# VEE Forecasting Project

February 17,

2012

*Candidate: Muhamed Umer Majeed*
*Email: umermajeed@hotmail.com*
*Session: Fall 2011*

# TABLE OF CONTENTS

## 1.0 Introduction

Tourism is a popular global activity and has become vital for many countries due to the large sums of money being spent on goods and services. Canada depends heavily on tourism and consequently forms a large industry that heavily impacts its economy.

Since Canada shares the largest undefended border in the world with the US, a large portion of tourists visiting Canada are Americans. That being said, it would be very beneficial for Canadians to be able to forecast the number of tourists entering Canada from the US on a monthly basis. This prediction could aid Canadian economy, ensuring that adequate resources are available to accommodate the US visitors in the upcoming months.

I decided to use a dataset from Statistics Canada[1], which encapsulates the number of entrants per month coming from the US into Canada, who stay overnight. I obtained a little over 35 years of data starting from 1972 up until April, 2007. To ensure I modeled the data correctly I have defined a training set from January 1972 till December 2005 and have identified my testing set as that from January 2006 up until April 2007.

Looking at an initial plot (see Section 2.1 below) of the data we see that there is an obvious seasonal component with some sort of trend. Possible models to account for these observances are causal, autoregressive integrated moving average (ARIMA) and seasonal autoregressive integrated moving average (SARIMA).

---

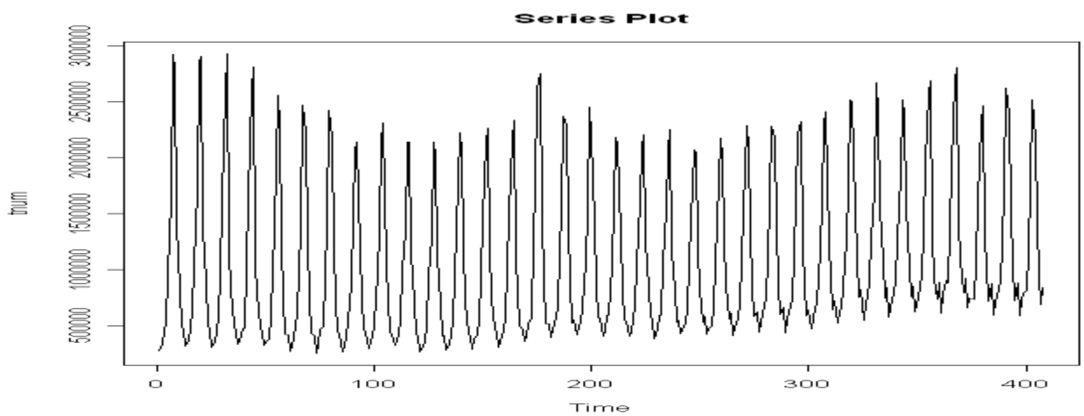
[1] http://www.StatCan.ca

# 2.0 Detailed Results

## *2.1 Data Transformation*

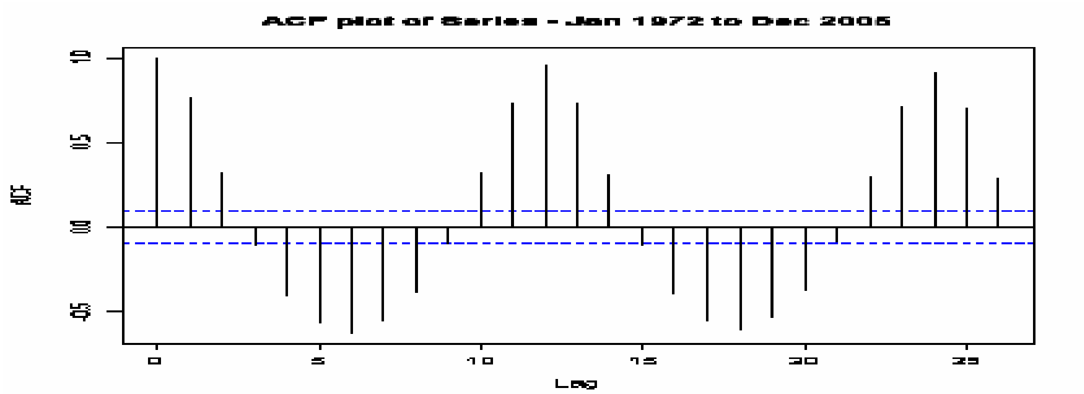
Looking at a time series plot (below) we see that the data is clearly non-stationary; with clear trends and strong seasonality. The ACF plot (see Figure 1B below) confirms this notion of a cyclical trend and seasonality.

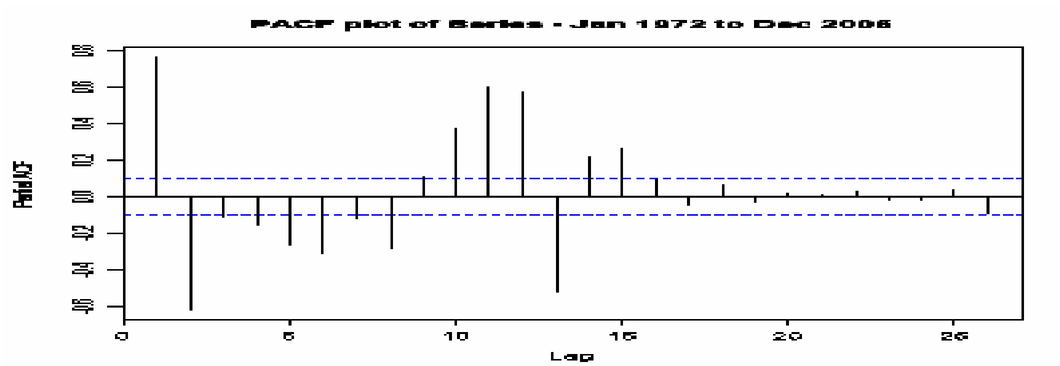
**<u>Figure 1:</u>** The time series plot of the number of US entrants into Canada per month




Time series plot of Number of Tourists from January 1972 – December 2005

**<u>Figure 1B</u>:** An ACF and PACF plot of data from January 1972 – December 2005

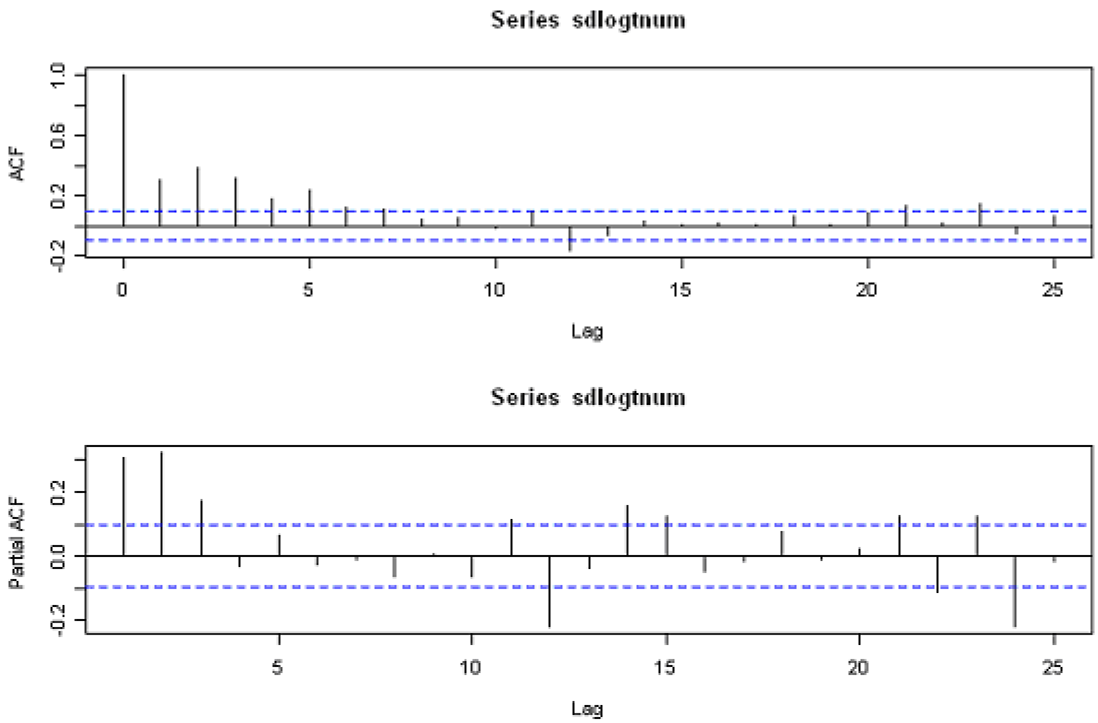



To remove the heteroscedasticity from the data, I applied 2 transformations, the log and square root (see Figure 2A in appendix A). The transformations were not enough to remove the strong seasonality in the data, leading me to perform seasonal differencing on the data at lag 12. Looking at the seasonally differenced and transformed time series plots (see Figure 2B in Appendix A), the square root-transformed tourist visits appear heteroscedastic. In contrast, the seasonally differenced log-transformed tourist visits are almost homoscedastic. This led me to continue my investigation with the log-transformed data.

Looking at the ACF plot (see Figure 3 below) of log-transformed tourist data, I noticed that the seasonal differencing removes a lot of seasonality, but there still exist some minor trend in the data.

**Figure 3:** ACF and PACF plots of seasonally differenced and logged data

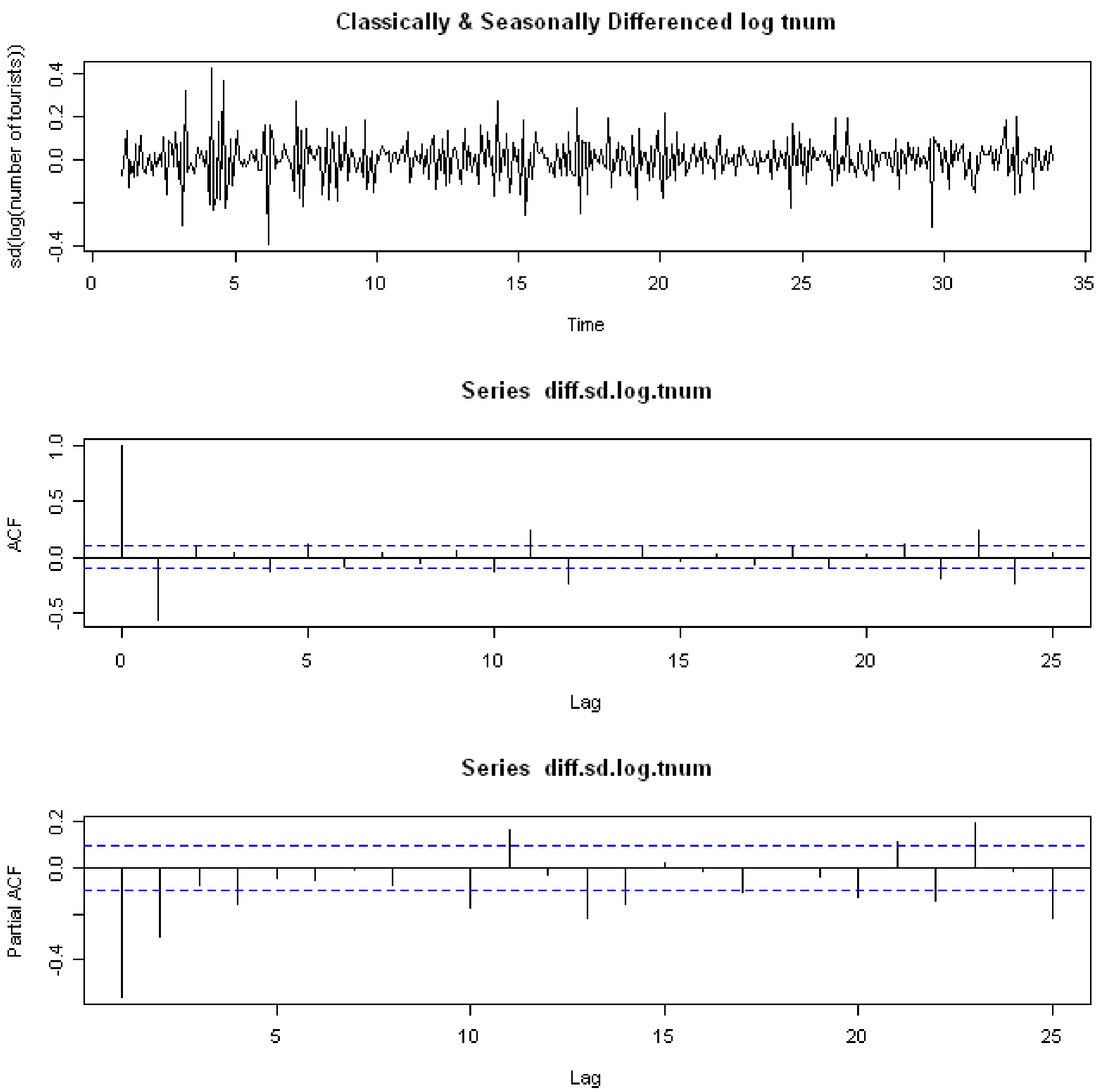



However, I felt confident that after further application of ordinary (classical) differencing to the seasonal log-transformed data, I could select an appropriate model to represent our dataset. I was not wrong in my judgment and did, in fact, find a model that fits the data well, as demonstrated in the next section.

## *2.2 Model Selecting*

To determine the various candidates for the best model, we look at the ACF and PACF plots of classical and seasonal log-transformed data (see Figure 4a below). There is significant correlation at lag 1 in the ACF plot, while the PACF plot shows significant correlation until the $4^{th}$ lag. There are a few other significant correlations at higher lags, however they are likely due to sampling error. The observations made from the ACF and the PACF plots lead us to select p = 4 and q = 1 for the model.

**Figure 4a:** Time series plot, ACF, and PACF of classically and seasonally differenced logged data




We also look at the seasonal log-transformed data (see Figure 4b below). There is significant correlation until lag 5 in the ACF plot. In contrast, the PACF plot shows significant correlation at lags 1, 2, and 3. We look at p = 3, and q = 5 in our model, but we also try p = 5 and q = 3 to account for sampling error.

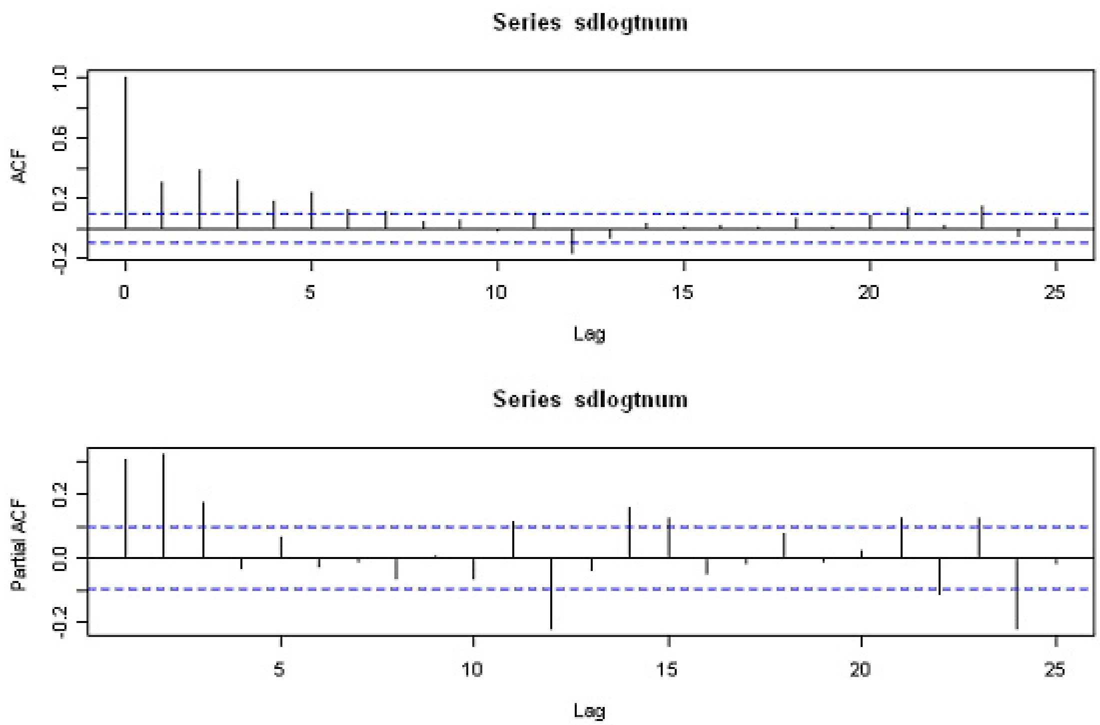



**Figure 4b:** Time series plot, ACF, and PACF of classically and seasonally differenced logged data

To determine the best model, we look at various ARIMA (p, 1, q) and examine the AIC and the $\sigma^2$ of the models, selecting the model with the lowest AIC. We look at a few models - seasonal and non-seasonal – in an attempt to fit our data. Some of the seasonal models that I looked at include:

- SARIMA (4,1,1) * (3,1,5)
- SARIMA (4,1,1) * (5,1,3)
- SARIMA (4,1,1) * (4,1,3)
- SARIMA (4,1,1) * (3,1,4)
- SARIMA (4,1,1) * (6,1,3)
- SARIMA (4,1,1) * (5,1,4)

Below is a summary of the models:

| Model | Seasonal | Classical & Seasonal | AIC | $\sigma^2$ | Log likelihood |
|---|---|---|---|---|---|
| 1 | (4,1,1) | (3,1,5) | -1028.38 | 0.003916 | 528.19 |
| **2** | **(4,1,1)** | **(5,1,3)** | **-1071.08** | **0.003367** | **549.54** |
| 3 | (4,1,1) | (4,1,3) | -1068.07 | 0.00342 | 547.03 |
| 4 | (4,1,1) | (3,1,4) | -1052.49 | 0.003607 | 539.25 |
| 5 | (4,1,1) | (6,1,3) | -1058.28 | 0.003497 | 544,14 |
| 6 | (4,1,1) | (5,1,4) | -1083.45 | 0.003156 | 556.73 |

Based on the lowest AIC and $\sigma^2$, we chose SARIMA (4,1,1) * (5,1,3). Next, I conducted a residual analysis to test out the model.

### *2.3 Model Testing*

SARIMA (4,1,1) * (5,1,3) is selected as the final model to forecast the number of tourists. It is important to test the model residuals for homoscedasticity, correlation and normality.

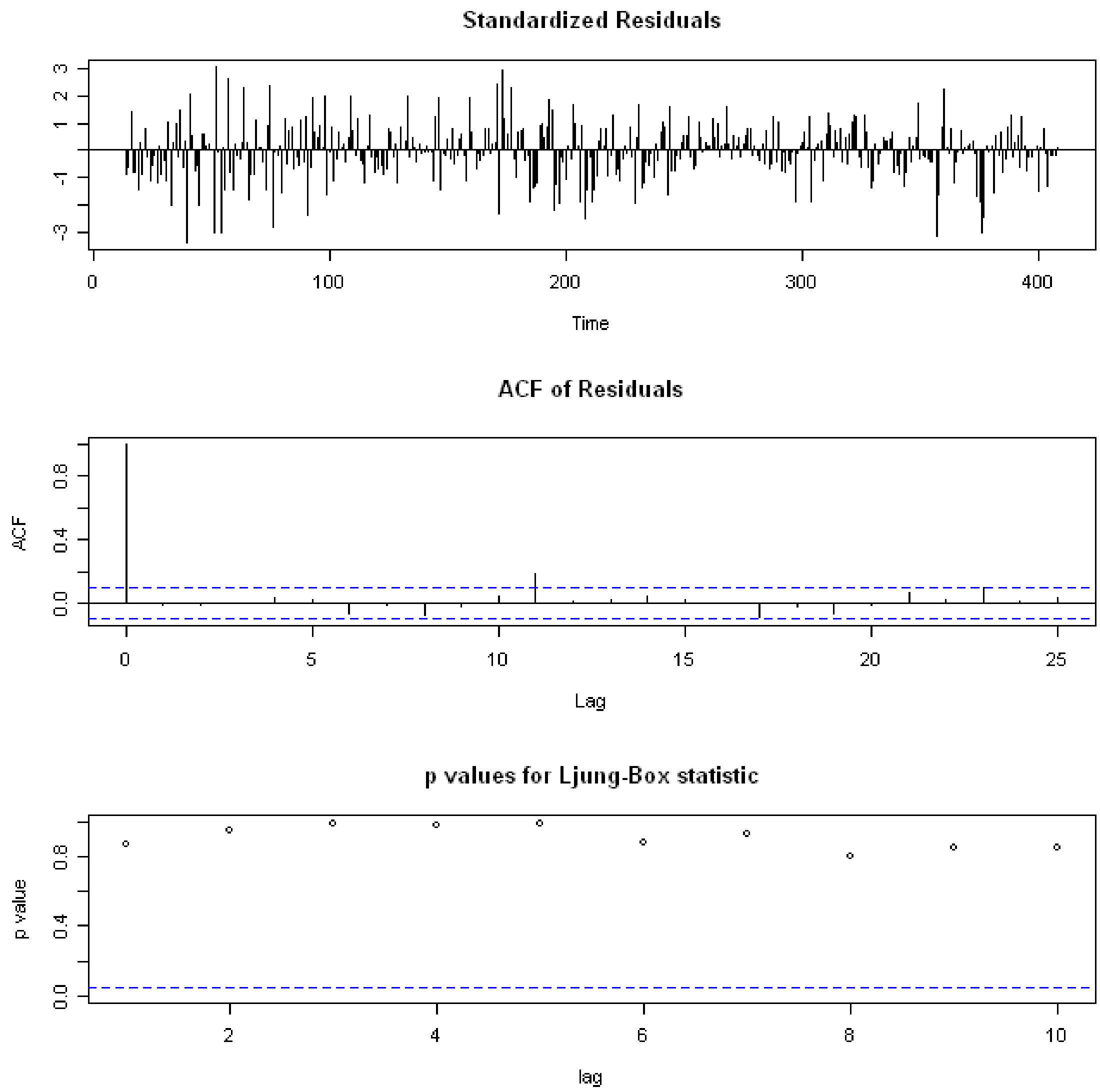



**Figure 6**: Tsdiag for SARIMA (4,1,1) * (5,1,3)

Now, looking at the Ljung-Box (see Figure 6 above), we see that none of the p-values are significant enough to reject the normality hypothesis. Meanwhile, the ACF of the residuals shows no significant correlations, except at lag 11, we regard this as error, and the standardized residuals chart shows no clear trends.  These charts suggest that the residuals are normally distributed and independent. The result of the Shapiro-Wilk test, with the SW = 0.9843, and the QQ-plot (see Figure 7 in Appendix A) also confirms the normality assumptions.

Therefore, we can say with confidence that our model meets the OLS assumptions.

# 3.0 Conclusion

After conducting a residual analysis of our model, SARIMA (4,1,1) * (5,1,3), I can conclude that the model chosen is a good fit for the data. The model residuals, being white noise, meet all the OLS assumptions. I used the model to predict the number of entries per month from the US into Canada, staying overnight, over the testing period from January 2006 to April 2007. The table below includes the true monthly tourist values over the testing period, as well as the predicted ones. We notice that the predicted values of the number of entries from our model are very close estimates of the true values from the data. The true values from the data lie within the 95% prediction interval of our predicted values. Therefore, we can conclude that the model used is very capable of predicting future values.

| **Month** | **Value** (1) | **Predicted Value** (2) | **Error** \|(1) – (2)\| | **95% Prediction Interval** | **Correct Prediction within 95%** |
|---|---|---|---|---|---|
| Jan 2006 | 555,207.0 | 565,847.3 | 10,640.30 | (505,047.8 , 633,966.0) | Yes |
| Feb 2006 | 632,462.0 | 684,700.6 | 52,238.60 | (604,939.1 , 774,978.7) | Yes |
| Mar 2006 | 722,212.0 | 764,204.2 | 41,992.20 | (667,117.3 , 875,420.3) | Yes |
| Apr 2006 | 831,343.0 | 777,010.6 | 54,332.40 | (672,631.0 , 897,587.9) | Yes |
| May 2006 | 1,149,851.0 | 1,076,711.5 | 73,139.50 | (928,812.8 , 1,248,160.8) | Yes |
| Jun 2006 | 1,725,335.0 | 1,677,050.7 | 48,284.30 | (1,441,490.1 , 1,951,195.3) | Yes |
| Jul 2006 | 2,294,960.0 | 2,285,811.7 | 9,148.30 | (1,959,646.5 , 2,666,264.0) | Yes |
| Aug 2006 | 2,111,749.0 | 2,210,867.4 | 99,118.40 | (1,890,954.0 , 2,584,904.2) | Yes |
| Sep 2006 | 1,373,629.0 | 1,236,838.7 | 136,790.30 | (1,055,342.9 | Yes |

| | | | | | |
|---|---|---|---|---|---|
| | | | | , 1,449,547.8) | |
| Oct 2006 | 923,754.0 | 862,292.1 | 61,461.90 | (734,083.9 , 1,012,891.9) | Yes |
| Nov 2006 | 681,779.0 | 654,746.9 | 27,032.10 | (556,106.7 , 770,883.6) | Yes |
| Dec 2006 | 852,780.0 | 795,149.4 | 57,630.60 | (673,795.1 , 938,360.2) | Yes |
| Jan 2007 | 526,722.0 | 537,090.1 | 10,368.10 | (448,761.7 , 642,803.8) | Yes |
| Feb 2007 | 586,835.0 | 657,175.8 | 70,340.80 | (545,926.5 , 791,095.6) | Yes |
| Mar 2007 | 672,289.0 | 764,075.9 | 91,786.90 | (630,786.3 , 925,530.5) | Yes |
| Apr 2007 | 733,374.0 | 751,871.8 | 18,497.80 | (617,374.8 , 915,669.4) | Yes |

## 4.0 General Discussion

In the tourism industry, weather and economy have significant impacts when forecasting. As Canada's tourism industry largely relies on the US, the industry is also largely affected by American economic conditions. Some major economic shocks that have shaped the tourism industry in the past include:

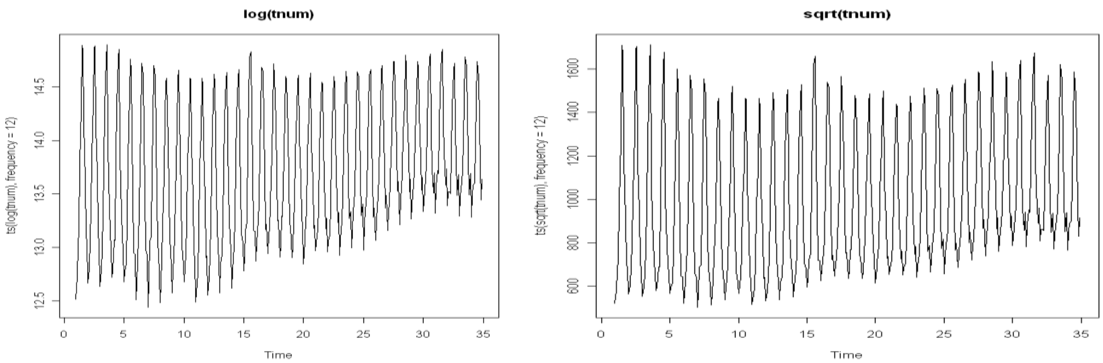
- Policy changes for border crossing (passport requirements)
- The terrorist attacks of 9/11
- The US Dollar exchange rate

With mix of unpredictable and predictable variables, as mentioned above, a forecasting in the tourism industry may indeed be more difficult than first imagined. This is the biggest limitation of our model. Even though the data accounts for these occurrences with spikes and dips, there is really no way quantify shocks previously observed. Therefore we cannot forecast tourist numbers with absolute certainty taking into account current economic conditions.

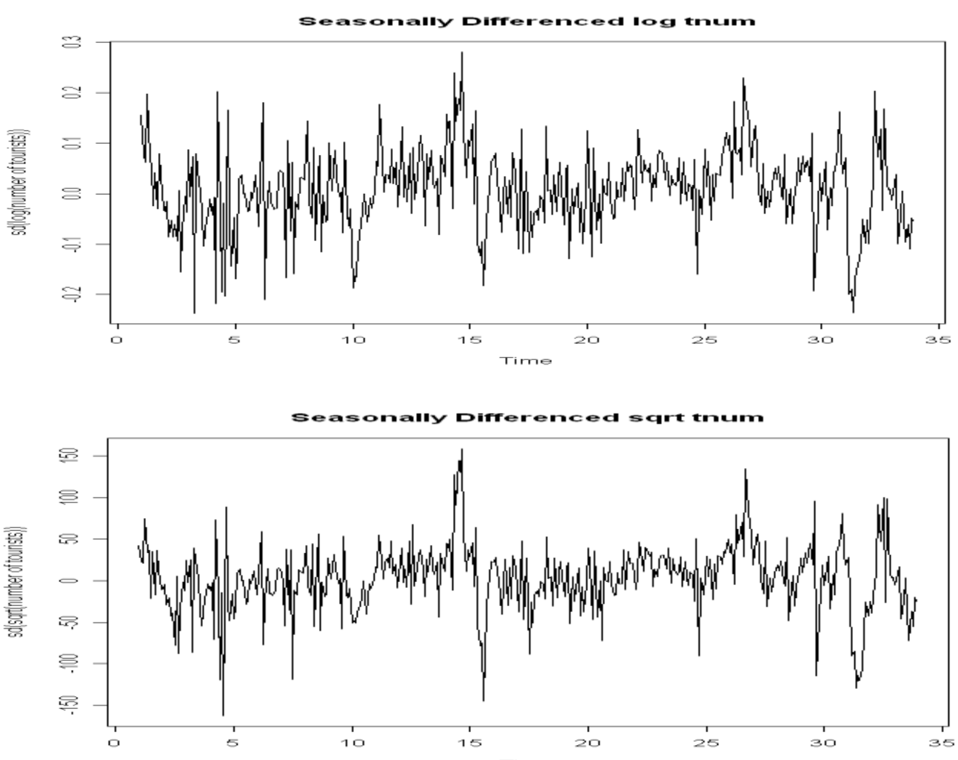
# Appendices

## *Appendix A: List of Figures*

**<u>Figure 2A</u>:** A time series plot the transformation of our data (Log and Square Root)




**<u>Figure 2B:</u>** A Time series plot of seasonally differenced and logged & seasonally differenced and square rooted data



**<u>Figure 5A</u>:** Residual Plots

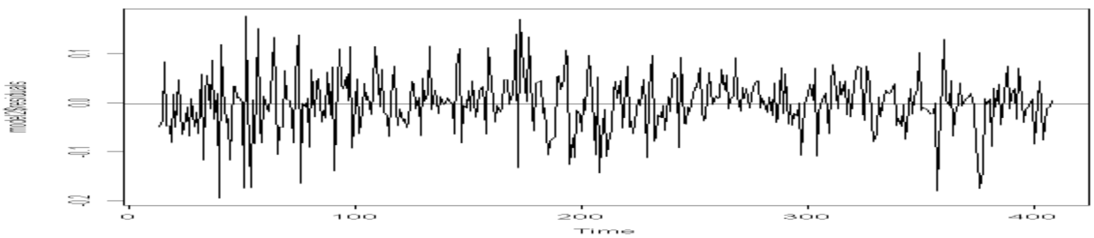


**Figure 5B:** ACF, PACF of residuals and Runs Test for SARIMA (4,1,1) * (5,1,3)

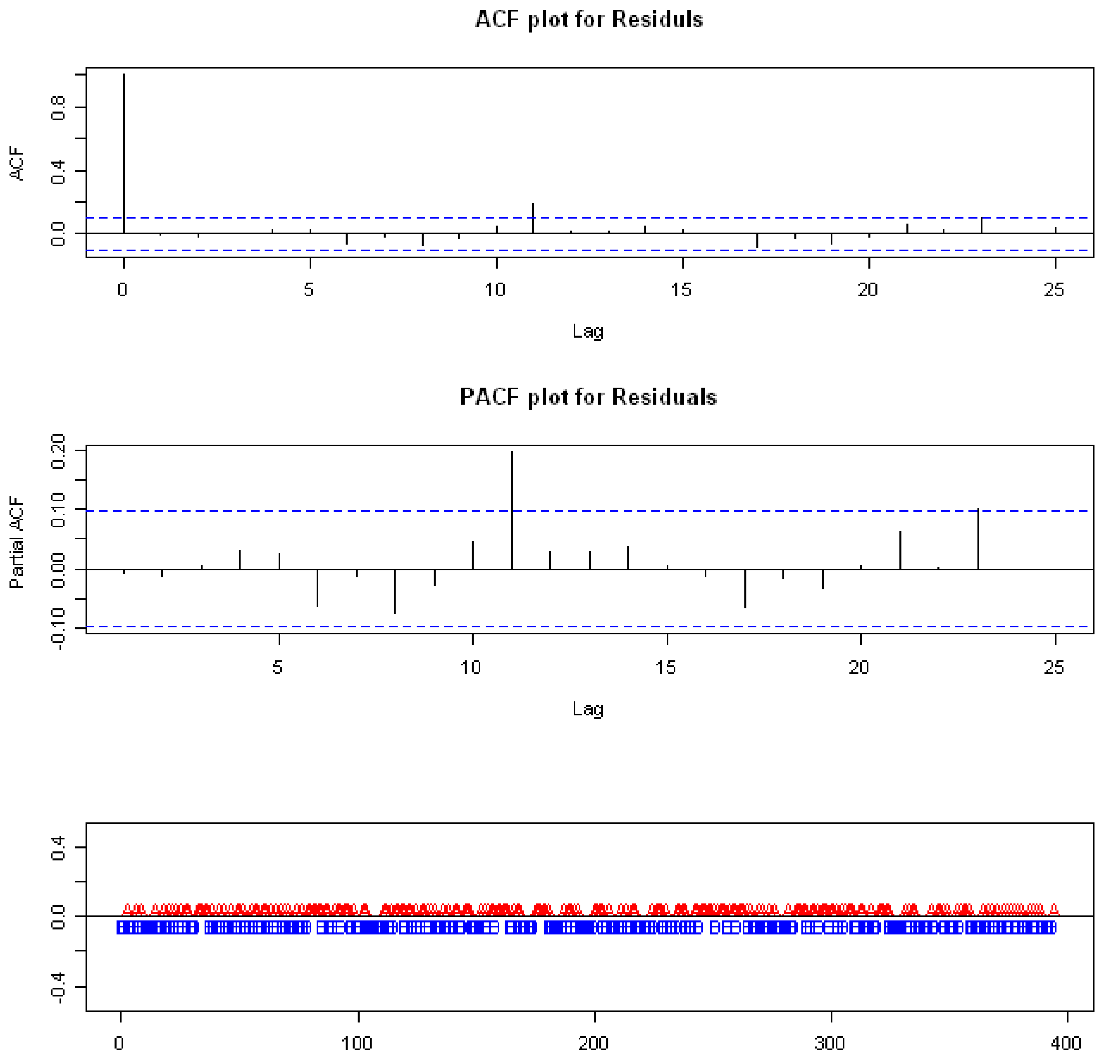
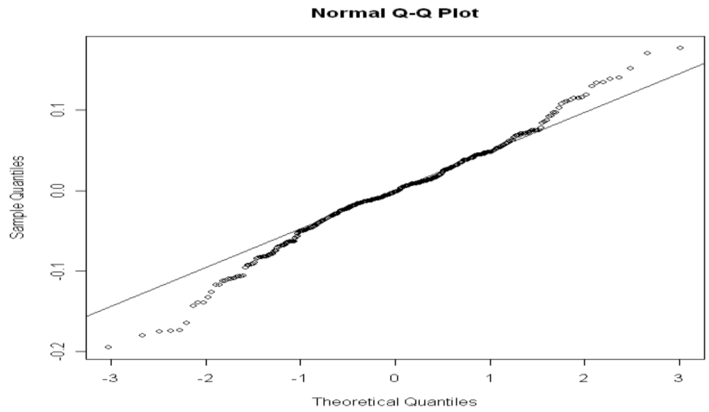


**Figure 7:** QQ plot for SARIMA (4,1,1) * (5,1,3)




Runs Test

```
> runs.test(model2$residuals)
$R
[1] 188

$E
[1] 198.4987

$z
[1] -1.057847

$p.value.1t
[1] 0.1450625
```

Shapiro-Wilks Test

```
> shapiro.test(model2$residuals)

        Shapiro-Wilk normality test

data:  model2$residuals
W = 0.9843, p-value = 0.0002742
```

# *Appendix B: Outputs for Different Models Considered*

## MODEL 1

```
> model1

Call:
arima0(x = log(tnum), order = c(4, 1, 1), seasonal = list(order = c(3, 1, 5),
    period = 12))

Coefficients:
         ar1     ar2     ar3      ar4      ma1     sar1     sar2     sar3     sma1    sma2    sma3     sma4    sma5
      0.1962  0.2744  0.1201  -0.1018  -0.8855  -0.0361  -0.3850  -0.2031  -0.4937  0.1839  0.0568  -0.1019  0.2380
s.e.  0.0524  0.0607  0.0572   0.0565   0.2189   0.0552   0.0349   0.0396   0.0050  0.0315  0.0197   0.0385  0.0068

sigma^2 estimated as 0.003916:  log likelihood = 528.19,  aic = -1028.38
```

## MODEL 2

```
> model2

Call:
arima0(x = log(tnum), order = c(4, 1, 1), seasonal = list(order = c(5, 1, 3),
    period = 12))

Coefficients:
         ar1     ar2    ar3      ar4    ma1     sar1     sar2     sar3     sar4     sar5     sma1    sma2    sma3
      0.3027  0.2305  0.048  -0.1117  -0.87  -0.1740  -0.6577  -0.5542  -0.4884  -0.0366  -0.4563  0.4355  0.2953
s.e.  0.0061  0.0194    NaN      NaN    NaN   0.0244   0.0134   0.0142   0.0243   0.0185   0.0036  0.0056  0.0067

sigma^2 estimated as 0.003367:  log likelihood = 549.54,  aic = -1071.08
Warning message:
In sqrt(diag(x$var.coef)) : NaNs produced
```

## MODEL 3

```
> model3

Call:
arima0(x = log(tnum), order = c(4, 1, 1), seasonal = list(order = c(4, 1, 3),
    period = 12))

Coefficients:
         ar1     ar2     ar3      ar4      ma1     sar1     sar2     sar3     sar4     sma1    sma2    sma3
      0.2732  0.2251  0.0644  -0.1022  -0.8620  -0.2738  -0.5971  -0.6104  -0.5092  -0.3133  0.3382  0.3915
s.e.  0.0504  0.0600  0.0555   0.0535   0.2148   0.0355   0.0316   0.0312      NaN   0.0008  0.0009  0.0031

sigma^2 estimated as 0.00342:  log likelihood = 547.03,  aic = -1068.07
Warning message:
In sqrt(diag(x$var.coef)) : NaNs produced
```

## MODEL 4

```
> model4

Call:
arima0(x = log(tnum), order = c(4, 1, 1), seasonal = list(order = c(3, 1, 4),
    period = 12))

Coefficients:
         ar1     ar2     ar3      ar4      ma1    sar1     sar2     sar3     sma1    sma2    sma3     sma4
      0.2558  0.2426  0.0629  -0.0968  -0.8733  0.2100  -0.3932  -0.6055  -0.7696  0.3565  0.4495  -0.2679
s.e.  0.0524  0.0630  0.0567   0.0540   0.2494  0.0093   0.0357   0.0005   0.0305  0.0175  0.0120   0.0295

sigma^2 estimated as 0.003607:  log likelihood = 539.25,  aic = -1052.49
```

## MODEL 5

```
> model5

Call:
arima0(x = log(tnum), order = c(4, 1, 1), seasonal = list(order = c(6, 1, 3),
    period = 12))

Coefficients:
         ar1     ar2     ar3      ar4      ma1     sar1    sar2    sar3    sar4    sar5    sar6     sma1     sma2
      0.2148  0.2299  0.0883  -0.0959  -0.8520  -0.2199  0.0787  0.0310  0.0339  0.3691  0.4149  -0.3752  -0.3992
s.e.  0.0446  0.0233  0.0210   0.0186   0.1133   0.0231     NaN  0.0134     NaN  0.0302  0.0731   0.0573   0.0499
        sma3
      0.0429
s.e.  0.0503

sigma^2 estimated as 0.003497:  log likelihood = 544.14,  aic = -1058.28
Warning message:
In sqrt(diag(x$var.coef)) : NaNs produced
```

## MODEL 6

```
> model6

Call:
arima0(x = log(tnum), order = c(4, 1, 1), seasonal = list(order = c(5, 1, 4),
    period = 12))

Coefficients:
         ar1     ar2     ar3      ar4      ma1     sar1     sar2     sar3     sar4     sar5    sma1    sma2
      0.3143  0.2513  0.0733  -0.1072  -0.8900  -0.6370  -0.8266  -0.4389  -0.9000  -0.1500  0.0825  0.3063
s.e.  0.0519  0.0581  0.0569   0.0556   0.2315   0.0563   0.0559   0.0598   0.0308   0.0585  0.0287  0.0377
         sma3    sma4
      -0.0905  0.6751
s.e.   0.0216  0.0217

sigma^2 estimated as 0.003156:  log likelihood = 556.73,  aic = -1083.45
```

# *Appendix C: Prediction*

```
> prediction2
Time Series:
Start = 409
End = 432
Frequency = 1
 [1]  565847.3  684700.6  764204.2  777010.6 1076711.5 1677050.7 2285811.7 2210867.4 1236838.7
[10]  862292.1  654746.9  795149.4  537090.1  657175.8  764075.9  751871.8 1087091.6 1669711.7
[19] 2253932.9 2292375.7 1182953.2  874877.3  647369.1  780357.6
> lowerbound2 <- exp(predict24$pred - 1.96*predict24$se)
> lowerbound2
Time Series:
Start = 409
End = 432
Frequency = 1
 [1]  505047.8  604939.1  667117.3  672631.0  928812.8 1441490.1 1959646.5 1890954.0 1055342.9
[10]  734083.9  556106.7  673795.1  448761.7  545926.5  630786.3  617374.8  889105.1 1360248.0
[19] 1829698.5 1854583.3  953807.8  703086.2  518550.4  623054.6
> upperbound2<-exp(predict24$pred + 1.96*predict24$se)
> upperbound2
Time Series:
Start = 409
End = 432
Frequency = 1
 [1]  633966.0  774978.7  875420.3  897587.9 1248160.8 1951105.3 2666264.0 2584904.2 1449547.8
[10] 1012891.9  770883.6  938360.2  642803.8  791095.6  925530.5  915669.4 1329165.8 2049580.1
[19] 2776530.3 2833513.1 1467149.0 1088643.7  808189.1  977375.0
```

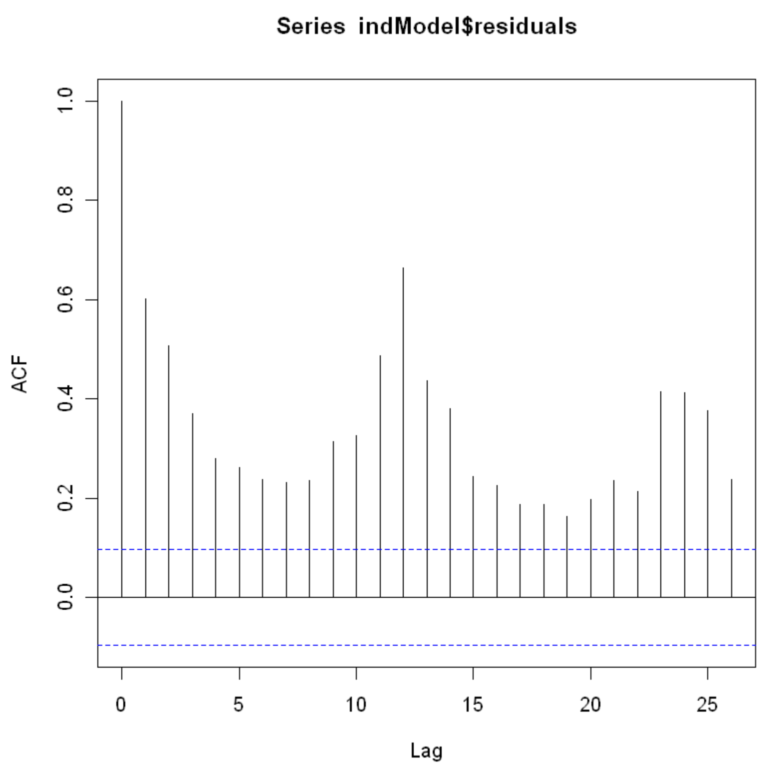
## *Appendix D: Regression Models*

**Regression Models**

***Indicator Variable Model with a Time Component***

```
regData<-read.csv("updatedData.csv", header = TRUE)

MONTH<-MONTH[1:408]
YEAR<-YEAR[1:408]

TIME<-seq(1,408,1)

JAN<-MONTH==1
FEB<-MONTH==2
MAR<-MONTH==3
APR<-MONTH==4
MAY<-MONTH==5
JUN<-MONTH==6
JUL<-MONTH==7
AUG<-MONTH==8
SEP<-MONTH==9
OCT<-MONTH==10
NOV<-MONTH==11
DEC<-MONTH==12

indModel<-
lm(NUM~JAN+FEB+MAR+APR+MAY+JUN+JUL+AUG+SEP+OCT+NOV+TIME)

summary(indModel)

Call:
lm(formula = NUM ~ JAN + FEB + MAR + APR + MAY + JUN + JUL +
   AUG + SEP + OCT + NOV + TIME)

Residuals:
   Min      1Q  Median      3Q     Max
-384519  -67208  -20142   54981  699398

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  434876.21   27057.05  16.073  < 2e-16 ***
JANTRUE     -177446.65   34025.40  -5.215 2.97e-07 ***
FEBTRUE      -98419.81   34024.32  -2.893  0.00403 **
MARTRUE      -27090.57   34023.35  -0.796  0.42637
APRTRUE       80204.65   34022.48   2.357  0.01889 *
```

```
MAYTRUE      493908.07   34021.71  14.517  < 2e-16 ***
JUNTRUE      991328.61   34021.05  29.139  < 2e-16 ***
JULTRUE     1794789.92   34020.49  52.756  < 2e-16 ***
AUGTRUE     1768556.75   34020.03  51.986  < 2e-16 ***
SEPTRUE      705350.56   34019.67  20.734  < 2e-16 ***
OCTTRUE      266969.80   34019.41   7.848 4.02e-14 ***
NOVTRUE      -28095.22   34019.26  -0.826  0.40938
TIME            738.40      58.98  12.519  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 140300 on 395 degrees of freedom
Multiple R-Squared: 0.9602,     Adjusted R-squared: 0.959
F-statistic: 794.8 on 12 and 395 DF,  p-value: < 2.2e-16

acf(indModel$residuals)
```

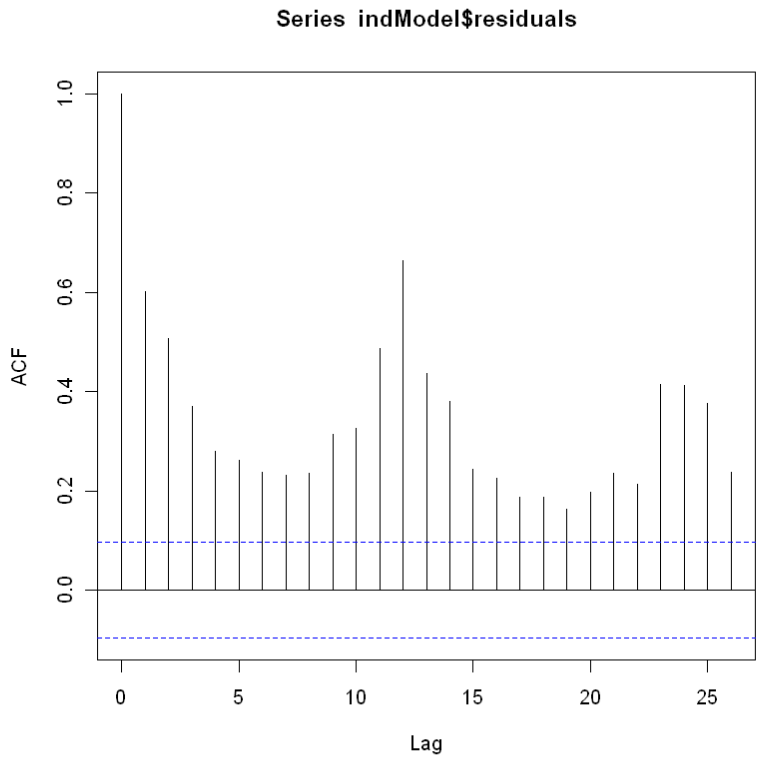



While the indicator regression model has good explanatory properties (R-Squared is high) and most of the indicator variables are significant, the residual standard error is very high, 140300 compared to 0.003367 for our chosen model2. Meanwhile, the acf plot of the residuals shows that there is a high positive correlation between the error terms at all lags.

***Indicator Model with a Year Component***

```
indModel<-
lm(NUM~JAN+FEB+MAR+APR+MAY+JUN+JUL+AUG+SEP+OCT+NOV+YEAR)
summary(indModel)

Call:
lm(formula = NUM ~ JAN + FEB + MAR + APR + MAY + JUN + JUL +
   AUG + SEP + OCT + NOV + YEAR)

Residuals:
   Min      1Q  Median      3Q     Max
-384519  -67208  -20142   54981  699398

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.703e+07  1.408e+06 -12.098  < 2e-16 ***
JANTRUE     -1.856e+05  3.402e+04  -5.455 8.66e-08 ***
FEBTRUE     -1.058e+05  3.402e+04  -3.110  0.00201 **
MARTRUE     -3.374e+04  3.402e+04  -0.992  0.32196
APRTRUE      7.430e+04  3.402e+04   2.184  0.02955 *
MAYTRUE      4.887e+05  3.402e+04  14.367  < 2e-16 ***
JUNTRUE      9.869e+05  3.402e+04  29.010  < 2e-16 ***
JULTRUE      1.791e+06  3.402e+04  52.650  < 2e-16 ***
AUGTRUE      1.766e+06  3.402e+04  51.900  < 2e-16 ***
SEPTRUE      7.031e+05  3.402e+04  20.669  < 2e-16 ***
OCTTRUE      2.655e+05  3.402e+04   7.804 5.42e-14 ***
NOVTRUE     -2.883e+04  3.402e+04  -0.848  0.39719
YEAR         8.861e+03  7.078e+02  12.519  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 140300 on 395 degrees of freedom
Multiple R-Squared: 0.9602,     Adjusted R-squared: 0.959
F-statistic: 794.8 on 12 and 395 DF,  p-value: < 2.2e-16
```

acf(indModel$residuals)




Again, while this variation of the indicator model has good explanatory properties, it is no better than the first regression model and still has a very high standard error. The residuals also seem to be highly correlated.

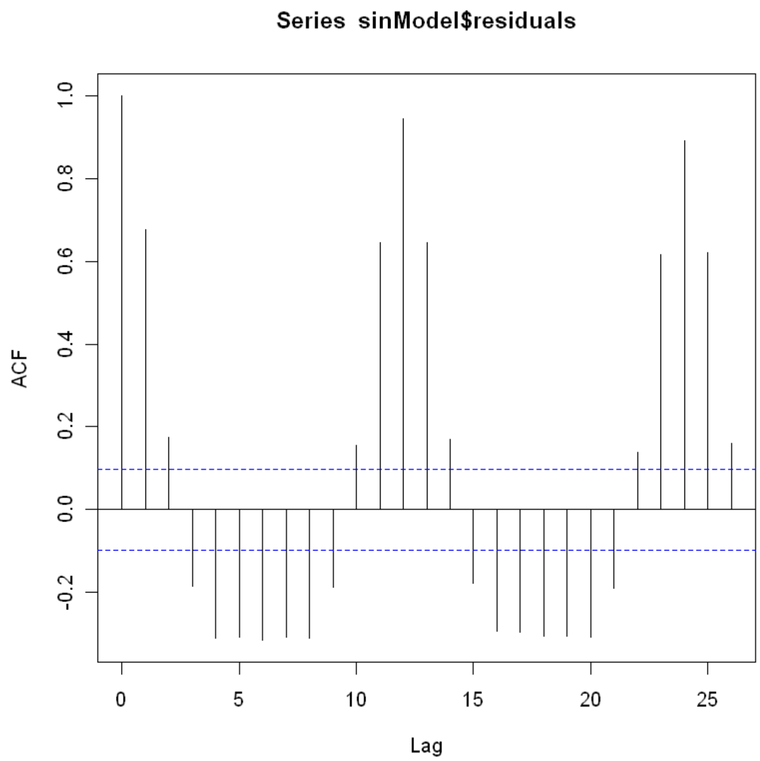
***Sinosoidal Model***

```
cos1<-cos(2*pi*MONTH/12)
sin1<-cos(2*pi*MONTH/12)
sinModel<-lm(NUM~cos1+sin1)
summary(sinModel)

Call:
lm(formula = NUM ~ cos1 + sin1)

Residuals:
    Min      1Q  Median      3Q     Max
-925240 -377157   -2416  239518 1527486

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1066717      25218   42.30   <2e-16 ***
cos1         -664513      35663  -18.63   <2e-16 ***
```

sin1 NA NA NA NA
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 509400 on 406 degrees of freedom
Multiple R-Squared: 0.461, Adjusted R-squared: 0.4596
F-statistic: 347.2 on 1 and 406 DF, p-value: < 2.2e-16

acf(sinModel$residuals)




The sinusoidal model proves to be both a poor fit and has a high standard error. Moreover, the residuals seem to be highly correlated and seem to display a seasonal pattern.