Final Method chosen

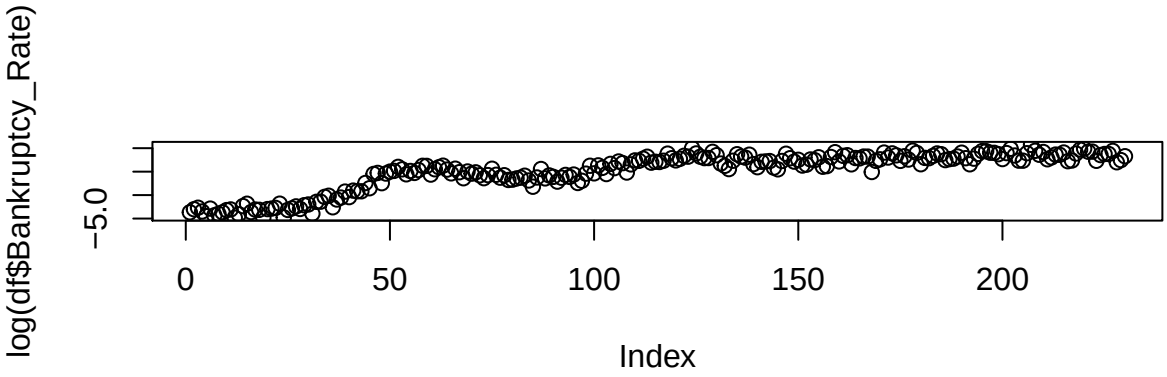
Evelyn Peng

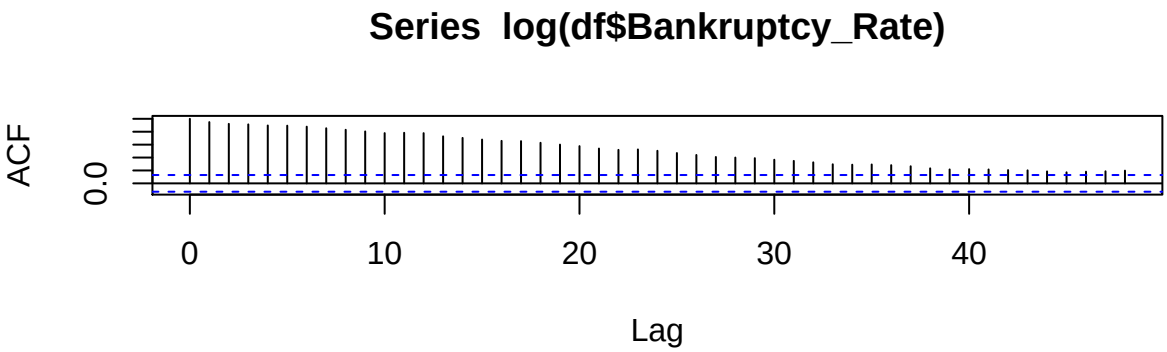
November 27, 2016

ARIMA model

Check for non-constant variance and apply suitable transformation

Plot time series data and ACF to see if the data is stationary. We found a log transformation is needed.

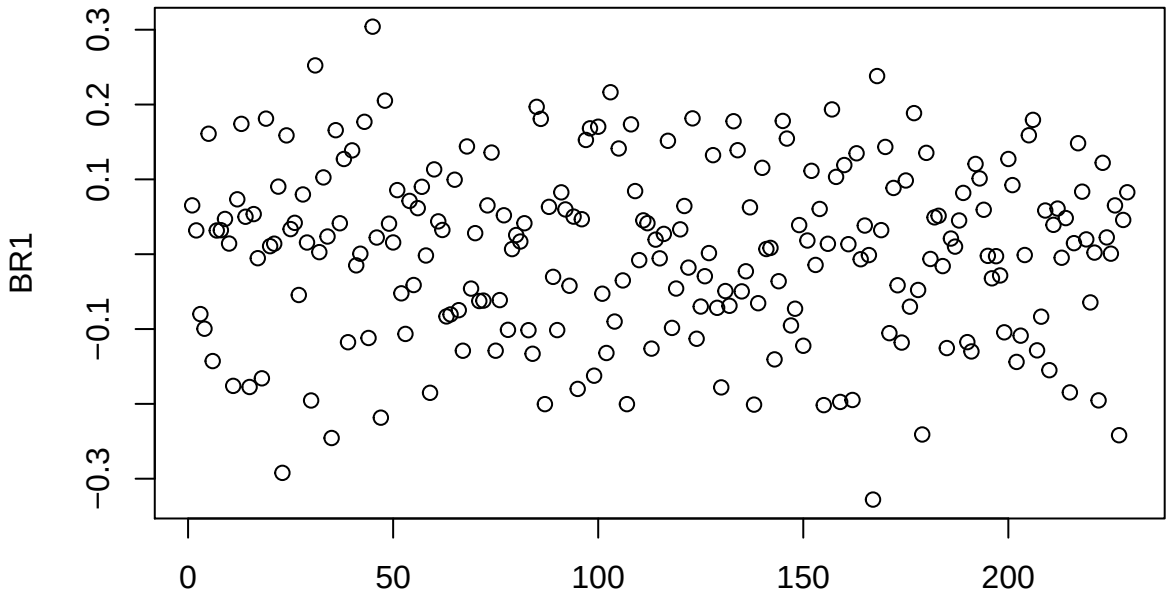




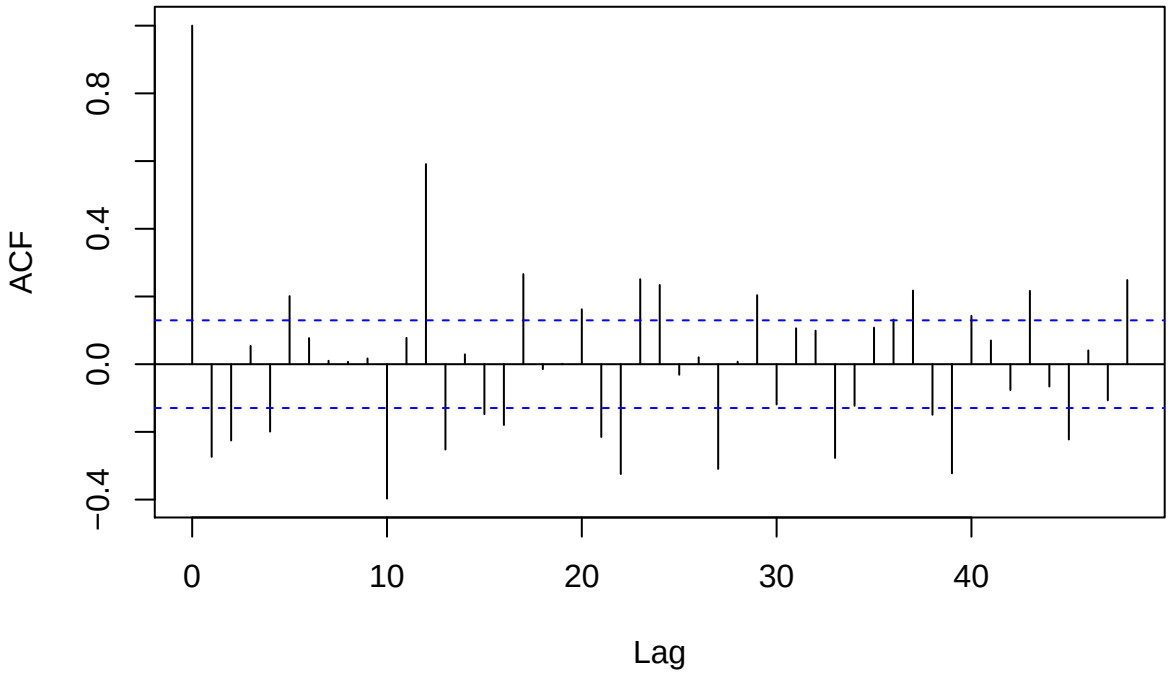

```
##                      Dickey-Fuller
## "Test Statistic:"         "-1.3595"        "P-value:"          "0.845"
```

Check for seasonal or non-seasonal trend

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 = 0.845. 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.



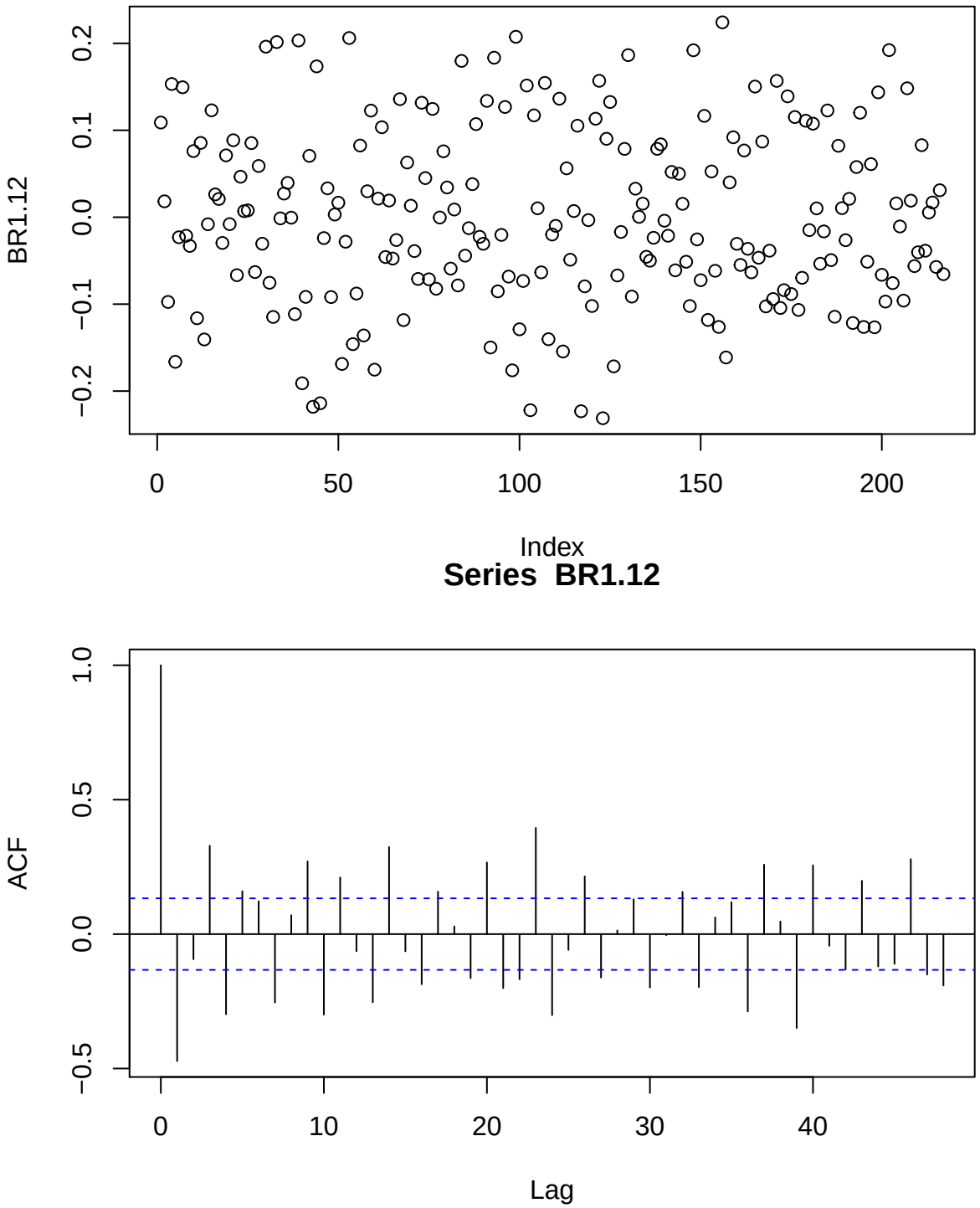
Series BR1




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

```
##                       Dickey-Fuller
## "Test Statistic:"          "-6.8225"        "P-value:"            "0.01"
```

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



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.

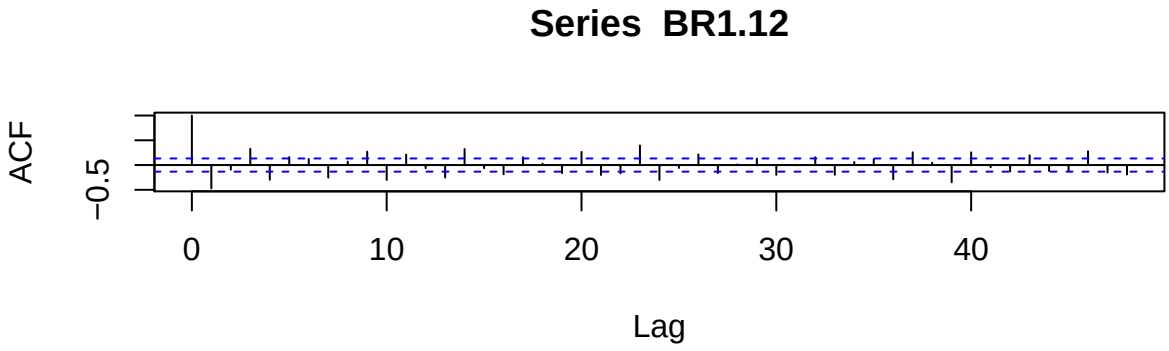
```
## Series: log(df$Bankruptcy_Rate)
## ARIMA(3,1,4)
##
## Coefficients:
##          ar1      ar2      ar3      ma1     ma2     ma3      ma4
```

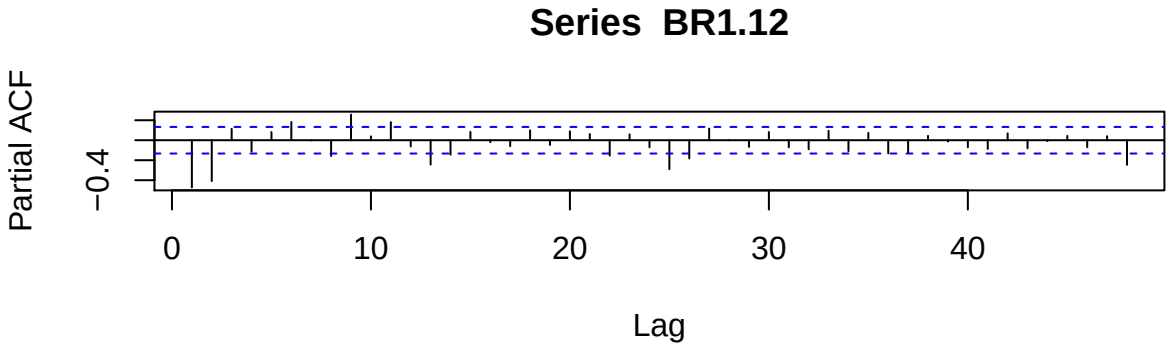
```
##       0.2918  -0.2557  -0.7007  -0.8012  0.1617  0.8083  -0.4336
## s.e.  0.0932   0.0917   0.0905   0.0893  0.1046  0.0934   0.0522
##
## sigma^2 estimated as 0.00786:  log likelihood=231.22
## AIC=-446.43   AICc=-445.78   BIC=-418.96

## [1] 0
```

Identify (p,q,P,Q) according to ACF/PACF plots

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.





Fit propsed model and iterate to the optimal model

(Using nested for loop) [should not included in the final report], (p,q,P,Q) with value (8,2,1,2) has the lowest AIC value, this servers as our first candidate model. And (6, 5, 2, 2) server as our second candidate model.

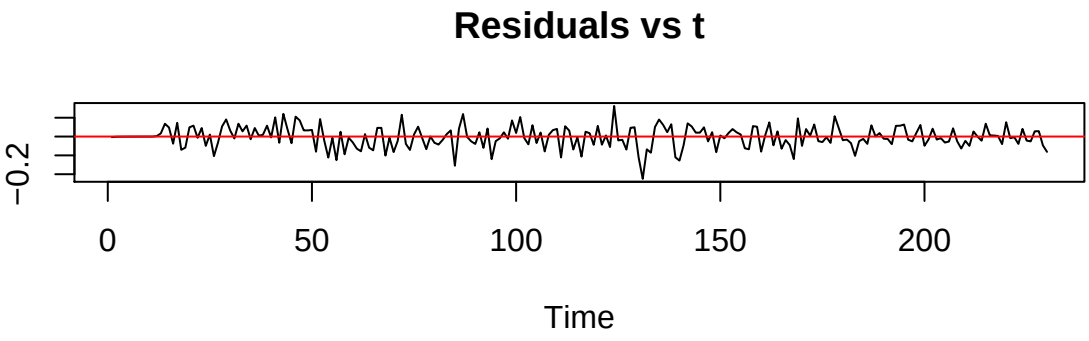
```
## Warning in arima(log(df$Bankruptcy_Rate), order = c(6, 1, 5), seasonal =
## list(order = c(2, : possible convergence problem: optim gave code = 1
```

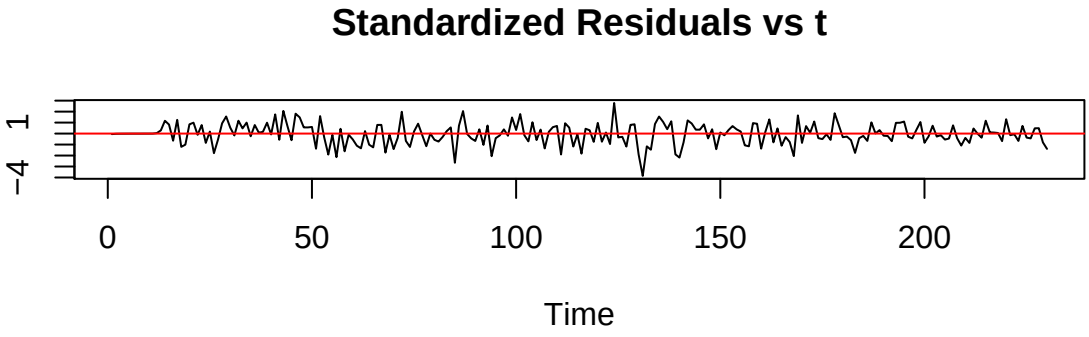
	(6,1,5)X(2,1,2)	(7,1,2)X(0,1,2)	(7,1,3)X(1,1,1)	(8,1,2)X(1,1,2)
$\hat{\sigma}^2$	0.003402	0.003367	0.003367	0.003342
log likelihood	298.64	298.64	298.46	299.75
AIC	-560.28	-573.28	-570.93	-571.5

Check fit with residual assumption or not

From the comparison table, we could clearly confirm the first model perform better than the otheres. Although the AIC is not the smallest among these four models, but it has least $\hat{\sigma}^2$ and largest log likelihood value. And it only traded off AIC for less than 0.1. We therefore choose the parameter of arima model (6,1,5)X(2,1,2) accordingly. With the candidate model arima (6,1,5)X(2,1,2), we start checking if it satisfy the formal and informal residual diagnostics.

i. Zero-Mean

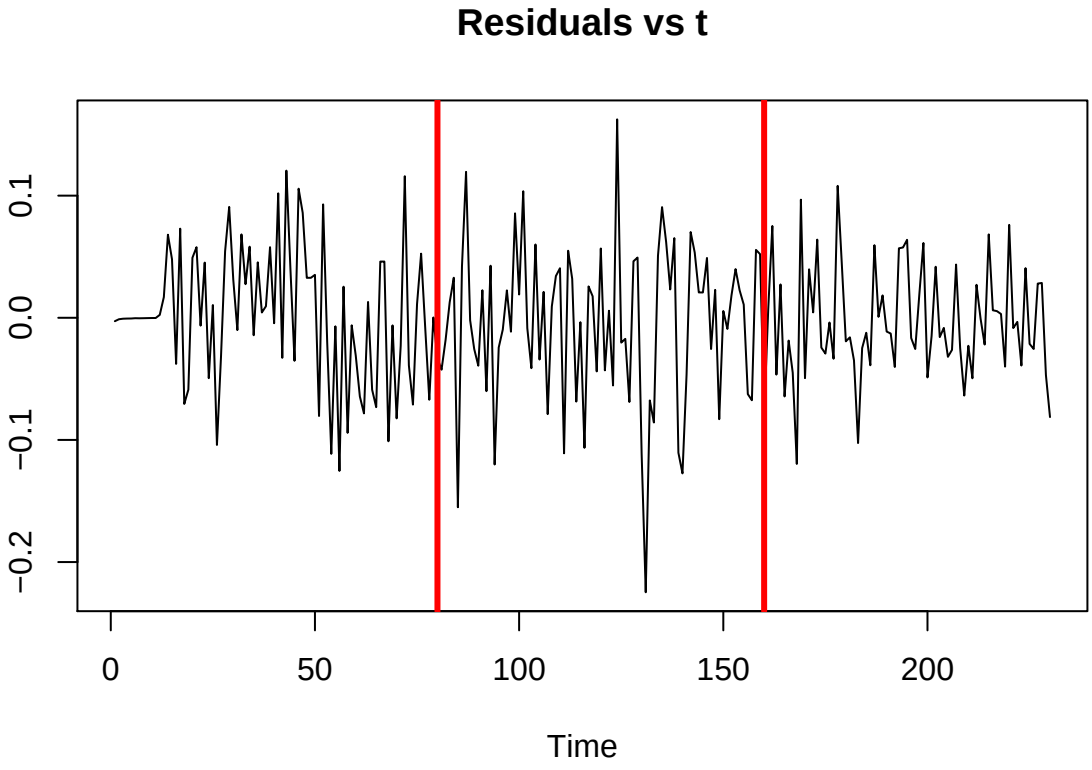





```
##                                t
## "Test Statistic:"        "-0.6305"       "P-value:"          "0.529"
```

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 = 0.529 indicating we should not reject the null hypotheses. The true mean is equal to 0 with more than 95 percent confidence level.

ii. Homoscedasticity

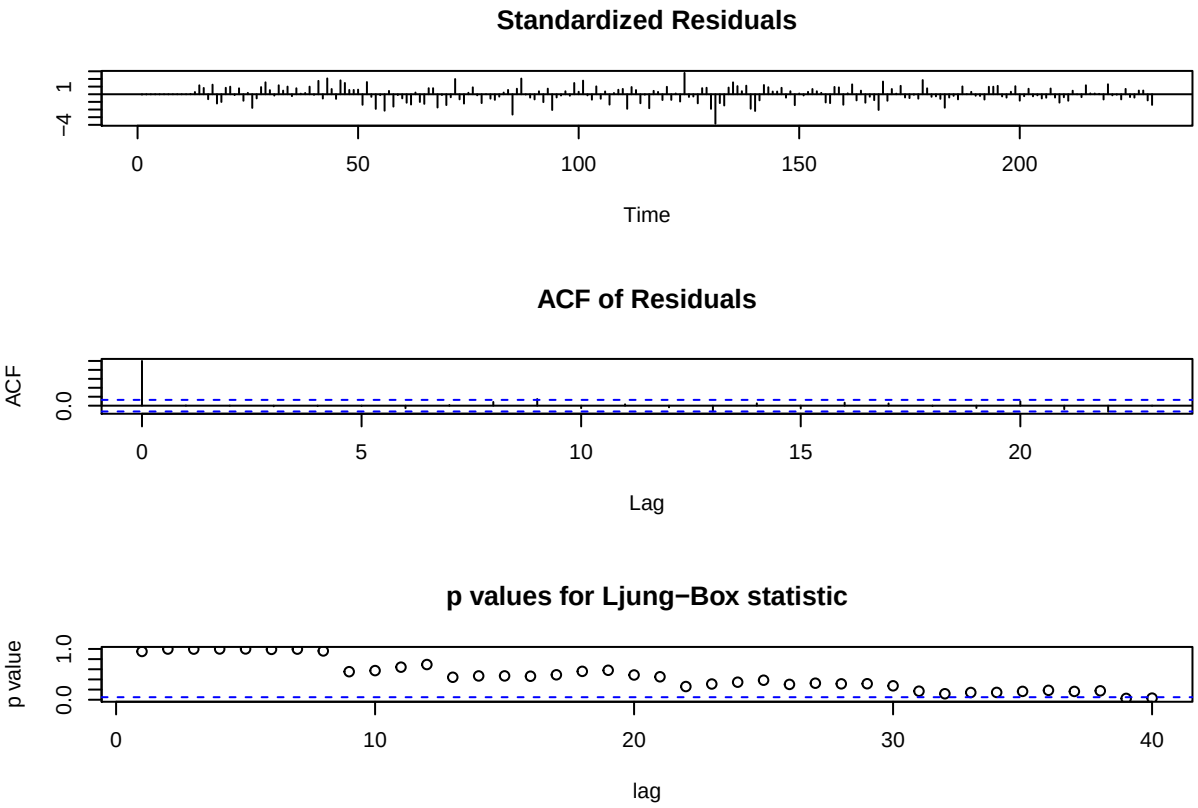



```
##                       Test Statistic
## "Test Statistic:"          "3.8038"        "P-value:"          "0.0237"

##                       Bartlett's K-squared
##    "Test Statistic:"            "10.0313"           "P-value:"
##
##             "0.0066"
```

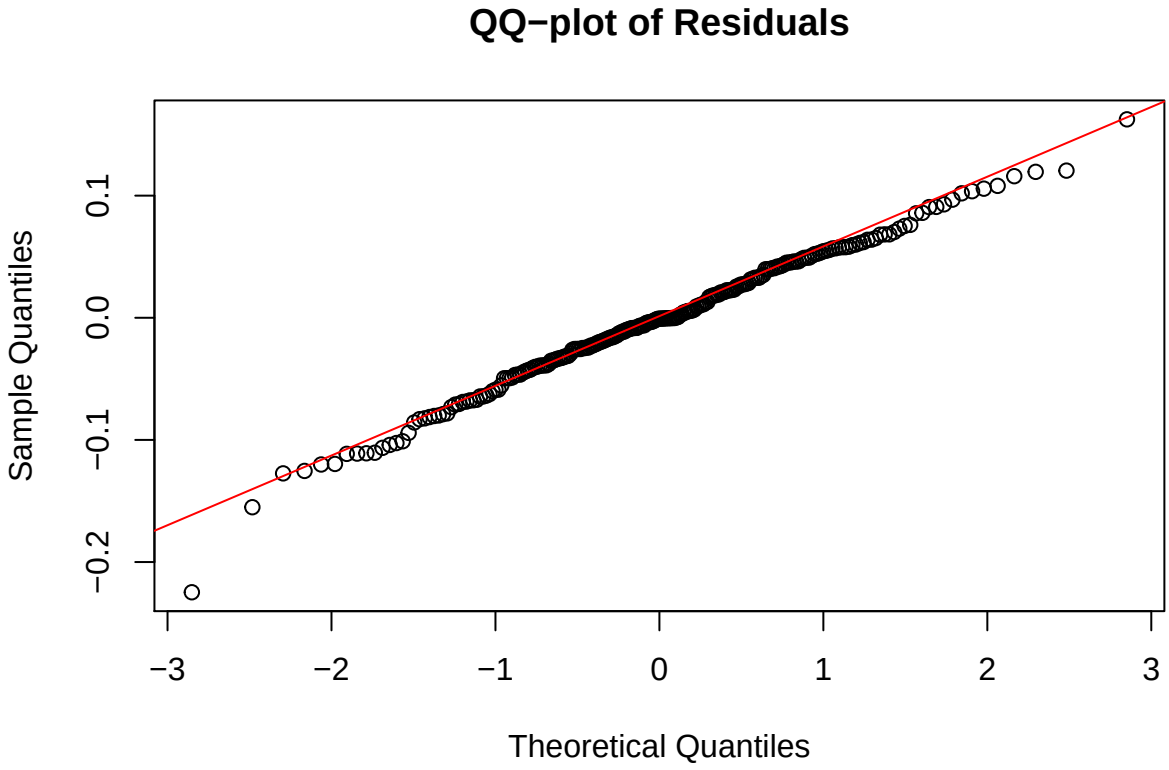
From the plot of standardized residuals and time t, we see no obvious differences variance between groups. I divided the group by 3. The variance stays constant for all the groups. And both levene test (0.6463) and Bartlett test (0.3828) gave p-value more than 0.3, indicating we should not reject the null hypothesis that all variance are the same across groups.

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:"          "0.9918"        "P-value:"          "0.2248"
```

The qqplot seems quite good, the empirical dots lies on the theoretical normal distribution line. And the formal test for normality-shapiro test, gave a p-value = 0.1755 suggesting not reject the null hypothesis. The residuals are normally distributed.

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.

ARIMAX

Considering there are external time series influences the bankruptcy rate, we take into account their effection when building the model by multivariate time seriese model. First, we start with ARIMA model. Continue using what we got earlier from SARIMA model, we fit in the data with parameter of arima model (6,1,5)X(2,1,2) and other time series: unemployment rate, population and house price index. Comparing all the combination of these three additonal predictors, the one with unemployment rate and house price index has the least AIC among all candidates. However, the AIC is bigger than the one without using external time series variables. As ARIMAX model address on the situation when these external variables are exougenous, this might not fit with our data as these variables might influence bankruptcy rate and bankruptcy rate would influnce these variables too.

VAR

It's valid to assume these external variables are endogenous, we then use vetor autoregression model to fit our data as next step. Using Akaike Information Criterion, Hannan-Quinn Criterion, Schwarz criterion and Akaike's Final Prediction Error Criterion, we choose p equals to 6 as this is the smallest number Schwarz criterion suggests. We should keep our model as simple as possible.

Final model choosing

Using RMSE to compare above four models: ARIMA, Exponential smoothing, ARIMAX and VAR. The RMSE was calculated based on the validation data which is that last 20% of training data. We conclude the ARIMA model is our final choose because it has the lowest RMSE among four models.

	ARIMA	Exponential Smoothing	ARIMAX	VAR
RMSE	0.006506753	0.01333646	0.03077996	1.000244