Time Series Analysis of Severity Data

****

December 31, 2014

The data under investigation is monthly claim severities for recreational vehicle insurance covering a period from 2004 to 2014. I will use the data from 2004 to 2013 to create the model and use the 2014 data for validation purposes. For our purposes lets define severity as:

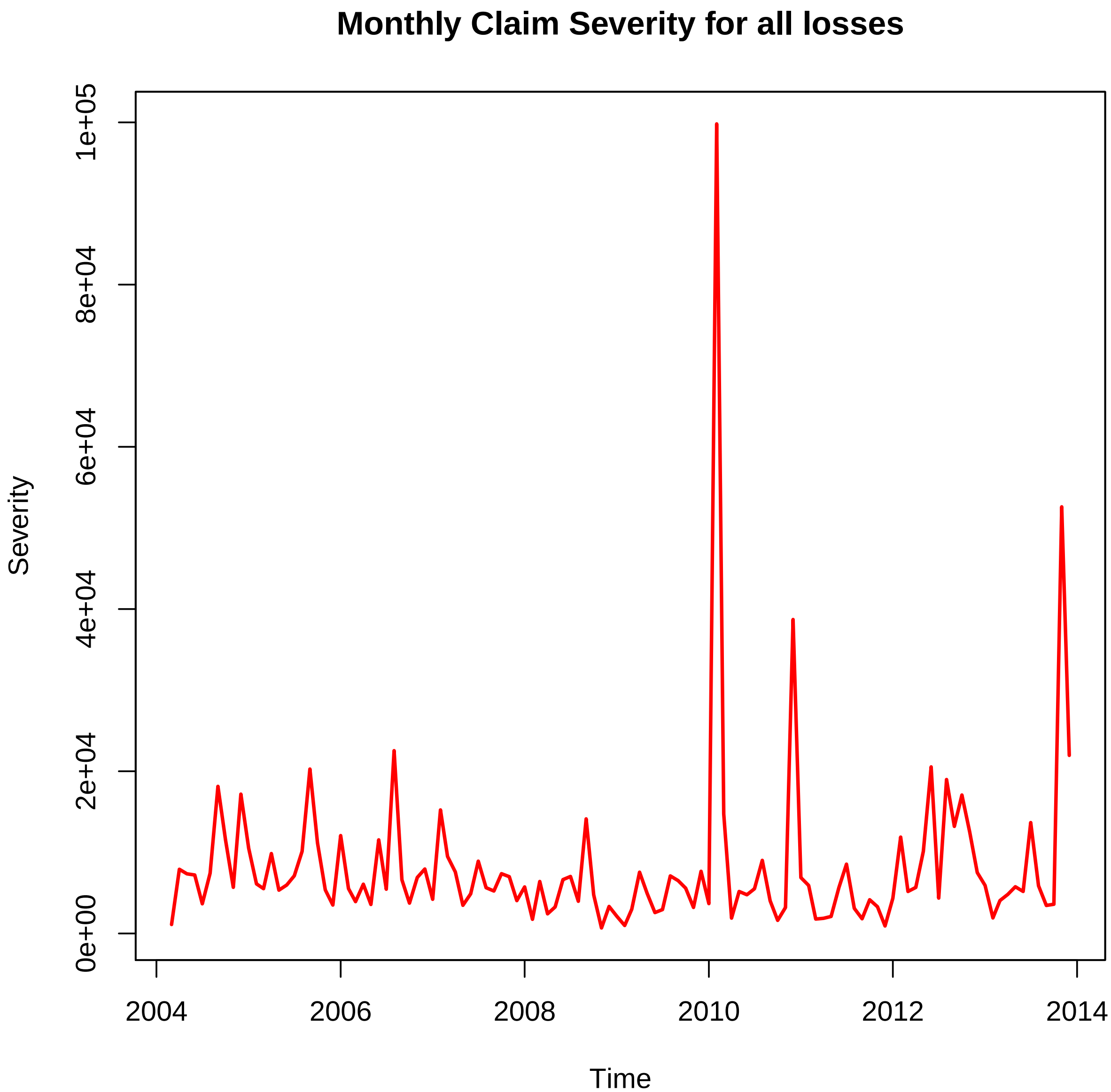
$$\text{Severity} = \frac{\text{total losses}}{\text{number of claims}}$$

This data has all losses, in cluding catastrophe losses, up to one million dollars. This was not by my choice it was already in this state.

Summary of the data.

```
##      Index                  tsev
##  Min.   :2004-03-01   Min.   :  681
##  1st Qu.:2006-08-08   1st Qu.: 3698
##  Median :2009-01-16   Median : 5685
##  Mean   :2009-01-15   Mean   : 8304
##  3rd Qu.:2011-06-23   3rd Qu.: 8396
##  Max.   :2013-12-01   Max.   :99811
```

Plot of the data



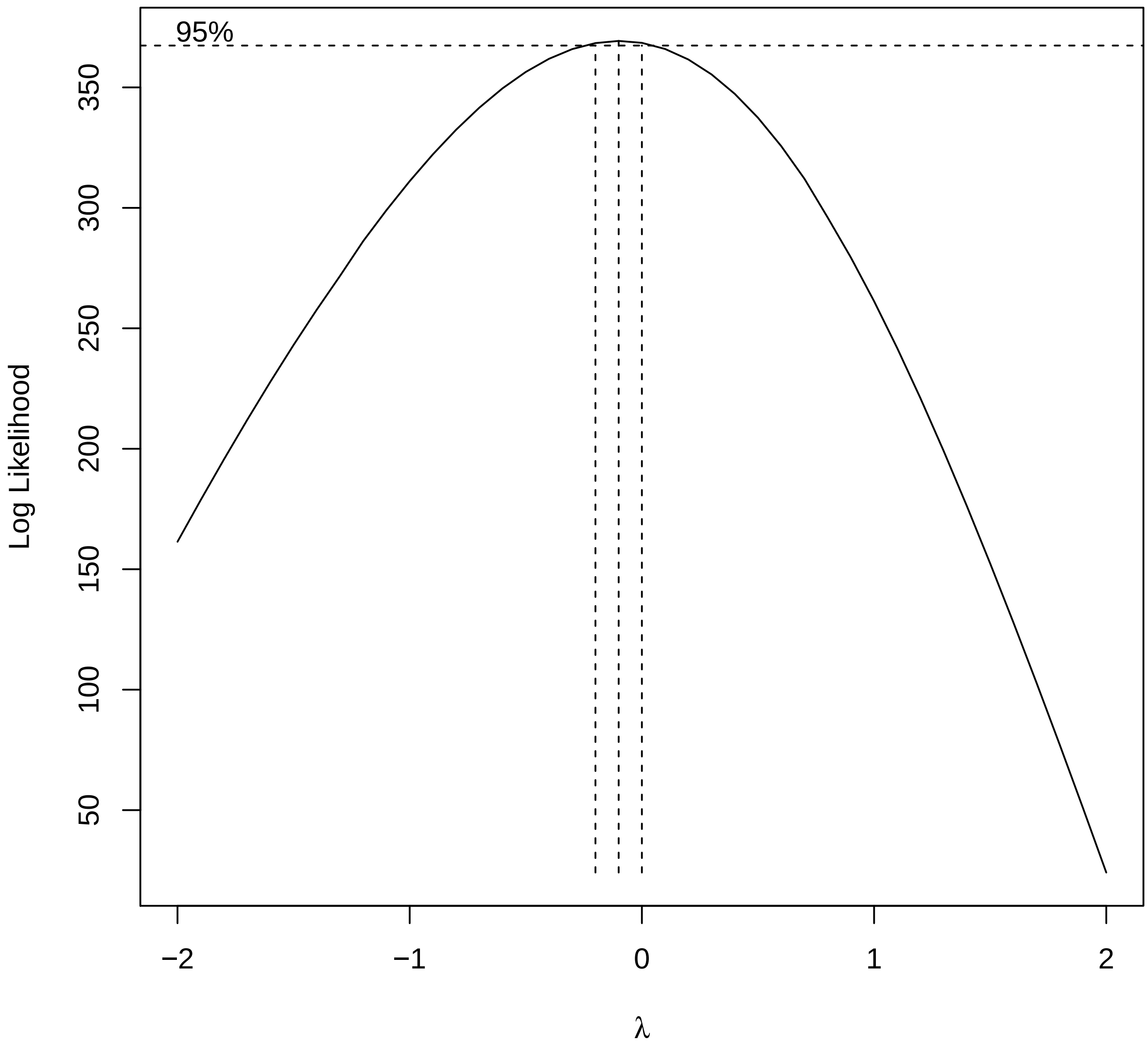
The variation seems to increase towards the end of the time series so a transformation will be used. As suggested in Cryer & Chan I will try a power series transformation to rein in the large values. Since all of the data points are positive, there is no need to shift the data.

$$g(x) = \begin{cases} \frac{x^{\lambda}-1}{\lambda} & \text{for } \lambda \neq 0 \\ \log(x) & \text{for } \lambda = 0 \end{cases}$$

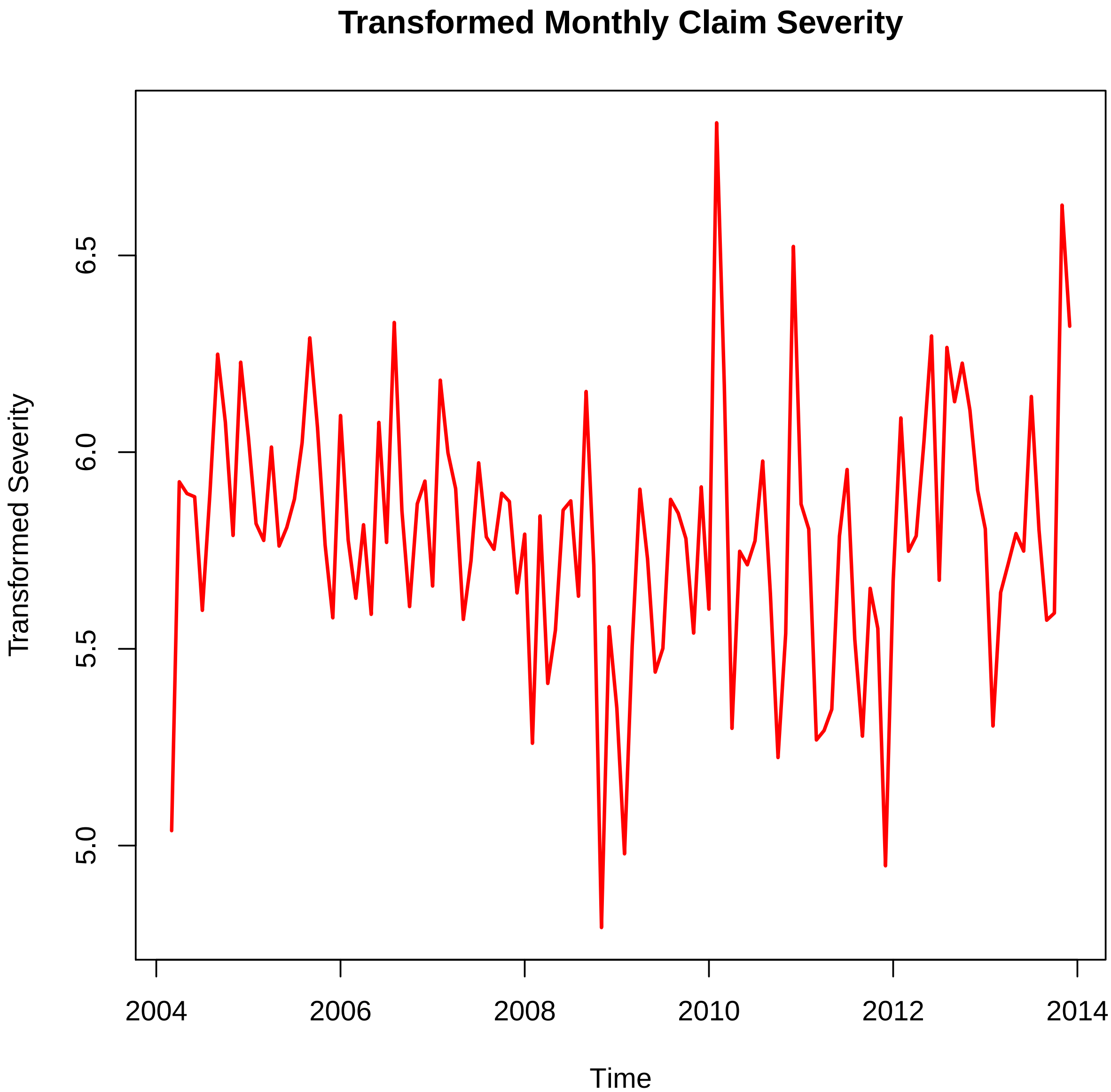
The R function maximizes the loglikelihood of g(x) relative to the a normal likelihood function. The output below is the MLE and confidence interval for the MLE. There is also a graphical representation of the likelihood function.

```
## $mle
## [1] -0.1
##
```

```
## $ci
## [1] -0.2  0.0
```

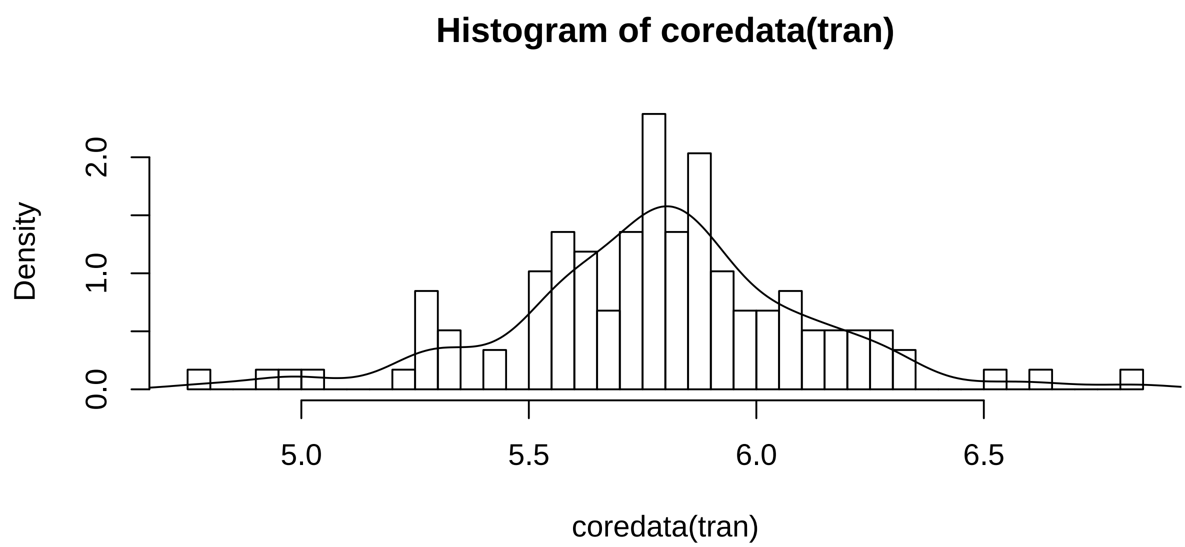


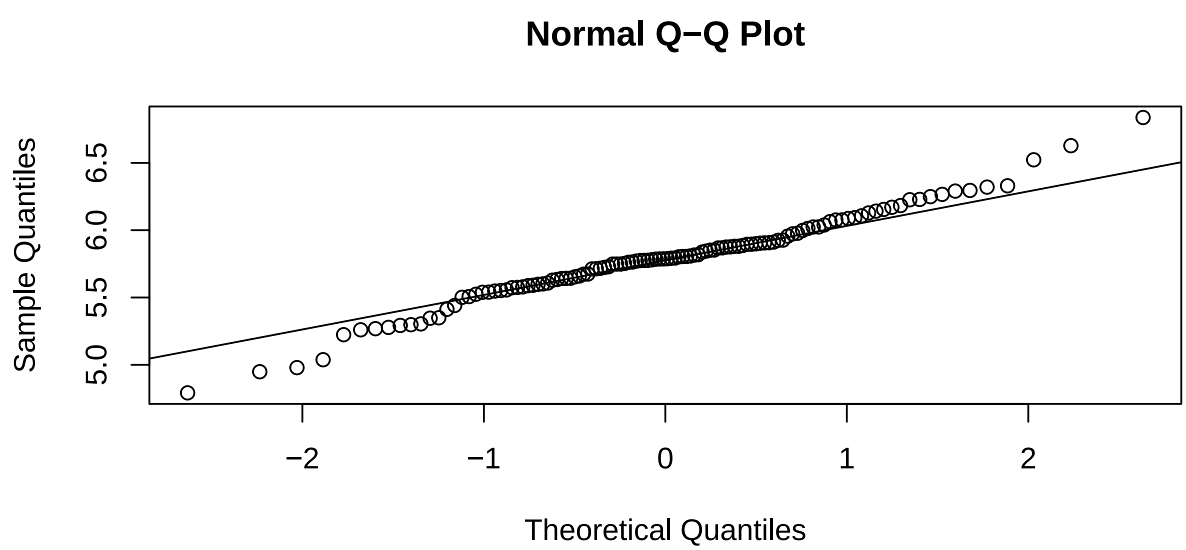
This suggests that using λ = -0.1 is the best choice. But anything in the confidence interval of $(-0.2, 0)$ will suffice technically.



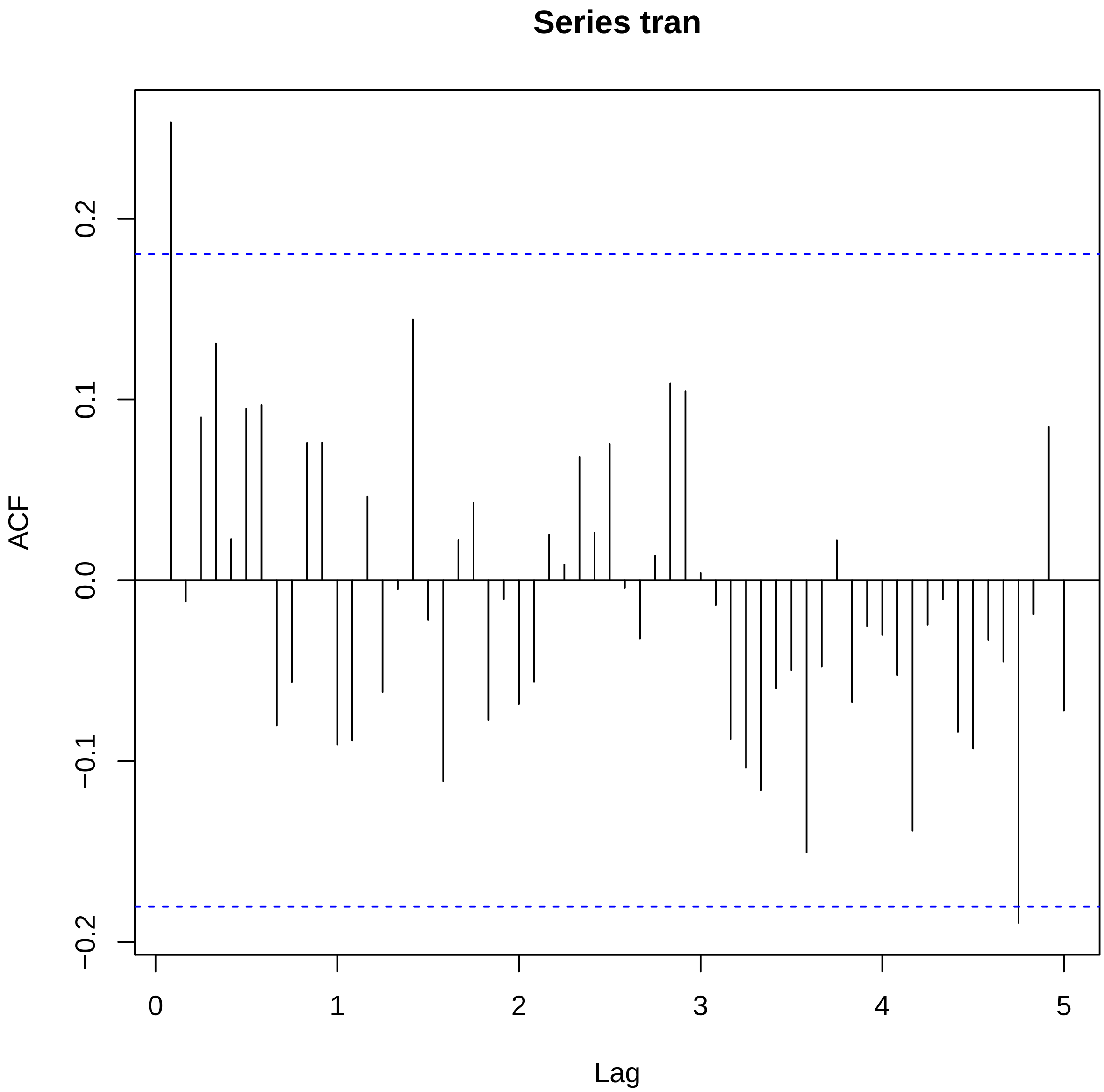
At this point we have succefully transformed the data near enough to stationarity to work with it effectively. There is no apparent trend over time. There is no readily apparent seasonality. The varaince appears relatively constant but this is difficult to judge.

Below is a histogram and Q-Q Plot of the tranformed data. There is still some heaviness in the: see Q-Q plot.

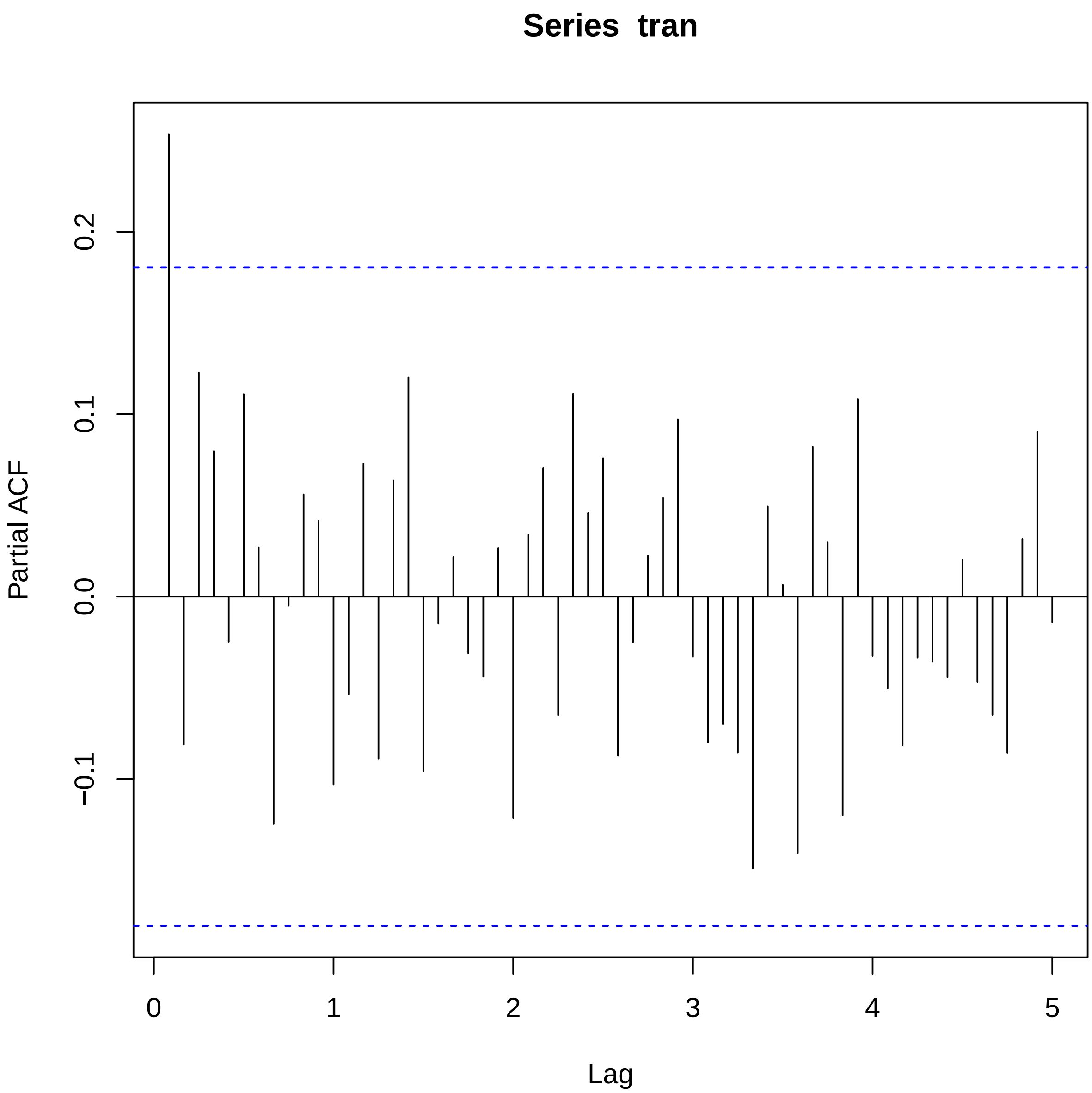




Next we look at the correlation structure of the series and begin to figure out the best value for p and q in our ARMA(p,q) model.



There is some significant positive correlation at lag 1. There is something of an alternating and tapering pattern which would suggest an AR model. However, the tapering is not consistent, so there be a MA componenet also. Note that lag is in terms of years in the figure above.



As with the ACF the PACF suggest we investigate AR(1), MA(1), and ARMA(1,1) models

As noted in the text the sample PACF and ACF do well for diagnosing AR(p) and MA(q) but are not as good for identifying the appropriate paraments form mixed ARMA models. We can use the Extended Autocorrelation function as a tiebreaker for choosing values of p and q to test.

```
eacf(tran)

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

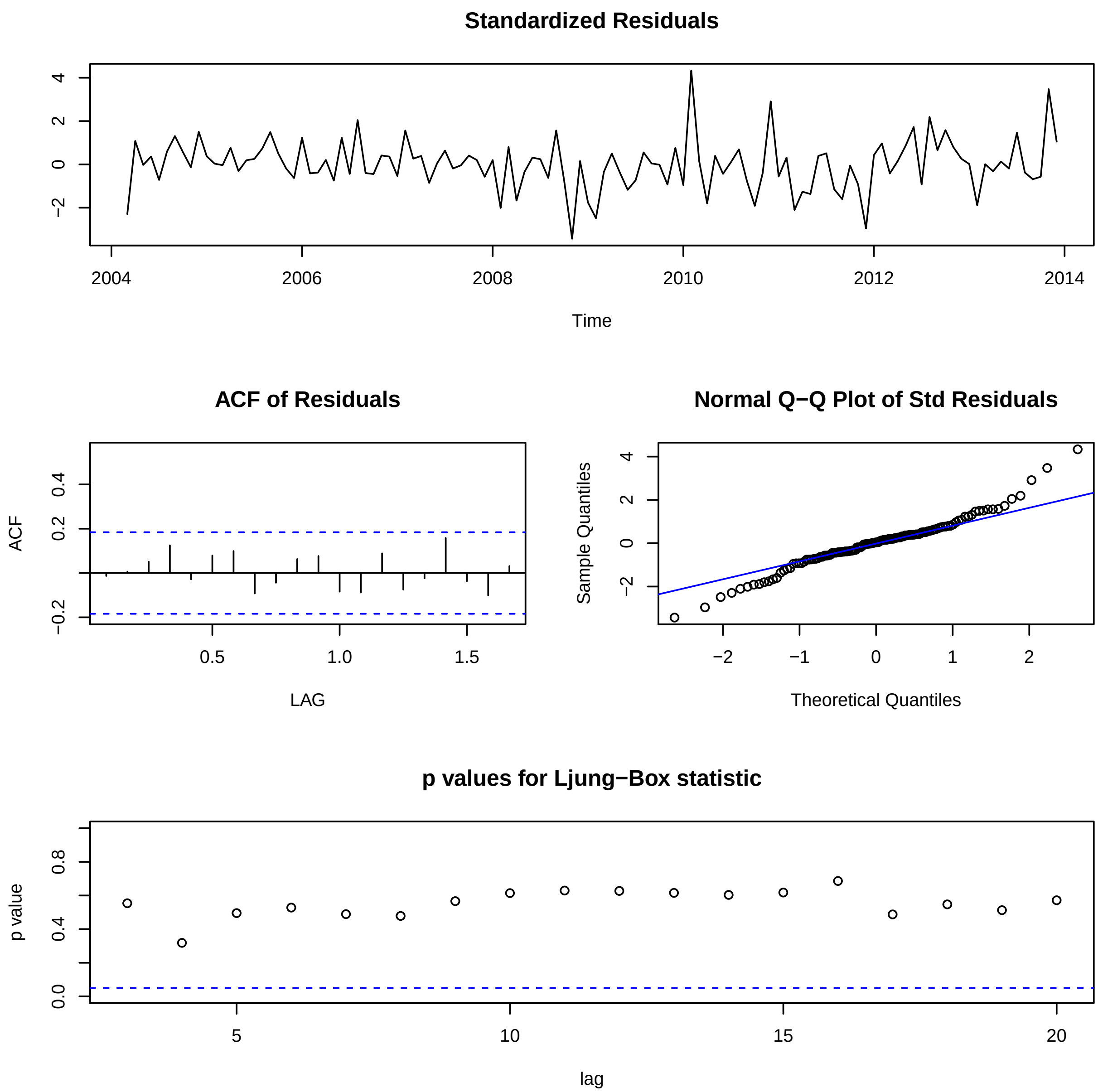
```
## 5 x x x x o o o o o o o o  o  o  o
## 6 o x o x o o o o o o o o  o  o  o
## 7 o x o o o o o o o o o o  o  o  o
```

This suggests that a ARMA(1,1) or MA(1) model would be most appropriate but we can also try the AR(1) model for comparison.

Comparison of the three models.

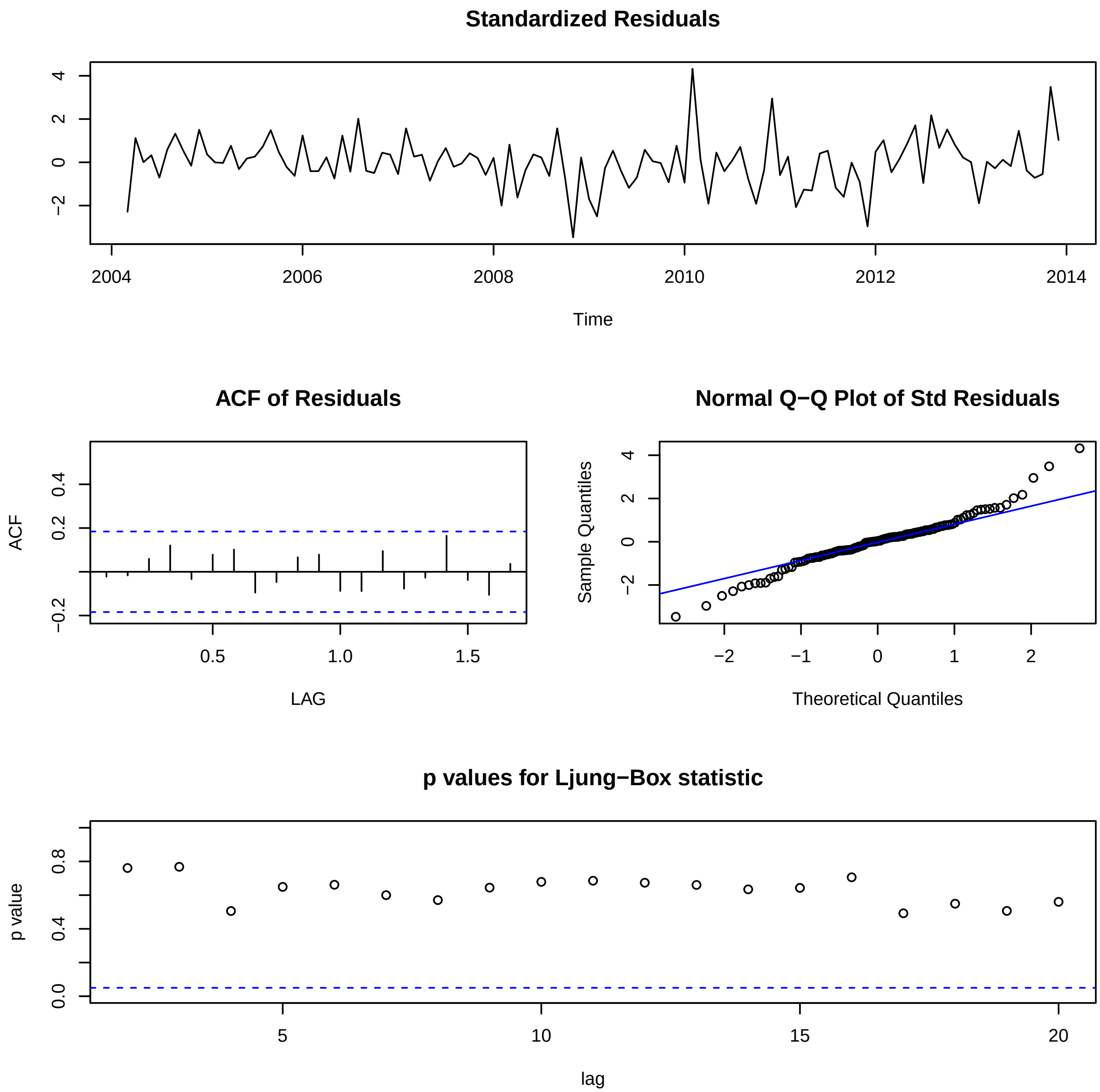
We can see in the coefficients section that the "ar1" coefficient is not significantly different from zero, so we can ignore the ARMA(1,1).

```
##
## Call:
## arima(x = tran, order = c(1, 0, 1))
##
## Coefficients:
##          ar1    ma1  intercept
##       -0.081  0.381      5.785
## s.e.   0.270  0.246      0.037
##
## sigma^2 estimated as 0.0974:  log likelihood = -30.06,  aic = 66.12
```



We can see in the coefficients section that "ma1" coefficient is significantly differnet from zero and the ACF of the residuals looks good.

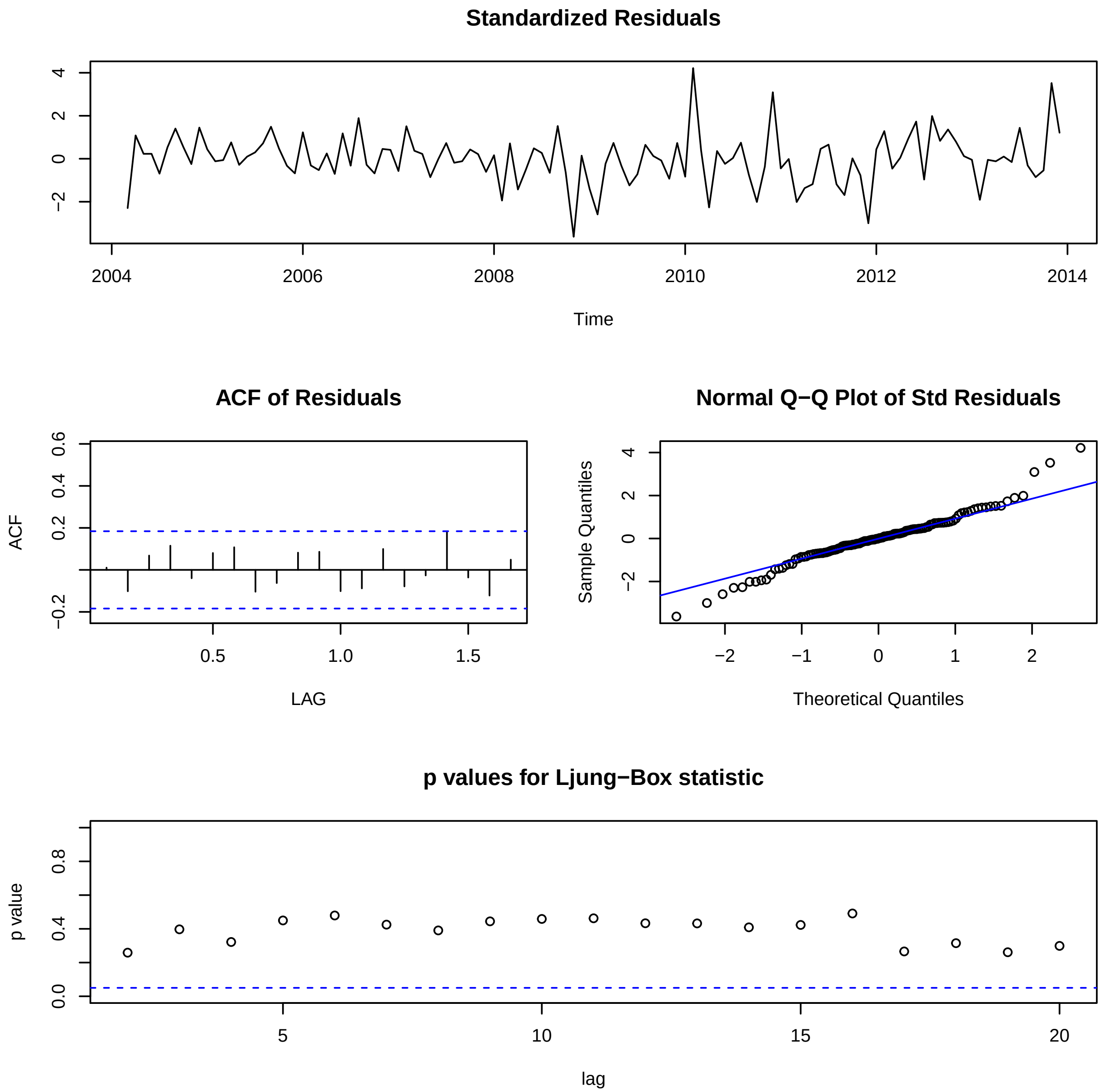
```
##
## Call:
## arima(x = tran, order = c(0, 0, 1))
##
## Coefficients:
##          ma1  intercept
##        0.309      5.785
## s.e.  0.092      0.038
##
## sigma^2 estimated as 0.0974:  log likelihood = -30.1,  aic = 64.2
```



We can see in the coefficients section that "ar1" coefficient is significantly differnet from zero and the ACF of the residuals loks good also. So we have the "ar1" model and the "ma1" model. We could pick based on the model with the least information loss or we could compare the predictive power of the two and use that.

```
##
## Call:
## arima(x = tran, order = c(1, 0, 0))
##
## Coefficients:
##          ar1  intercept
```

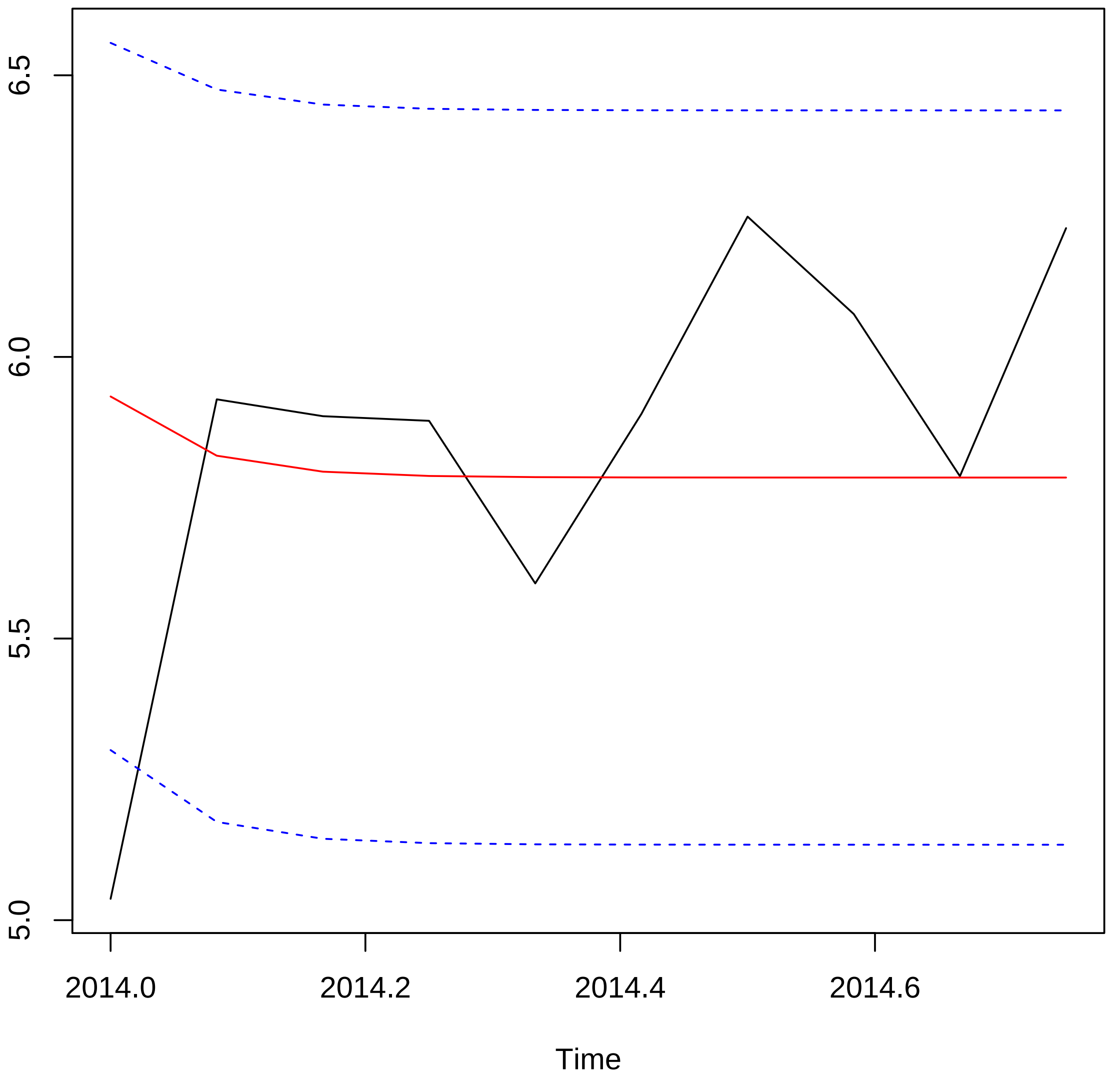
```
##        0.269      5.786
## s.e.  0.091      0.039
##
## sigma^2 estimated as 0.0985:  log likelihood = -30.74,  aic = 65.49
```



The following is the prediction for 2014 compared with the actual data that covers January 2014 through October 2014.This first plot is the actual (transformed) data for 2014 in black overlayed with the AR(1) prediction in red and the standard error in dashed blue lines.

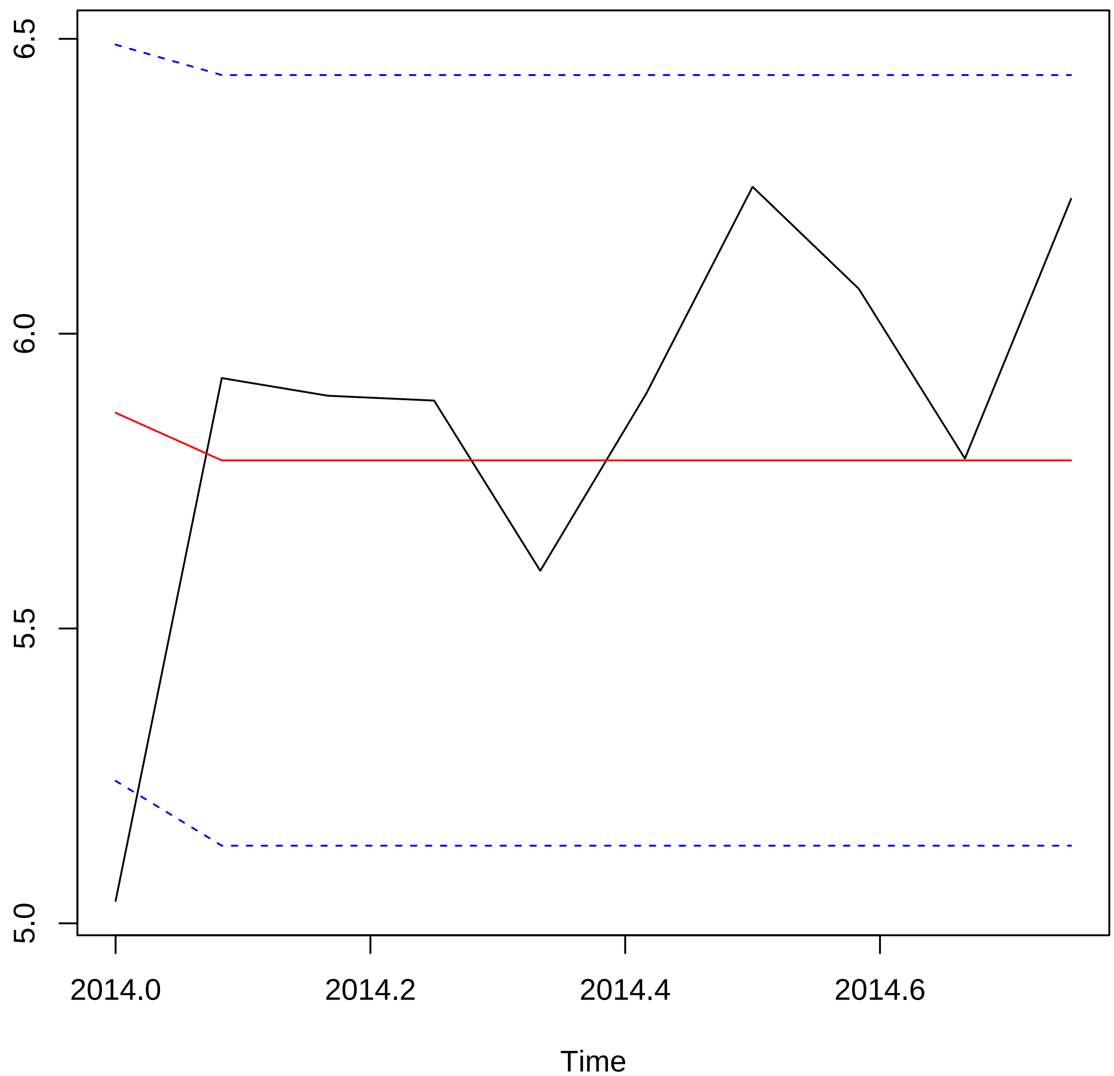
The number shown below is the MSE of the data versus the fitted model.

```
## [1] 0.1368
```

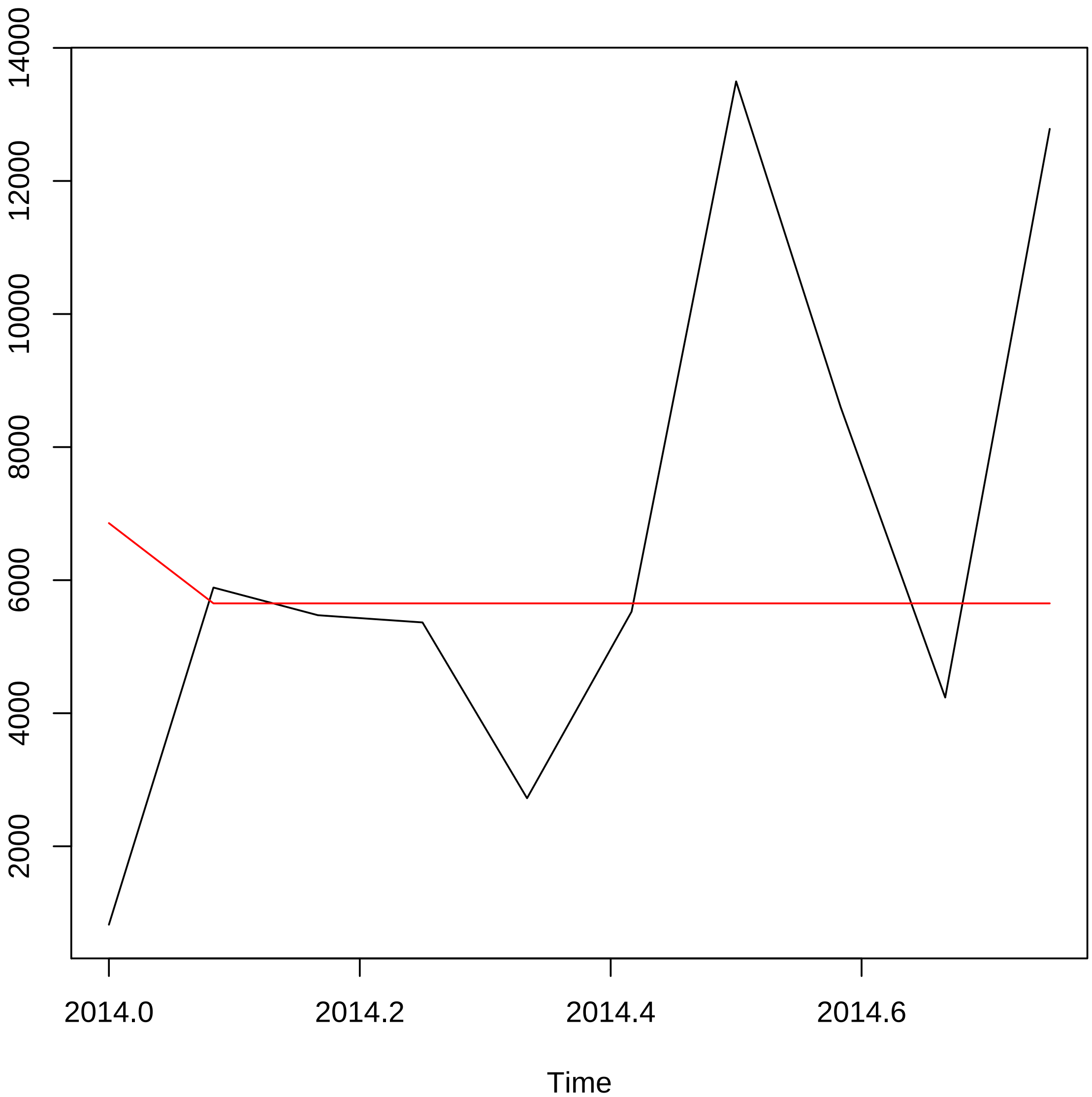


Below is the same type of plot but for the MA(1) model. The number shown below is the MSE of the data versus the fitted model.

```
## [1] 0.1272
```



The MA(1) model had the lower information loss and lower MSE. Now a plot of the original data vs the untransformed fitted model.



Conclusion

Based on my analysis I would use the MA(1) to forecast claim severity. Based on the lower AIC and MSE on the test data, it appears to be a better fit. A similar method could be used on this data to get the predictions for 2015. We would not necessarily end up using a MA(1) again but the same methods could be used to arrive at an efficient model.