Time Series Analysis on the Number of Taiwanese Visitors Travelling to USA

Introduction

The scientific question motivating my work is: **"through time series analysis, can I find factors or leading indicators that affect the number of Taiwanese people travelling to the United States?"** The relations between Taiwan, China, and the United States have always been one of the most important matters Taiwanese citizens care about. Though the official relations between the government of Taiwan and the federal government of the United States terminated due to the recognition of Beijing since 1979, Taiwan and the U.S. have maintained unofficial relations ever since. Because of the historic background and geographic location of Taiwan, most our nationals are pro-USA. Therefore, when it comes to pursuing higher education, vacation, or business, the United States has often been put at a high priority in Taiwanese people's minds. Hence, I am interested to find out if there are other factors that affect the number of Taiwanese people travelling to the U.S. from 1996-2015.

About Data

This monthly international visitation data are collected and reported from the National Travel and Tourism Office's (NTTO) Visitor Arrivals Program (I-94 Record) from the U.S.[1] The oldest data reported is from January 1996, and the most updated data collected is from April 2015. Thus, the total dataset I am analyzing contains 232 monthly figures of Taiwanese

[1] http://travel.trade.gov/research/monthly/arrivals/

nationals visiting the U.S. from January 1996 till April 2015. In addition, I collected another time series of monthly Consumer Price Index (CPI) in Taiwan[2], from January 1996 till April 2015. I calculated the percentage change of CPI for each month and later I would study the relationship between the number of visitors and CPI percentage change. A summary statistics for both time series is shown in figure 1.

```
> summary(cpi)
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-0.0190600 -0.0038000  0.0011100  0.0008745  0.0061340  0.0380900
> summary(visit)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   9451   20160   27090   28410   33520   63400
```

Figure 1

Exploratory Data Analysis

As always, a careful inspection of the time series plot is the first step. Figure 2 shows clear seasonality. June, July, and August always have the highest values, which respond to the summer vacation. In addition, for students going abroad for summer school or graduate studies, July and August are the months when they usually travel. Note that there are other spikes at January or February. These months are Chinese New Year, which usually take place in the late January or early February, according to Chinese lunar calendar.

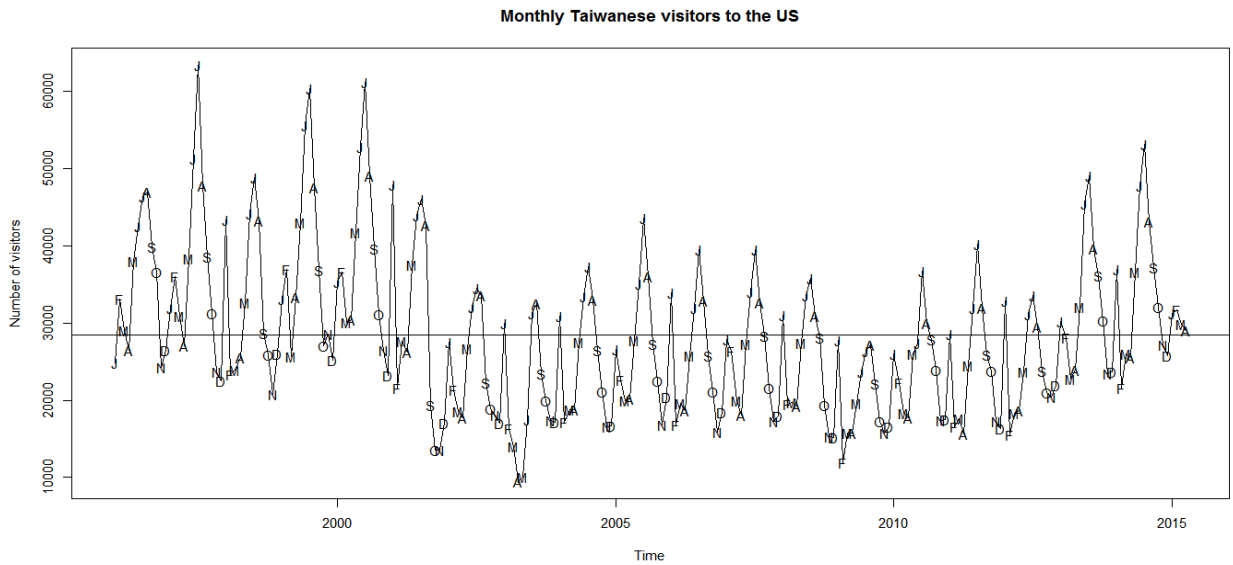

Figure 2

[2] National Statistics, Republic of China (Taiwan): http://www.stat.gov.tw/np.asp?ctNode=485

Next, I decomposed the series by loess method, and saw non-stationary trend and two seasonal cycles — summer vacation and Chinese New Year. After examining the trend in figure 3 and figure 2, I found the downward turns around September 2011 and the 1st quarter of 2003 are affected by the 9/11 terrorist act and SARS (Severe Acute Respiratory Syndrome) outbreak in south Asia, respectively. Thus, intervention analysis might be needed in later model fitting.

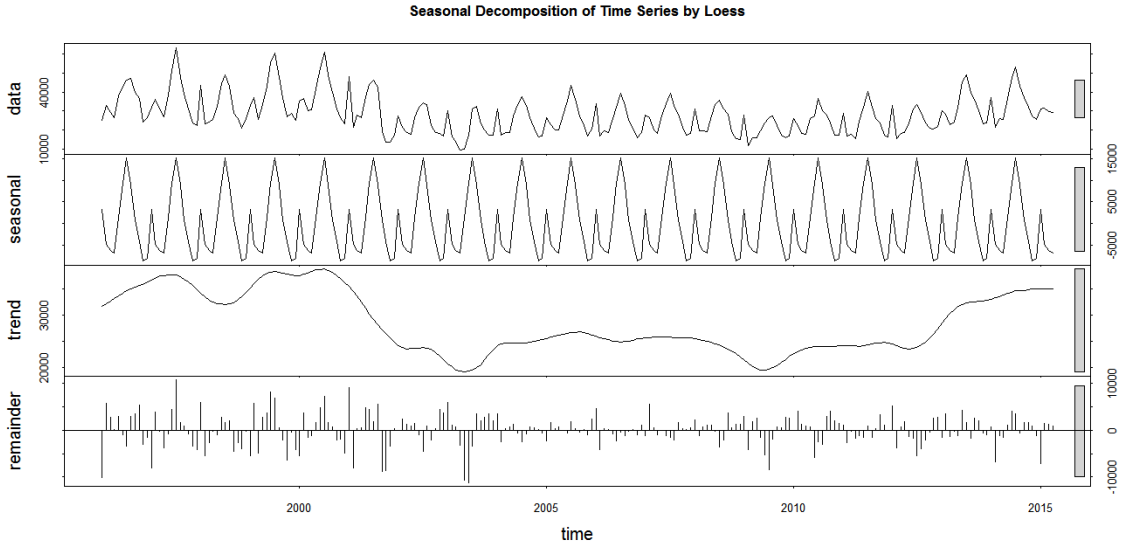

Figure 3

Lastly, to study whether CPI percentage change would be a leading indicator of visitation, I inspected the sample cross-correlation function based on the prewhitened data. Since the visitor data has many spikes and strong seasonality, I had it transformed by logarithm and took first and seasonal difference. The CCF is significant at lag 0 suggesting a strong contemporaneous positive relationship between CPI percentage change and visitors.

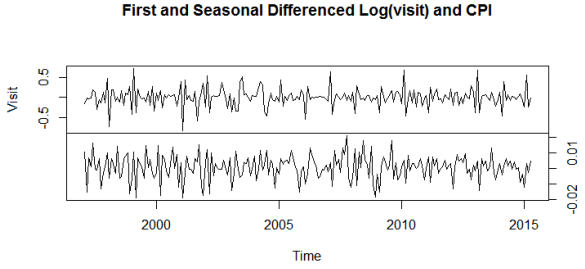

Figure 4

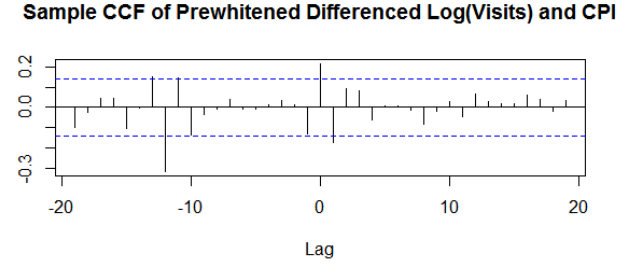

Figure 5

Model Specification

From data exploratory analysis, I leant that there were two interventions: the 911 attack and SARS outbreak. Therefore, I first specified a model for the pre-intervention data, and then incorporated 911 events, SARS, and CPI time series into the model. The ACF and PACF plots of pre-intervention data imply that there are some autoregressive order and moving-average order. Thus, I compute AICs for 36 models by changing regular orders p, q and seasonal orders P, Q ranging from 0~5. The model with the smallest AIC is specified as seasonal ARIMA $(2,1,1)x(2,1,1)_{12}$. *(The AIC calculation results is ignored, but codes are provided in the end)*

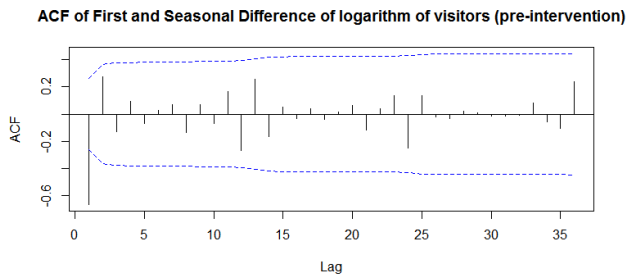

Figure 6

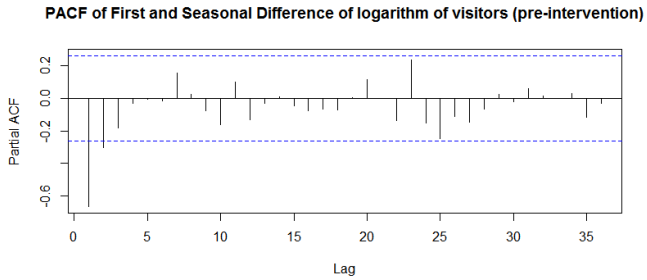

Figure 7

Next, I looked at the intervention data. The terrorist acts had lingering depressing effects on air traffic. Thus the intervention may be specified as an AR(1) process with the pulse input at September 2001. But the unexpected turn of 9/11 events had a strong instantaneous chilling effect. Thus I modeled the 9/11 effects as: $m_t = \omega_0 P_t^{(T)} + \frac{\omega_1}{1-\omega_2 B} P_t^{(T)}$ where T denotes September 2011. $\omega_0 + \omega_1$ represents the instantaneous 9/11 effect, and, for $k \geq 1$, $\omega_1 (\omega_2)^k$ gives the 9/11 effect k months afterward.

Lastly, the SARS outbreak also impacted the air traffic. Unlike 9/11 event, SARS effects were

continuous for several months, and as soon as WHO declared the SARS outbreak contained, people's travel behavior returned to normal. Therefore, I encoded SARS effect as a new regression predictor, which equals to "1" during February 2003 to June 2003 when Taiwan was affected by SARS, and "0" otherwise. In conclusion, the model was specified as an SARIMA model (2,1,1) x $(2,1,1)_{12}$ plus 9/11 effects, and plus two regression predictors: SARS and CPI.

Model Fitting

Figure 8 gives the maximum likelihood estimates of the first fitted model (`m1.visit`).

Figure 8

```
> m1.visit

Call:
arimax(x = log(visit), order = c(2, 1, 1), seasonal = list(order = c(2, 1, 1),
    period = 12), xreg = data.frame(CPI = cpi, SARS = c(rep(0, 85), rep(1, 5),
    rep(0, (232 - 90)))), method = "ML", xtransf = data.frame(I911 = 1 * (seq(visit) ==
    69)), I911 = 1 * (seq(visit) == 69)), transfer = list(c(0, 0), c(1, 0)))

Coefficients:
         ar1     ar2      ma1     sar1     sar2     sma1     CPI     SARS  I911-MA0  I911.1-AR1  I911.1-MA0
      0.0296  0.0536  -0.6606  -0.4097  -0.3509  -0.4948  3.0131  -0.5031    1.0515      0.4212     -1.4985
s.e.  0.1433  0.1059   0.1209   0.1416   0.1057   0.1680  1.0308   0.0679    0.6707      0.1672      0.6540

sigma^2 estimated as 0.01513:  log likelihood = 141,  aic = -260.01
```

Considering the magnitude of standard errors, the estimates of `ar1, ar2,` and `I911-MA0` are not significantly different from zero. Hence, a model fixing these coefficients to be zero was subsequently fitted and reported in figure 9. Note that though the log likelihood has decreased slightly, `m2.visit` model has smaller AIC than `m1.visit` model.

Figure 9

```
> m2.visit

Call:
arimax(x = log(visit), order = c(2, 1, 1), seasonal = list(order = c(2, 1, 1),
    period = 12), xreg = data.frame(CPI = cpi, SARS = c(rep(0, 85), rep(1, 5),
    rep(0, (232 - 90)))), fixed = c(0, 0, NA, NA, NA, NA, NA, NA, 0, NA, NA),
    method = "ML", xtransf = data.frame(I911 = 1 * (seq(visit) == 69), I911 = 1 *
        (seq(visit) == 69)), transfer = list(c(0, 0), c(1, 0)))

Coefficients:
      ar1  ar2      ma1     sar1     sar2     sma1     CPI     SARS  I911-MA0  I911.1-AR1  I911.1-MA0
        0    0  -0.6945  -0.6308  -0.4675  -0.2147  2.4886  -0.5103         0      0.9648     -0.5032
s.e.    0    0   0.0609   0.1906   0.1224   0.2435  1.0359   0.0628         0      0.0320      0.0816

sigma^2 estimated as 0.01556:  log likelihood = 138.56,  aic = -261.12
```

However, after some model diagnostics, I detected some innovative outliers affecting the `m2.visit` model. Thus, a third model incorporating innovative outliers was fitted and reported in figure 9.(`m3.visit`) The AIC has been greatly reduced and log likelihood has also increased. Hence, the SARIMA model $(2,1,1)x(2,1,1)_{12}$ plus 9/11 effects, plus two regression predictors: SARS and CPI, and plus some innovative outlier was carried into later model diagnostic.

Figure 9

```
> m3.visit

Call:
arimax(x = log(visit), order = c(2, 1, 1), seasonal = list(order = c(2, 1, 1),
    period = 12), xreg = data.frame(CPI = cpi, SARS = c(rep(0, 85), rep(1, 5),
    rep(0, (232 - 90)))), fixed = c(0, 0, NA, NA, NA, 0, NA, NA, NA, NA, NA,
    NA, NA, NA, 0, NA, NA), method = "ML", io = c(26, 62, 88, 89, 122, 158),
    xtransf = data.frame(I911 = 1 * (seq(visit) == 69), I911 = 1 * (seq(visit) ==
        69)), transfer = list(c(0, 0), c(1, 0)))

Coefficients:
      ar1  ar2      ma1     sar1     sar2  sma1     CPI     SARS    IO-26    IO-62    IO-88    IO-89
        0    0  -0.6689  -0.7033  -0.4858     0  1.8283  -0.3831  -0.2975  -0.3229  -0.2825  -0.5166
s.e.    0    0   0.0582   0.0696   0.0691     0  0.9002   0.0613   0.1063   0.1000   0.0942   0.0949
      IO-122   IO-158  I911-MA0  I911.1-AR1  I911.1-MA0
     -0.2985  -0.3148         0      0.9629     -0.5029
s.e.  0.1056   0.0944         0      0.0308      0.0716

sigma^2 estimated as 0.01221:  log likelihood = 166.64,  aic = -307.28
```

Model Diagnostics

Figure 10 shows three diagnostic tools in one display. The standardized residuals plot shows that all residuals are all within magnitudes +/-3. The sample ACF of the residuals shows there

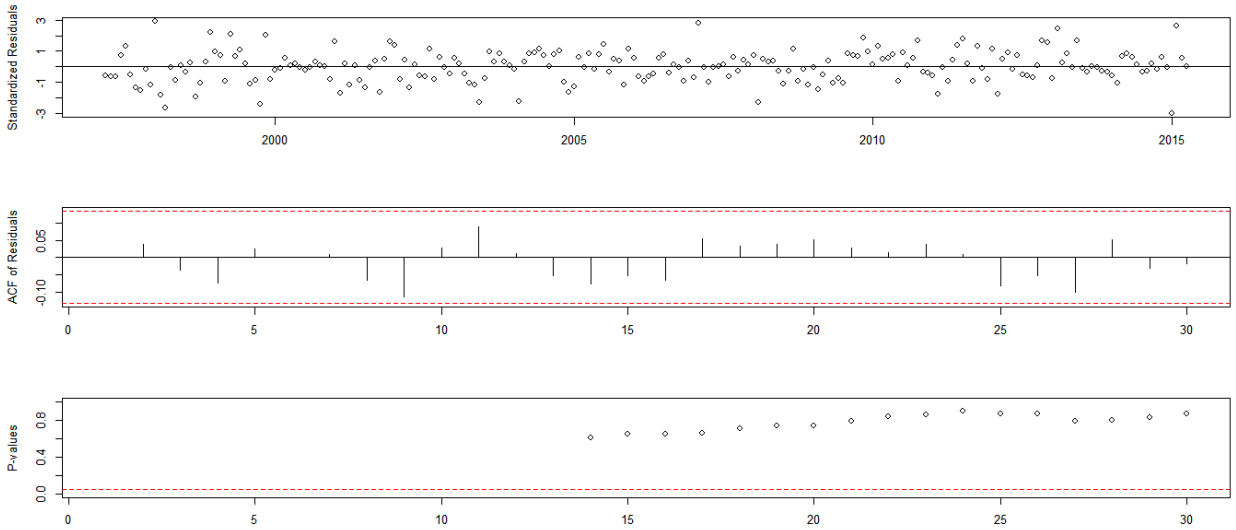

Figure 10

is no evidence of autocorrelation in the residuals of this model. Lastly, the p-value for the Ljung-Box test statistic are above 5% significant level so we have no evidence to reject the null hypothesis that the error terms are uncorrelated.

Next, I checked the normality of residuals. The QQ plot shows a straight line pattern. And the histogram is like bell-shaped. Both plots suggest the residuals are approximately normally distributed. In addition, the Shapiro-Wilk normality test applied to the residuals produces a test statistic of W = 0.98987, which corresponds to a p-value of 0.1043, and thus we would not reject normality.

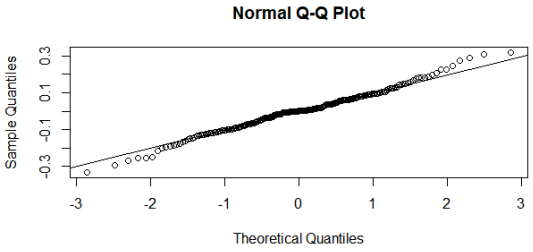

Figure 11

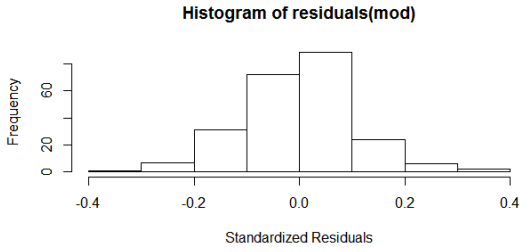

Figure 12

Lastly, I applied frequency analysis on the residuals to examine the white noise assumption. The smoothed periodogram of residuals is close to a constant line except that between frequency 0.2 and 0.3, the spectrum is smaller. However, considering the previous diagnostic results, which are all good, I would say the `m3.visit` model seem to capture the dependence structure of the number of Taiwanese visitors well. In addition, the three external factors – 9/11, SARS, and CPI all played

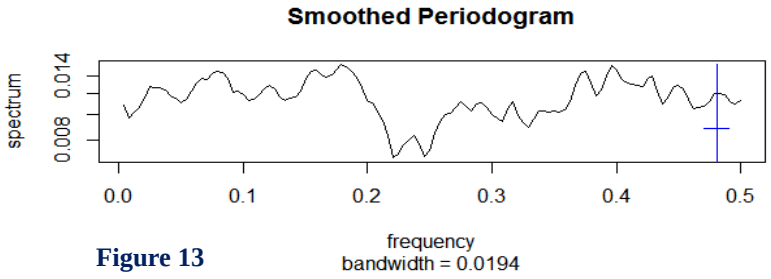

Figure 13

significant roles in model fitting.

Checking ARCH/GARCH

To examine whether there is evidence for ARCH in the residuals of our fitted SARIMA model (m3.visit), I applied McLeod.Li test. Figure 14 shows the p-value of 25 lags are all above 5% significant level, suggesting there is no evidence for ARCH. In addition, the ACF and PACF plots of the squared and absolute residuals of `m3.visit` model all suggest that the residuals are likely independently and identically distributed. Hence, there is no evidence to incorporate ARCH model into our SARIMA model.

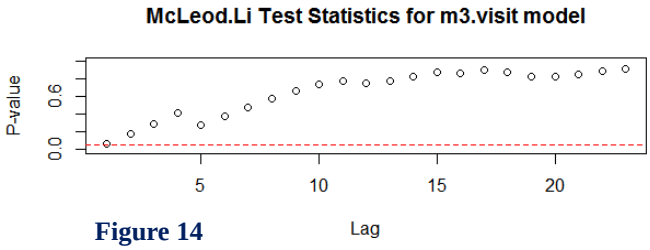

Figure 14

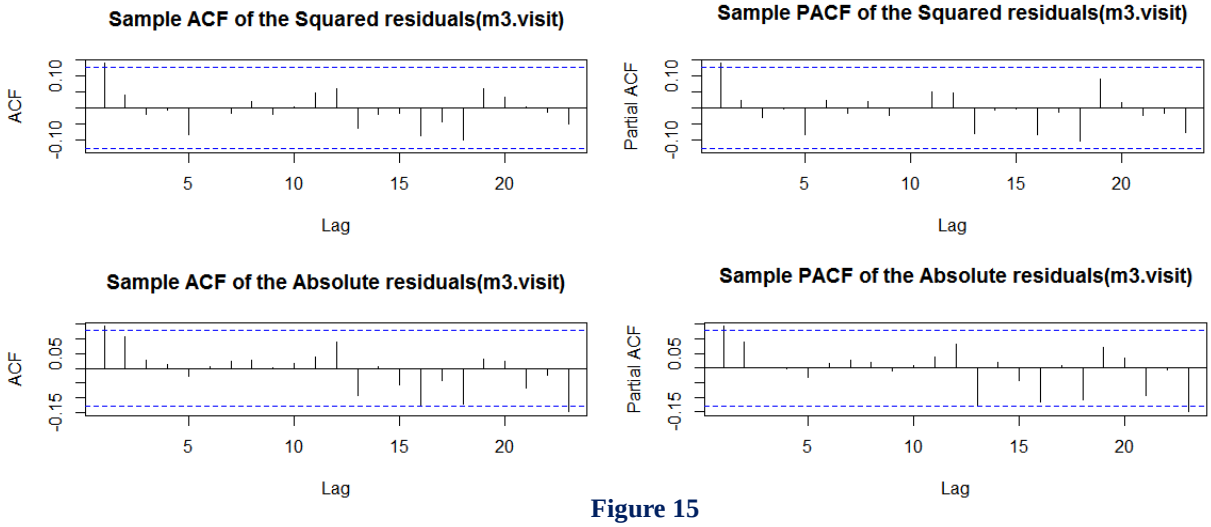

Figure 15

Conclusion

In summary, the `m3.visit` model has the smallest AIC and passed all model diagnostics, suggesting a good fit to the visitor data. Thus, the final model is denoted as $Y_t = m_t + \beta X_t + \delta$

$Z_t + N_t + IO_t$, where m_t is the change in the mean function due to 9/11 event, X_t is the regression predictor of CPI percentage change, Z_t is the regression predictor of SARS event, N_t follows a Seasonal ARIMA (2,1,1) x $(2,1,1)_{12}$ process, and IO_t denotes innovative outliers. The maximum likelihood estimates of each component is:

1. $m_t = \frac{\omega_1}{1-\omega_2 B} P_t^{(T)}$ where T denotes Sep. 2011, ω_1 = 0.9629, ω_2 = -0.5029 (Note that initially I model 9/11 intervention effect as, $m_t = \omega_0 P_t^{(T)} + \frac{\omega_1}{1-\omega_2 B} P_t^{(T)}$ but since the coefficient of ω_0 is not significantly from zero, ω_0 was fixed as zero in mode fitting. It seems that the instantaneous chilling effect for Taiwanese visitors wasn't as strong as I expected.
2. βX_t : X_t denotes the CPI percentage change at time t, β =1.8283
3. δZ_t : Z_t denotes SARS outbreak, which equals to "1" when t = February 2003 to June 2003, and equals to "0", otherwise. δ = - 0.3831
4. Nt is a SARIMA process, denotes as $Y_t = -0.7033Y_{t-12} - 0.4858Y_{t-24} + \varepsilon_t + 0.6689\,\varepsilon_{t-1}$
5. Innovative outliers are at t = 26 (Feb. 1998), 62 (Feb. 2001), 88 (April 2003), 89 (May 2003), 122 (Feb. 2006), 158 (Feb. 2009).

Note that many outliers fall in February. This might be mainly due to the effect of Chinese New Year, which usually falls in January but sometimes in February according to the Chinese lunar calendar. Since matching Chinese lunar calendar with regular calendar is an extremely complicated task, I utilized innovative outliers to deal with this variance in the analysis. This part could be further improved in future analysis.

The 9/11 intervention and regression predicators of CPI and SARS all contribute significantly to the model. The 9/11 event and SARS outbreak both have negative coefficients with the number of visitors, while CPI as positive coefficient. Therefore, we know that there were indeed external factors affecting the number of Taiwanese people travelling to the United States.

The open circles in the plot shown in figure 16 represent the fitted values from the final estimated model. They indicate generally good agreement between the model and the data.

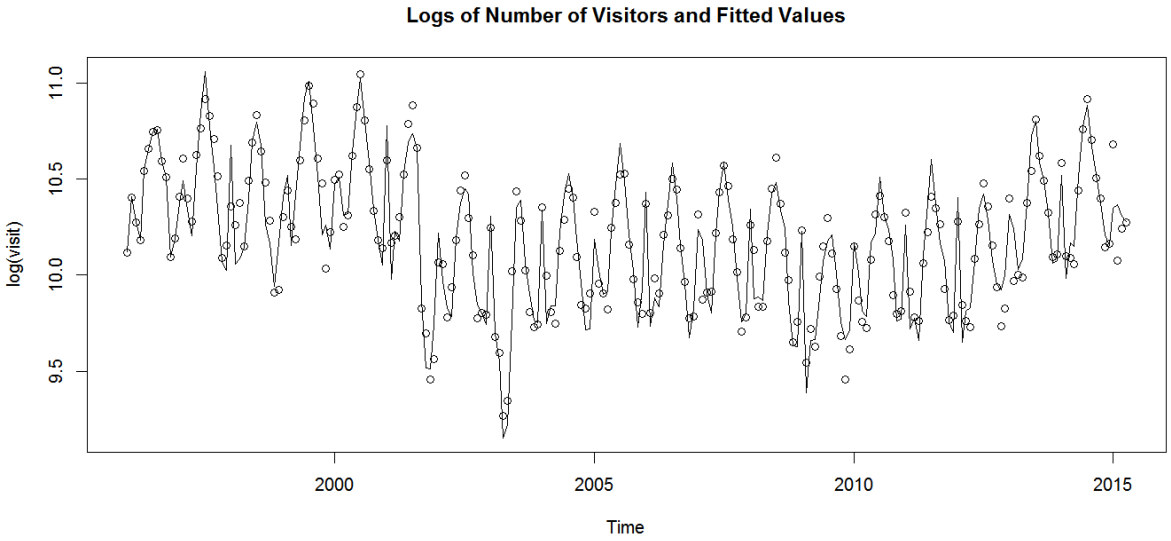

Figure 16

In summary, the answer to my question is, **"there are several external factors affecting the number of Taiwanese people travelling to the United States. The 9/11 event and SARS outbreak had negative impacts on people's travel tendency. The percentage change of CPI could be a leading indicator in predicting the number of Taiwanese visitors travelling to the U.S. There is a contemporaneous positive relationship (lag=0) between CPI percentage change and visitors."**