**시계열 분석을 통한 코로나19 확진자 수 예측(SAS이용)**

**3/19부터 12/23까지 코로나19 확진자 수, 총 279개의 데이터를 이용하여 시계열 분석함**

**# 파일 읽기**

**data** lab;

infile 'c:\covidnumber.txt';

input number @@;

date = intnx('day',**'16MAR20'D**,\_n\_-**1**);

format date date9.;

lnumber=log(number);

lnumber1=dif(lnumber);

**run**;

**# 원시계열 그림 확인하기**

symbol1 i=join v=none l=**1** c=black;

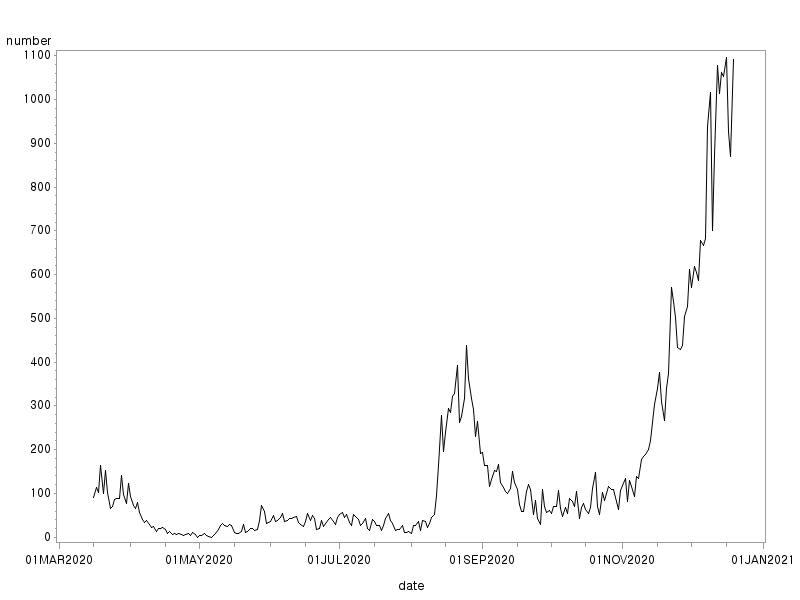
symbol2 i=join v=none l=**3** c=blue;

symbol3 i=join v=none l=**3** c=red;

**proc** **gplot** data=lab;

plot number\*date=**1**;

**run**;



추세요인과 불규칙요인이 존재해보이며, 정상 시계열이 아니라고 보여짐

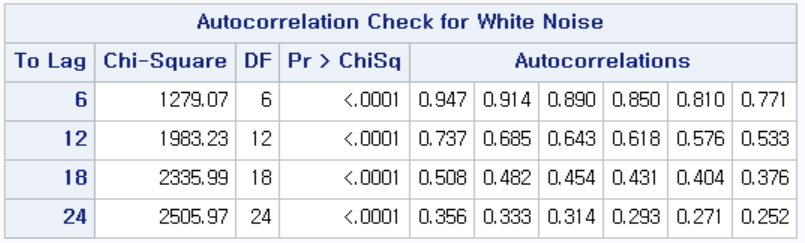
확률적 추세(인접 자료간의 강한 양의 상관으로 인해 추세가 당분간 있는 것처럼 보임)가 보이는 것 같아 이를 확인하기 위해 단위근 검정 시행

또한 시간에 따라 분산이 증가하는 경향이 보이므로 로그변환과 차분을 통한 분산안정화 필요

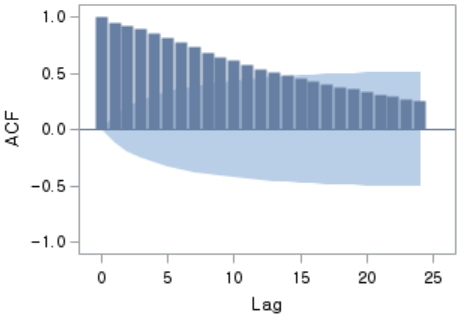
**# 정상성 확인**

**proc** **arima** data=lab;

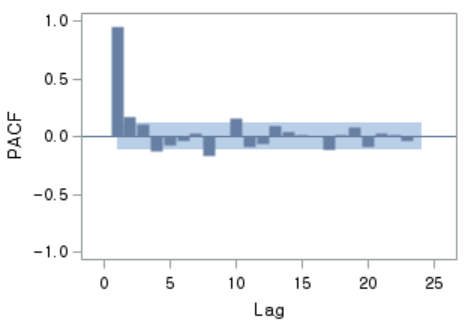
identify var=number nlag=**24**; **run**;



포트맨토 검정 값을 확인해보면, p-value가 작으므로 모든 시차에서 유의합니다. 귀무가설을 기각하므로 자기상관이 존재한다고 볼 수 있다(H0: 자기상관이 존재하지 않음, 데이터가 W.N.을 따름)



또한 ACF가 지수적으로 감소하지 않고, 매우 천천히 감소하므로 비정상시계열



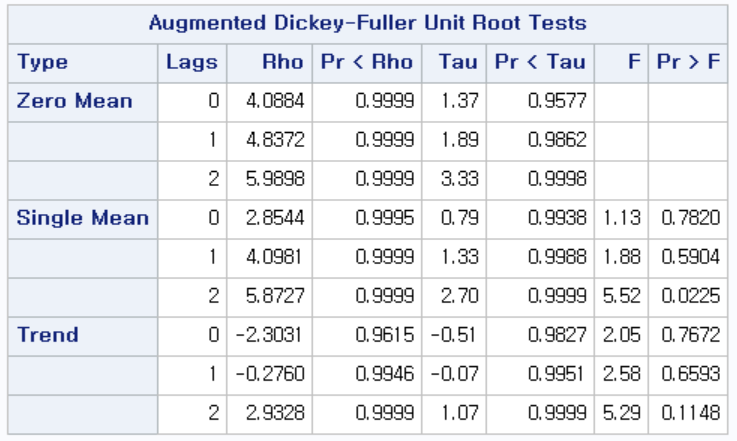
PACF를 보면 시차2에서 절단 (AR(2)과정으로 보이지만, ACF에서 이미 문제가 있음)

추가적으로 확률적 추세를 검정을 통해 확인하기 위해 단위근 검정을 시행해야 함

**# 단위근 검정**

𝐻0 ∶ = 1(단위근) vs 𝐻1 ∶ < 1 (정상)

ADF-test 결과



추세와 절편이 존재하는 시계열 데이터이므로 Trend의 p-value를 봐야한다.

p-value가 크므로 𝐻0을 기각하지 못한다.

따라서, 단위근이 존재하므로 확률적 추세를 가지는 비정상시계열이라 할 수 있다.

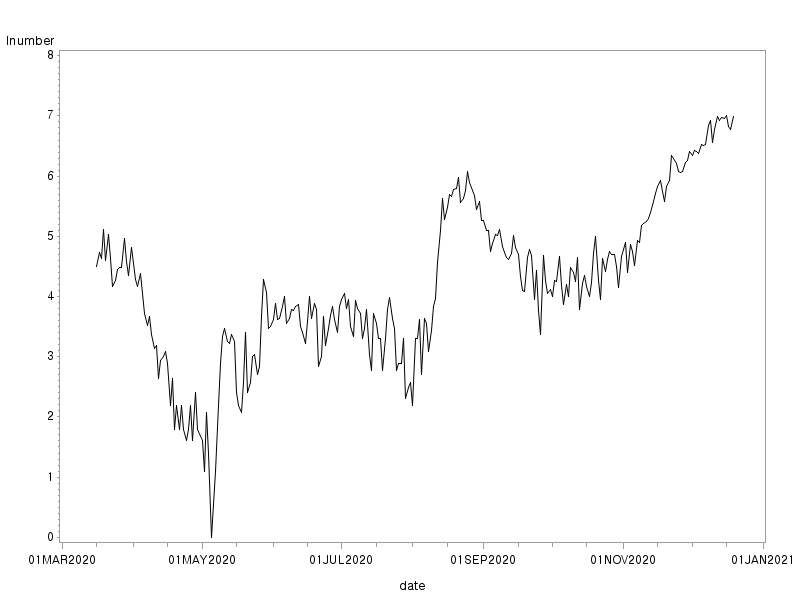
-> 로그변환과 차분을 통해 정상화시켜야 함

**# 비정상성 해결을 위한 로그변환**

**proc** **gplot** data=lab;

plot lnumber1\*date=**1**;

**run**;



로그변환 후에도 정상성을 만족하지 않음

**\***정상성

- 평균이 일정하다(평균이 시간 축에 평행, 뚜렷한 추세가 없음)

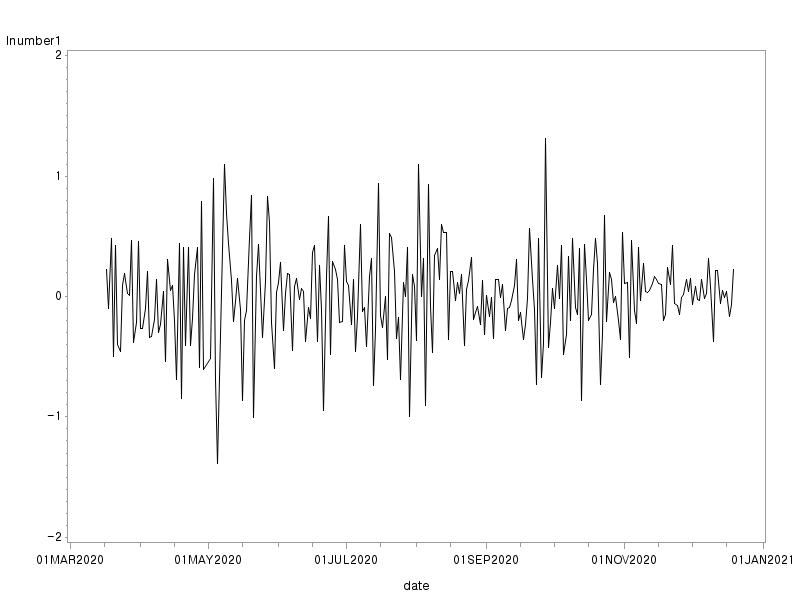
- 분산이 시점에 의존하지 않는다

- 공분산은 단지 시차에만 의존하고 시점 자체에는 의존하지 않는다

**# (로그변환 후) 1차 차분**

plot lnumber1\*date=**1**;

**run**;



평균 수준을 0으로 볼 수 있다(잔차 시계열이 0을 기준으로 랜덤하게 산포)

-> 로그변환과 1차 차분된 자료의 시계열그림: 평균 수준을 0으로 볼 수 있음

분산 또한 시차와 무관: 정상성 만족

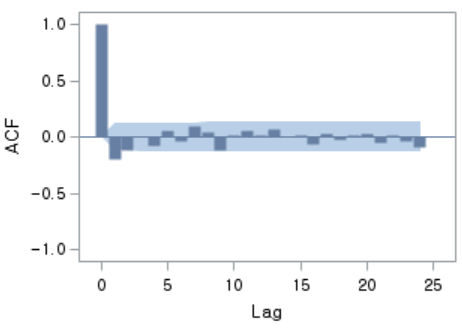
따라서 최종적으로 다음과 같은 모형을 적합한다.

**# 모형 적합 전과 후 비교**

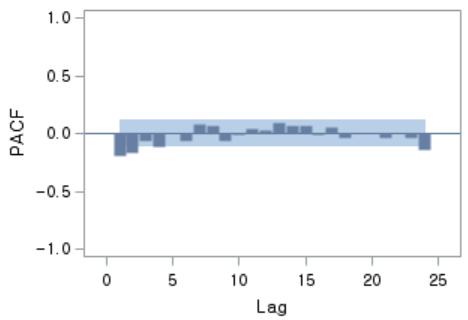
- 모형 적합 전

**proc** **arima** data=lab;

identify var=lnumber1 nlag=**24**;



ACF가 처음 시차 1개에서 유의함



PACF가 처음 시차 1개에서 유의함

- 모형 적합 후

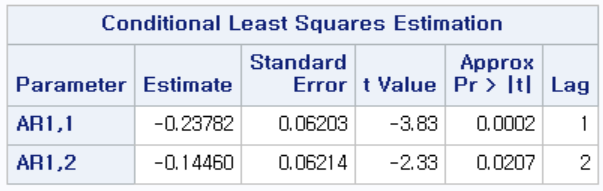
**proc** **arima** data=lab;

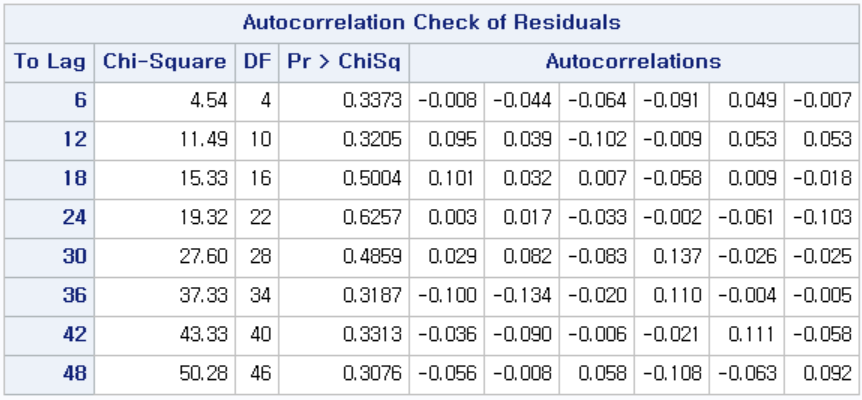
identify var=lnumber1 nlag=**24**;

estimate p=**2** plot noint;

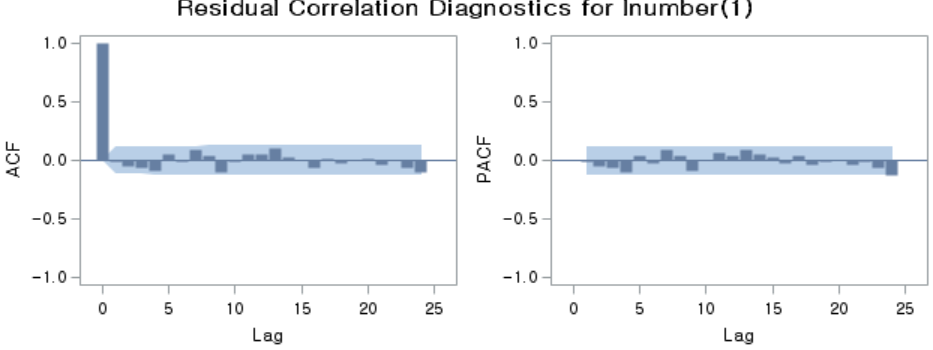
forecast lead=**20** out=fore2;

**run**; **quit;**



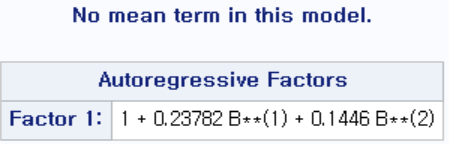


적합된 모형의 진단을 위한 포트맨토 검정 결과 p-value > 0.05 이므로 모든 시차에서 유의하지 않음



잔차의 ACF와 PACF도 모든 시차에서 유의하지 않음

따라서 모형이 적합하다고 판단



**적절한 모델:**

**# 예측**

**proc** **arima** data=lab;

identify var=lnumber(**1**) nlag=**24**;

estimate p=**2** plot noint;

forecast lead=**20** out= fore2;

**run**;

**quit**;

**data** fore2; set fore2; n=\_n\_;

**run**;

**proc** **gplot** data=fore2;

plot lnumber\*n=**1** forecast\*n=**2**/ overlay legend;

**run**;

로그변환한 후의 예측값(의 예측값)과 신뢰수준 95% 하에서의 신뢰구간

