<방한 관광객 수에 대한 시계열 데이터 분석>

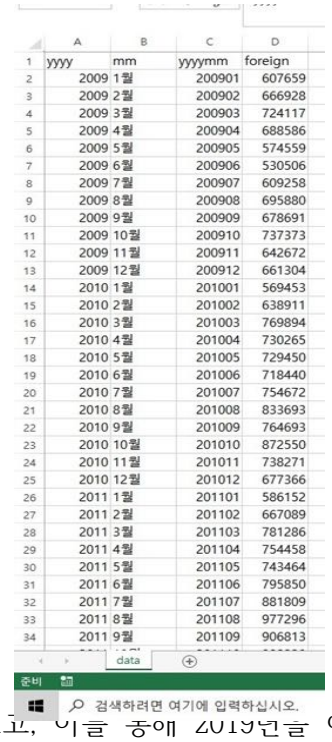
2022710230 통계학과 남 주연

<목차>

1.데이터 선택 배경, 출처

최근 2-3년 동안 코로나로 인해 해외여행이 거의 불가능 한 상황 이였다. 우리나라에서도 흔히 볼 수 있었던 외국 관광객들을 보지 못한지 오랜 시간이 흘렀다. 이로 인해 관광객 수가 급격히 줄고 이로 인한 수입이 줄어들어 특히 명동, 이태원등의 상권이 많이 줄어들었다. 이런 상황을 생각해 보니 코로나 이전에 우리나라 방문 관광객 수는 어떤 패턴을 가지고 있었는지 분석해 보고 싶다는 생각이 들어 주제를 정하게 되었다.

e-나라지표 사이트에서 년도 별 방한 관광객 수를 다운 받아 월별로 엑셀로 정리하여 데이터 분석 자료로 사용하였다.



	A	B	C	D
1	yyyy	mm	yyyymm	foreign
2	2009	1월	200901	607659
3	2009	2월	200902	666928
4	2009	3월	200903	724117
5	2009	4월	200904	688586
6	2009	5월	200905	574559
7	2009	6월	200906	530506
8	2009	7월	200907	609258
9	2009	8월	200908	695880
10	2009	9월	200909	678691
11	2009	10월	200910	737373
12	2009	11월	200911	642672
13	2009	12월	200912	661304
14	2010	1월	201001	569453
15	2010	2월	201002	638911
16	2010	3월	201003	769894
17	2010	4월	201004	730265
18	2010	5월	201005	729450
19	2010	6월	201006	718440
20	2010	7월	201007	754672
21	2010	8월	201008	833693
22	2010	9월	201009	764693
23	2010	10월	201010	872550
24	2010	11월	201011	738271
25	2010	12월	201012	677366
26	2011	1월	201101	586152
27	2011	2월	201102	667089
28	2011	3월	201103	781286
29	2011	4월	201104	754458
30	2011	5월	201105	743464
31	2011	6월	201106	795850
32	2011	7월	201107	881809
33	2011	8월	201108	977296
34	2011	9월	201109	906813

data

준비

검색하려면 여기에 입력하십시오.

2009-2018년 월별 관광객 수를 일단 분석해보고, 이를 통해 2019년을 예측해보고 실제 데이터와 비교해 본 후, 이 과정을 바탕으로 2020-2021년 코로나가 없었더라면 관광객 수가 어떠했을지 예측해 보는 방향으로 분석을 진행 하였다.

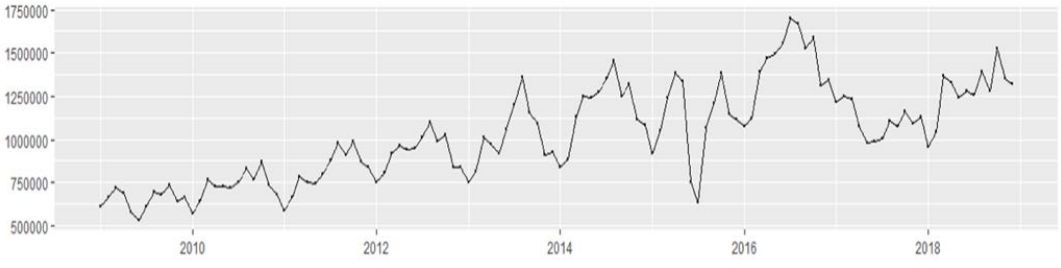
train=subset(data,yyyy<=2018)

test=subset(data,yyyy>=2019)

또한 우리나라에서 해외여행을 가는 출국 여행객 수와 방한 관광객 수가 서로 영향을 끼칠 것이라는 예상을 할 수 있다. 따라서 우리나라에서 출국 하는 여행객 수 데이터도 다운받아 출국 여행객 수와 방한 관광객 수 사이의 연관성도 분석해 볼 예정이다.

2. 방한 관광객 수 데이터 분석 및 모형 결정

2009-2018년 데이터를 그래프로 그린 결과를 보면 일정한 패턴을 가지고 증가 추세를 보이고 있지만, 2015년 5,6월 쯤 급감하는 추세를 보인다. 이는 메르스로 인한 것이라고 추정된다. 그 이후 2016년 에는 중국, 일본 시장에 맞는 유치 정책으로 인해 관광객이 대폭 상승 하였고, 2017년에 사드와 북핵 문제로 인해 감소하는 경향을 보이고 있다.



않고 계절성이 있는 것을 볼 수 있다.

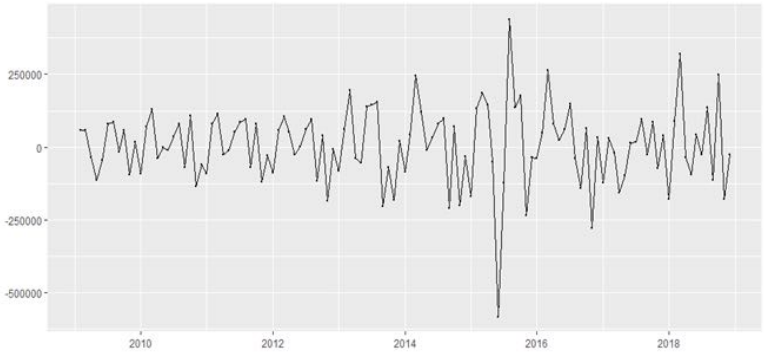
<피드백-1>

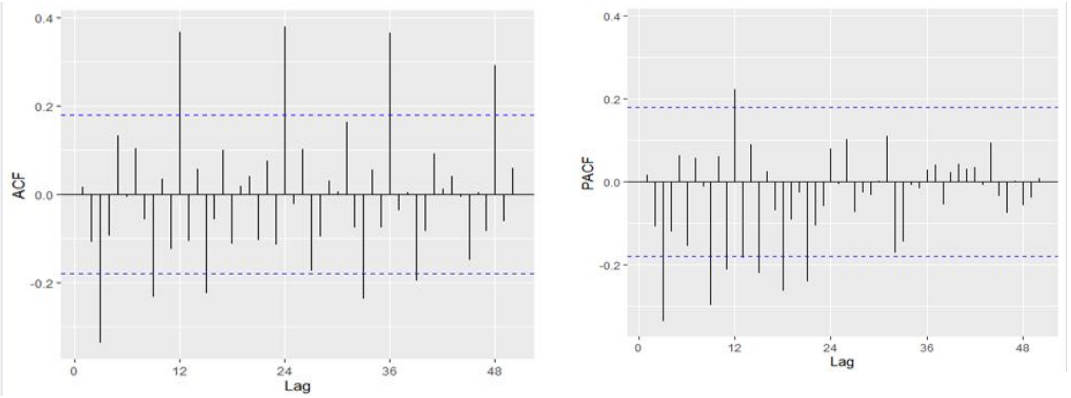
이 부분에서 stationary 검정을 단위근 검정을 통해 확인하는 방법에 대해 피드백을 받았다. 따라서 adf.test를 진행해 보았다.

```
        Augmented Dickey-Fuller Test

data:  data6
Dickey-Fuller = -3.6355, Lag order = 0,
p-value = 0.03328
alternative hypothesis: stationary
```

결과가 p-value 값이 0.03 정도이기 때문에 stationary 라고 할 수 없다.
따라서 1차 차분을 해보았다. 1차 차분을 한 경우 trend는 제거 되었지만, acf 그래프를 보면 12,24,36,48 등에서 계절성이 사라지지 않은 모습을 볼 수 있다.

<trend 제거>

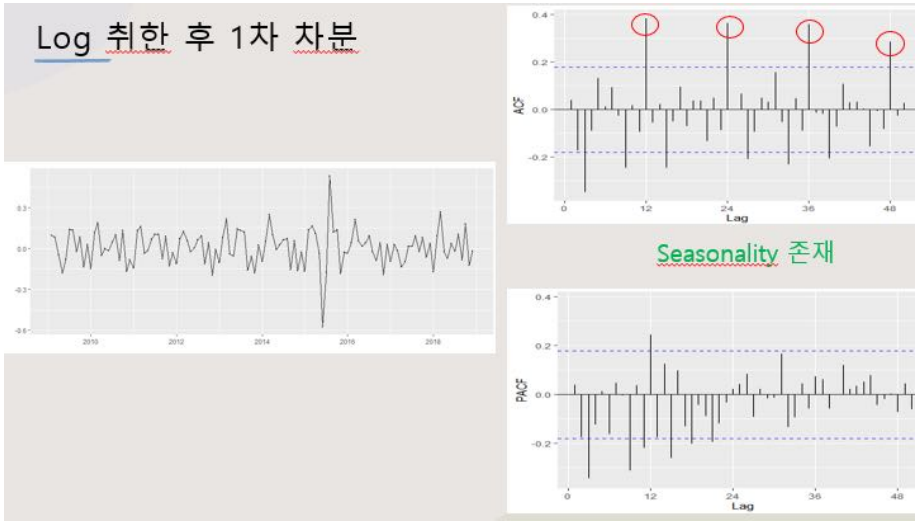


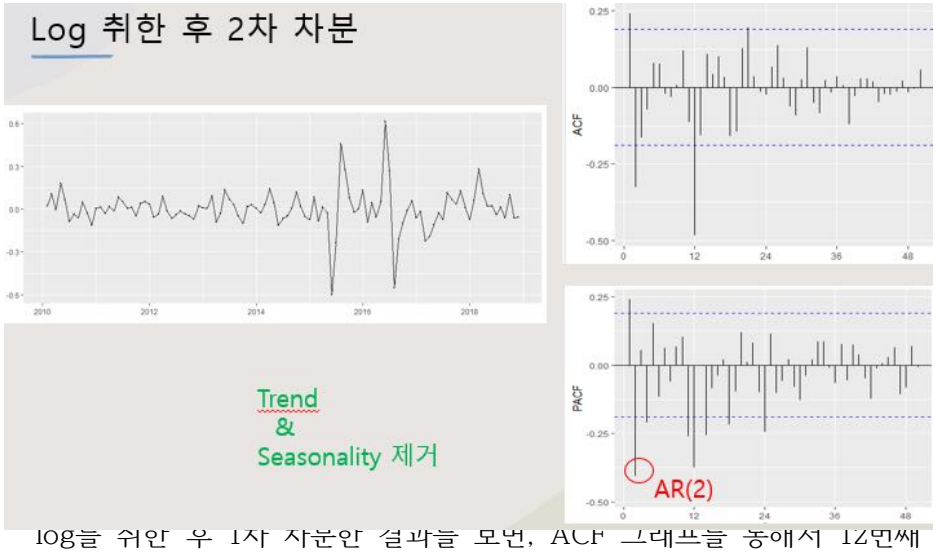


<seaonality 존재>

한 번 더 차분을 해 보았다. 2차 차분결과를 보면 trend 와 계절성이 거의 사라진 것을 볼 수 있다.

특정부분에서 변동폭이 일정하지 않아 데이터에 log를 취한 후 1차, 2차 차분을 진행해 보았다.





log를 취한 후 1차 차분한 결과를 보면, ACF 그래프를 통해서 12번째 LAG 마다 계절성이 있는 것이 드러나 계절성 차분을 한 번 더 진행하였다. 2차 차분 결과 ACF 그래프, PACF 그래프 모두에서 계절성에 대해 정상성 문제가 드러나지 않았다. log를 취한 것과 취하지 않은 결과가 크게 다르지 않은 것을 알 수 있다. 따라서 log를 취하고 2차 차분한 모형을 살펴보도록 한다.

<피드백-1> 받은 부분인 단위근 검정을 log를 취한 후 2 차 차분한 데이터에 대해서도 검정해 보았다.

```
        Augmented Dickey-Fuller Test

data:  diff_22
Dickey-Fuller = -7.5338, Lag order = 0, p-value = 0.01
alternative hypothesis: stationary

Warning message:
In adf.test(diff_22, k = 0) : p-value smaller than printed p-value
```

adf.test를 한 결과 p-value 가 0.01 이므로 stationary 하다고 할 수 있다.
PACF의 LAG2까지 0보다 큰 값이 드러나므로 계절성에 대한 모형은 AR(2)모형을 고려해볼 수 있다. 따라서 이에 맞게 여러 sarima 모형을 살펴 보았다. log를 취한후 1차 차분 한후 계절성에 대해 한번 더 차분하였기 때문에 SARIMA(0,1,0) * (2,1,0) ,SARIMA(1,1,0) * (2,1,0) ,SARIMA(2,1,0) * (2,1,0) 이렇게 3 모형을 비교해 보았다. 또한 Auto.arima를 이용한 모형은 SARIMA(3,1,1) * (2,0,0)였다. 이 모형들의 결과를 보면 SARIMA(2,1,0) * (2,1,0) 모형이 가장 aic 값도 작고 residual test의 p-value 값들도 0.05보다 크기 때문에 잔차가 iid 와 white noise를 성립 하기 때문에 이 모형을 결정하였다.

	Sarima(0,1,0)* (2,1,0)	Sarima(1,1,0)* (2,1,0)	Sarima(2,1,0)* (2,1,0)	Sarima(3,1,1)* (2,0,0)
aic	-153.1	-156.73	-178.82	-177.9
Ljung-Box Q	0.0012*	0.0024*	0.2251	0.5782
McLeod-Li Q	0.0962	0.1149	0.9944	0.9992
Turning points T	0.0014*	0.0057*	0.4238	0.0417*
Diff signs S	3e-04*	9e-04*	0.0833	0.8749
Rank P	0.6262	0.4013	0.5049	0.132

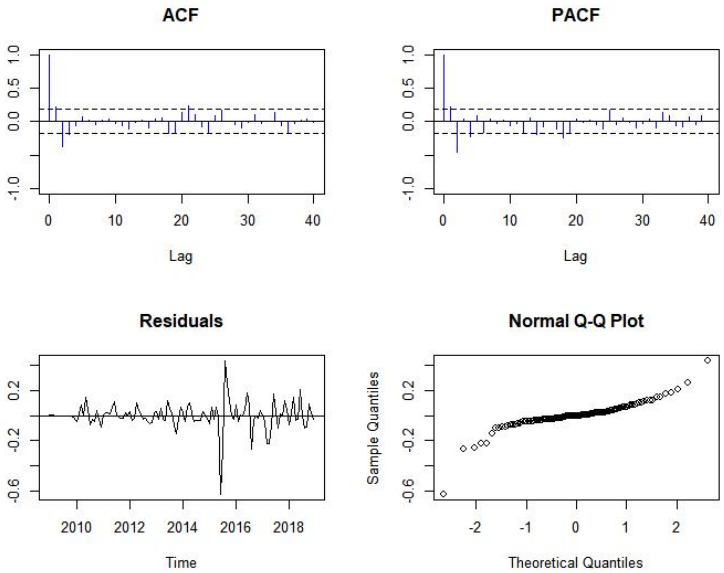
> 0.05

<피드백-2>
SARIMA 모형들의 RMSE 나 MSPE 지표 비교 해보기.

<피드백-3>
잔차 검정에서 잔차의 ACF, PACF, PLOT, QQ-PLOT 도 함께 보여주기.

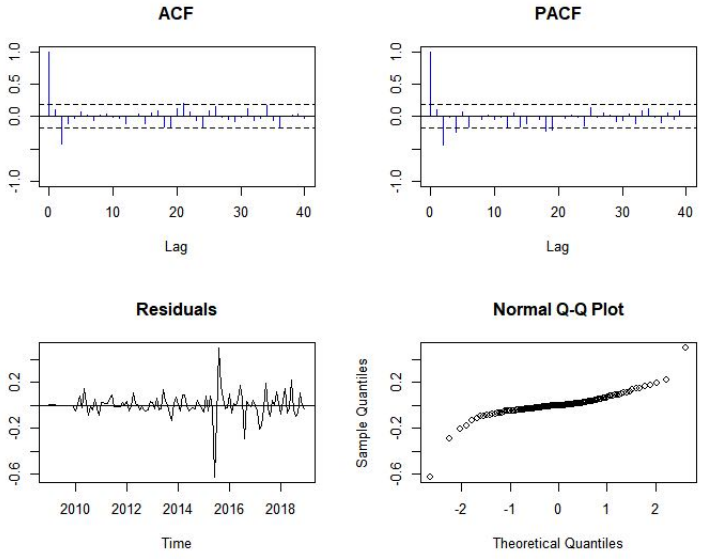
1) SARIMA(0,1,0) * (2,1,0)

```
                      ME       RMSE         MAE          MPE       MAPE       MASE
Training set 0.001278165 0.1056699 0.06171286 0.007525709 0.4456584 0.6309488
                    ACF1
Training set 0.2261279
```



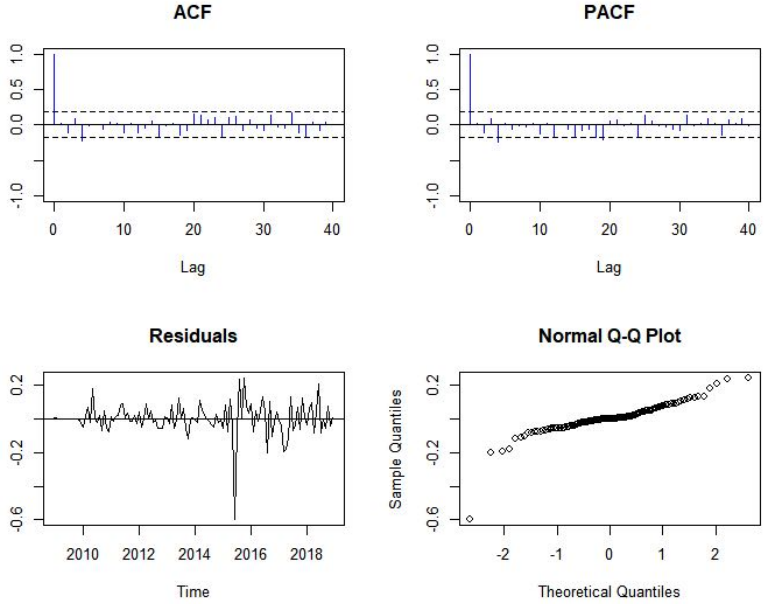
2) SARIMA(1,1,0) * (2,1,0)

```
                       ME       RMSE         MAE           MPE      MAPE       MASE
Training set 0.0009138279 0.1029104 0.05922826 0.005418801 0.427496 0.6055464
                    ACF1
Training set 0.1022469
```



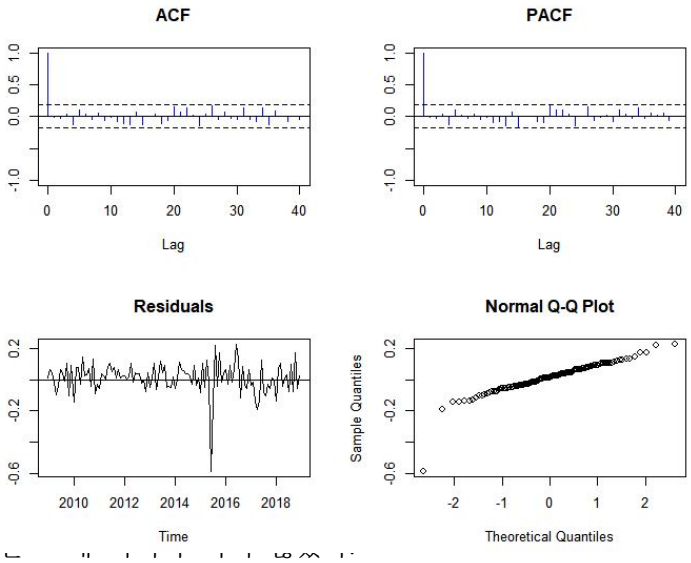
3) SARIMA(2,1,0) * (2,1,0)

```
                      ME        RMSE         MAE           MPE       MAPE
Training set 0.0009845183 0.09139282 0.05630149 0.005751149 0.4059554
                   MASE        ACF1
Training set 0.5756233 0.01622803
```



4) SARIMA(3,1,1)

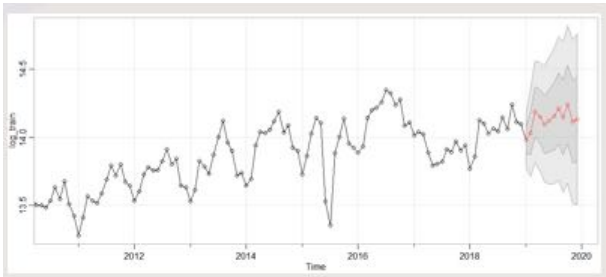
```
                      ME        RMSE         MAE          MPE       MAPE       MASE
Training set 0.00912795 0.09542504 0.06817989 0.06349114 0.4938801 0.6970673
                    ACF1
Training set -0.008673726
```



검정을 해본 결과 I 책하였다.
잔차들의 그래프에

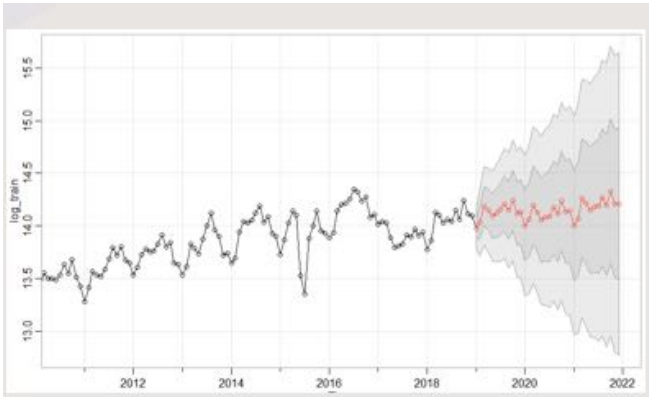
3. 데이터 예측

SARIMA(2,1,0) * (2,1,0) 모형으로 2019년을 예측한 결과를 살펴보면 아래와 같다. log를 취해서 진행했었기 때문에 정확한 수치를 위해서 다시 지수를 취해서 결과를 보았다.



	test.yyyymm	raw_2019	predict_2019
1	201901	1104803	1174875
2	201902	1201802	1242601
3	201903	1535641	1451543
4	201904	1635066	1398317
5	201905	1485684	1322150
6	201906	1476218	1358218
7	201907	1448067	1403954
8	201908	1586299	1485464
9	201909	1459664	1394531
10	201910	1656195	1530486
11	201911	1456429	1349935
12	201912	1456888	1373294

실제 데이터와 차이가 약간 씩은 있지만 2020,2021년도 예측하는 데에는 무리가 없어 보여 모형을 유지하였다. 이를 이용해 2020년도와 2021년도에 코로나가 없었다면 얼만 큼의 관광객이 방문 했을 지를 예측해 보았다.



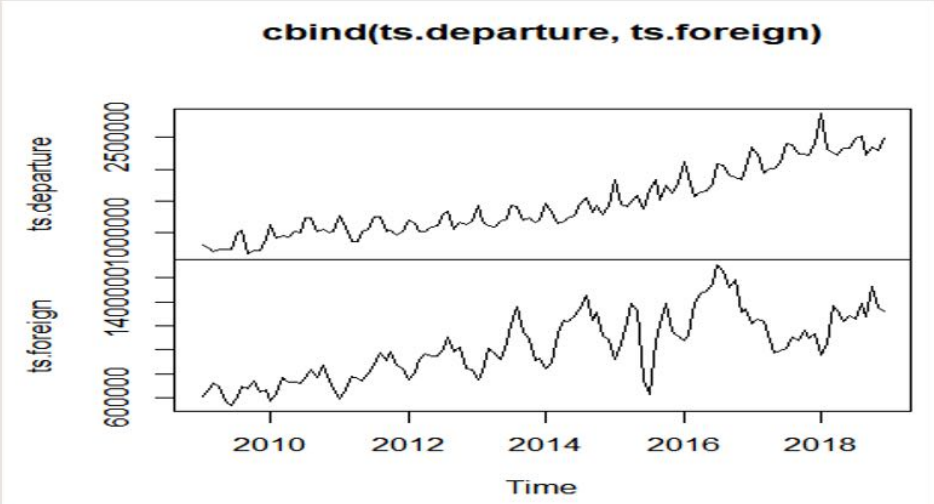
년도	실제 값	예측 값
202001	1272708	1205627
202002	685212	1277345
202003	83497	1466929
202004	29415	1371705
202005	30806	1271562
202006	36938	1304435
202007	61012	1313977
202008	68797	1431058
202009	65040	1351365
202010	61585	1522486
202011	61764	1372545
202012	62344	1386654
202101	58397	1205250
202102	65582	1290045
202103	74604	1558516
202104	70112	1494716
202105	74463	1397164
202106	77029	1438687
202107	83005	1448755
202108	97087	1570811
202109	89800	1466211
202110	92416	1669319
202111	94358	1482678
202112	90150	1486441

있다.

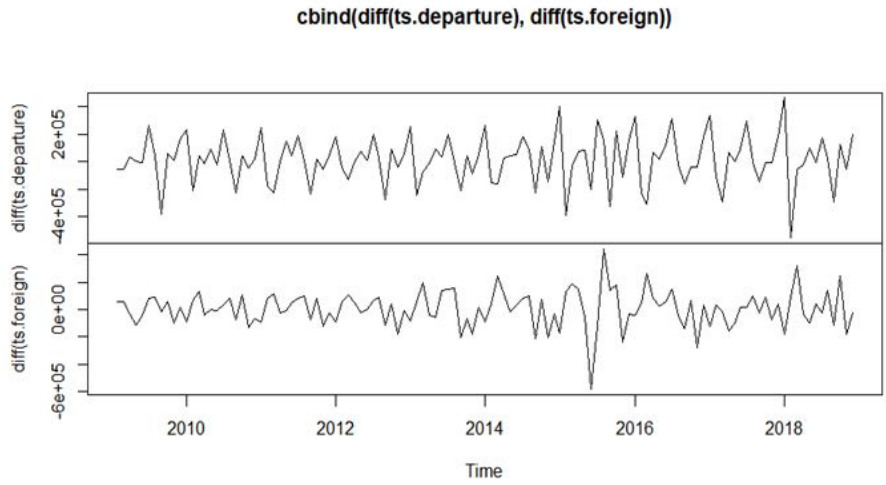
4. 방한 관광객 수와 해외 출국 여행객수 관계 분석

다음은 방한 관광객 수와 해외출국 여행객 수 사이에 연관성이 있음을 확인해 보고 싶어 var 모형을 이용하는 방법을 사용해 보았다. 앞의 데이터와 동일하게 2009-2018년 데이터를 이용하여 두 데이터를 합쳐 그래프를 그려 보았더니 증가하는 트렌드를 보였다.

2009-2018 월별 자료 사용



여기서 departure 이 우리나라에서 해외로 출국하는 여행객 수이고 foreign 이 방한 관광객 수 이다. 증가하는 추세가 보여 두 데이터를 1차 차분한 결과 트렌드가 사라졌다.



차분한 데이터를 var 모형을 사용하기 위해 var select를 이용하였더니 sc 값이 6이 나와 p에 6을 대입 한 후 잔차의 상관관계를 살펴보기 위해 serial.test를 진행 하였다.

```
> var.select <- VARselect(set.diff, lag.max = 8, type = "const")
> var.select$selection
AIC(n)  HQ(n)  SC(n) FPE(n)
     8      6      6      8
```

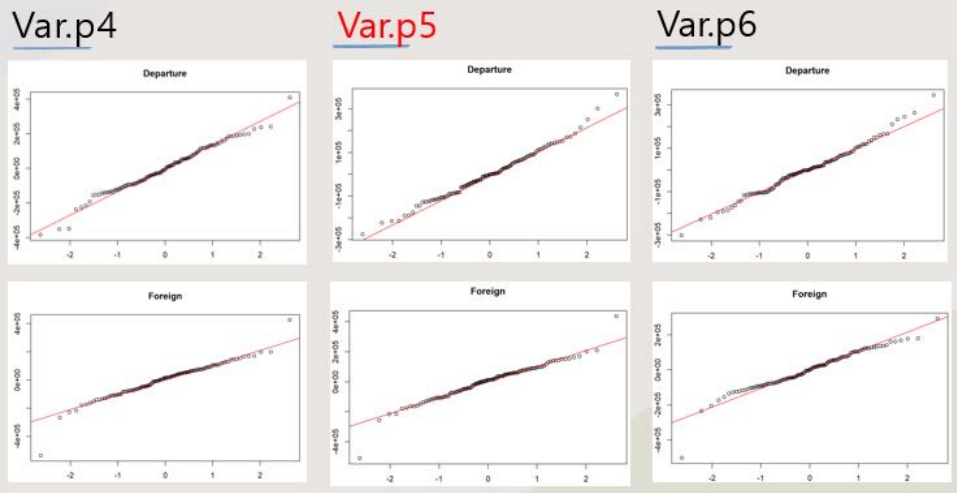
```
> var.p6 <- VAR(set.diff, p = 6, type = "const")
> serial.test(var.p6, type = "PT.asymptotic")

        Portmanteau Test (asymptotic)

data:  Residuals of VAR object var.p6
Chi-squared = 74.685, df = 40, p-value = 0.0007207
```

결과를 보면 p-value 값이 작아 p값들을 조정해가며 p=4, p=5 들도 진행 해 보았다.

```
> var.p4 <- VAR(set.diff, p = 4, type = "const")
> serial.test(var.p4, type = "PT.asymptotic")

        Portmanteau Test (asymptotic)

data:  Residuals of VAR object var.p4
Chi-squared = 124.19, df = 48, p-value = 1.134e-08
```

```
> var.p5 <- VAR(set.diff, p = 5, type = "const")
> serial.test(var.p5, type = "PT.asymptotic")

        Portmanteau Test (asymptotic)

data:  Residuals of VAR object var.p5
Chi-squared = 75.956, df = 44, p-value = 0.001962
```

하지만, p-value 값이 모든 경우 작게 나와 유의미한 결과가 나오지 않았다. 따라서 각각의 p값에 따른 잔차 그래프를 그려 보았다.



크게 차이가 나지는 않지만 p=5 값에서 조금 더 잘 맞아 보이고, 앞서 p값들 중에서 그나마 p-value 값이 큰 p가 5인 모형에서 서로 데이터가 영향을 끼치는지에 대해 확인해 보았다.

1) 해외 출국 여행객 수가 방한 관광객 수에 영향을 주는지

```
> causality(var.p5, cause = "diff.ts.departure.")
$Granger

        Granger causality H0: diff.ts.departure. do not
        Granger-cause diff.ts.foreign.

data:  VAR object var.p5
F-Test = 2.5659, df1 = 5, df2 = 206, p-value = 0.02814
```

2) 방한 관광객 수가 해외 출국 여행객 수데 영향을 주는지

```
> causality(var.p5, cause = "diff.ts.foreign.")
$Granger

        Granger causality H0: diff.ts.foreign. do not
        Granger-cause diff.ts.departure.

data:  VAR object var.p5
F-Test = 4.5084, df1 = 5, df2 = 206, p-value = 0.0006386
```

granger causality를 이용해 결과를 확인한 결과 p value 값이 0.05보다 작아 귀무가설을 기각할 수 있어, 해외출국 수, 방한 관광객 수가 서로 영향을 미치는 것을 알 수 있다. 서로 데이터가 영향을 끼친다는 결과는 얻었지만, var 모형 선택과정에서 p-value 값이 작았기 때문에 var 모형 사용이 이 두 데이터에서는 유의미한지에 대해서는 의문점으로 남았다.