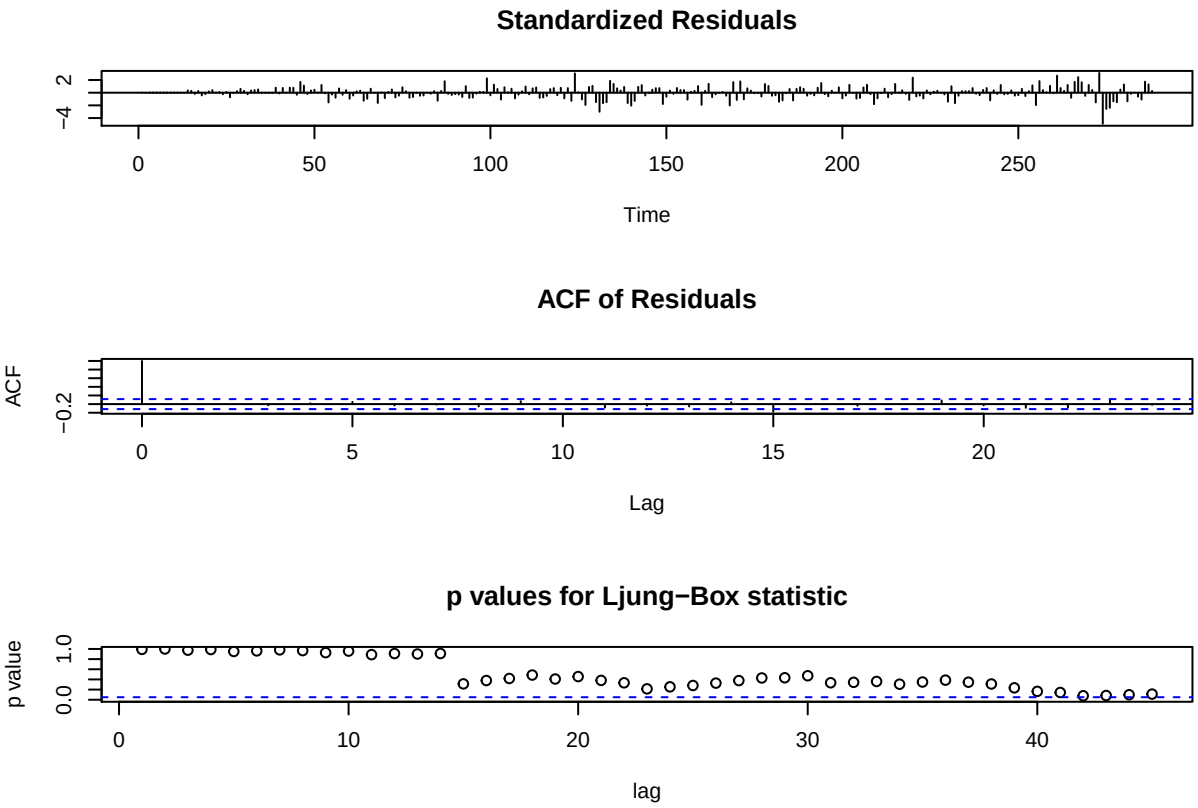



```
##                       Test Statistic
## "Test Statistic:"          "7.4013"        "P-value:"          "0.0069"

##                       Bartlett's K-squared
##    "Test Statistic:"             "9.9212"           "P-value:"
##
##             "0.0016"
```

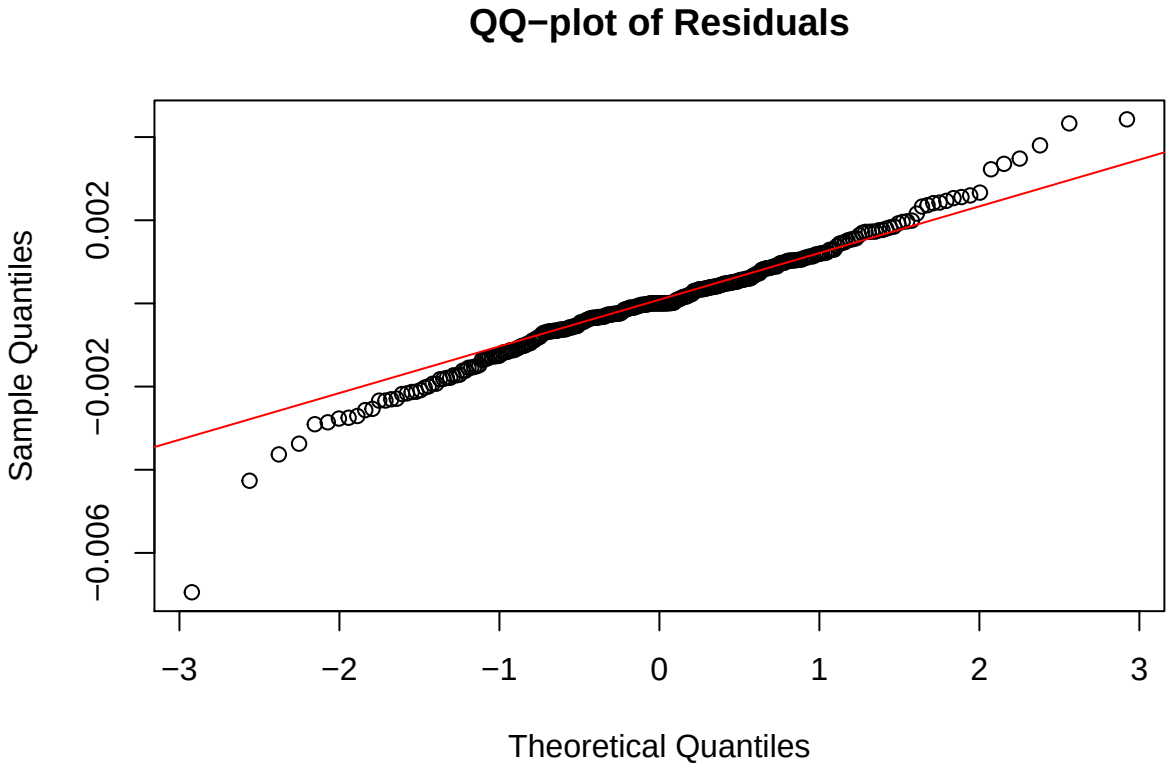
From the plot of standardized residuals and time t, we see slightly differences variance between groups. I divided the group by 3, each group last for around 100. The variance do not stay constant for all the groups. Levene test gave p-value = 0.0069 and Bartlett test gave p-value more than 0.0016, indicating we should reject the null hypothesis that all variance are the same across groups. However, the Bartlett test is sensitive to even slightly deviation from normal distribution, we should not trust the p-value. And the levene test gave a p-value not really small. Therefore, even though this model fails to pass leven test, the heteroskedasticity in this model is not too obvious.

iii. Zero-Correlation



The function tsdiag give the graph of ACF and Ljung-Box test all in one! From the ACF plots, it shows the residuals' are uncorrelated (only one spikes at 0 and no spikes afterwards) and for formal test of correlation, Ljung-Box test, all the p-value are larger than the critical value and we should not reject the null hypothesis that all correlations are equal to 0 because all the p value is above the confidence interval. The residuals do not have correlation for all lags.

iv. Normality

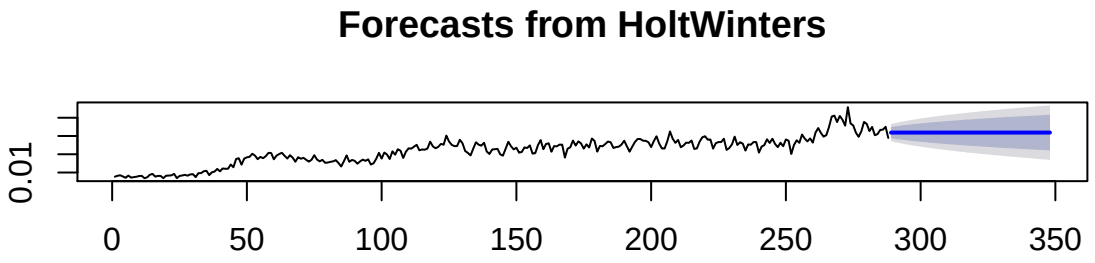


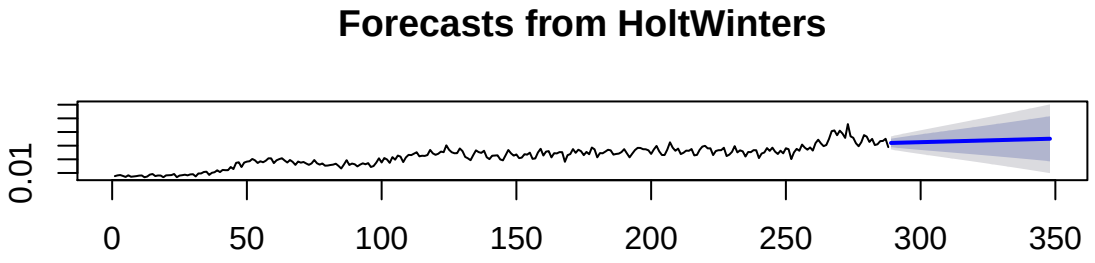

```
##                                   W
## "Test Statistic:"          "0.9739"        "P-value:"              "0"
```

The qqplot seems quite good, the empirical dots lies on the theoretical normal distribution line. Although the formal test for normality-Shapiro test, gave relatively small p value suggesting we should reject the null hypothesis, the visualization showing the residues are quite normally distributed. The normality test is not useful when dealing with real life problem. No real quantity is exactly normally distributed. The normal distribution is just a mathematical abstraction that's a good enough approximation in a lot of cases. Therefore, we should conclude our data is normal distributed enough by observing qqplot.

Exponential smoothing (Holt-Winters Methods)

From previous diagnosis, we found there are trend and seasonality in these data. Therefore, we adapt Triple Exponential Something method with multiple effect (apply on heteroskedastic data). However, after trying four different types of Holt-Winter method, there is no sign of seasonality. We cannot use Exponential Smoothing to predict future bankruptcy rates.


Forecasts from HoltWinters
0.01
0
50
100
150
200
250
300
350


Forecasts from HoltWinters
0.01
0
50
100
150
200
250
300
350