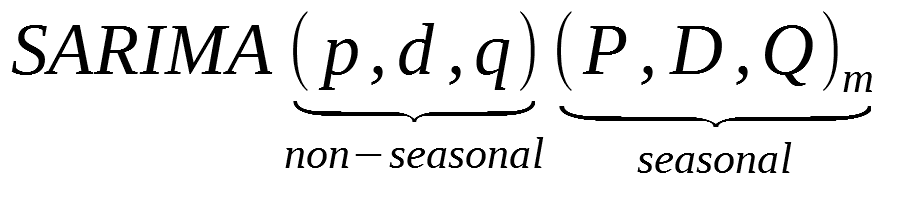
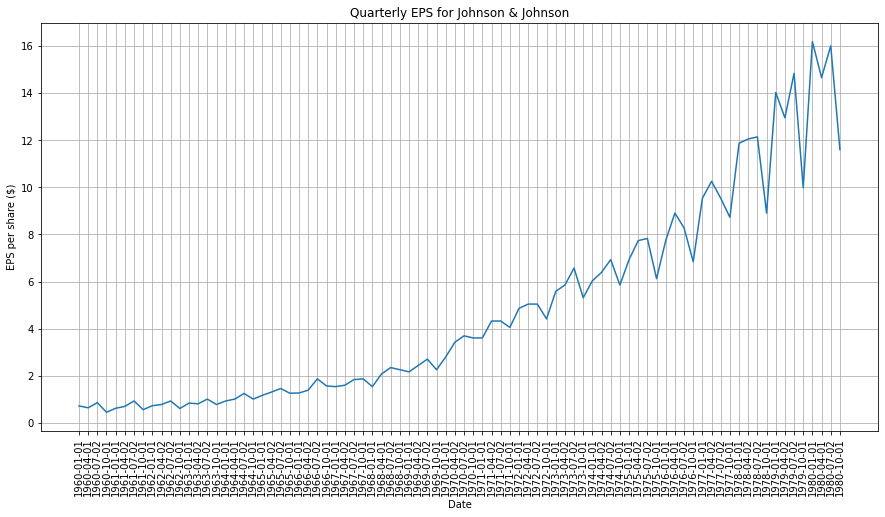
**3. 시계열 분석\_SARIMA**

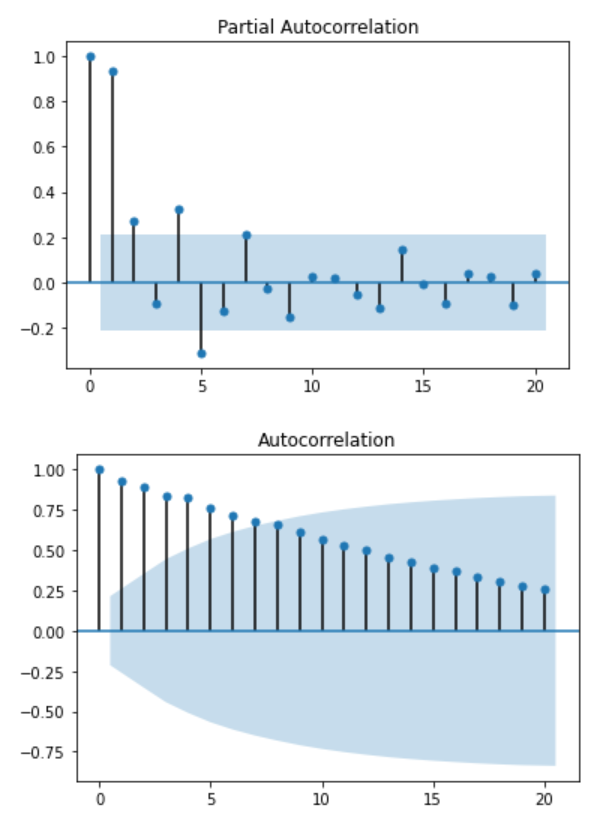
**2021. 03. 21**



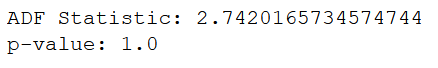


그래프로 확인한 바로 주어진 시계열 데이터는 정상성을 띠고 있지 않다. 그리고 분산이 증가하는 것으로 보아 이분산성 또한 가지고 있다고 할 수 있다.

확인을 위해 PACF와 ACF를 그려본다.



정확한 확인을 위해 Augmented Dickey-Fuller test를 진행한다.



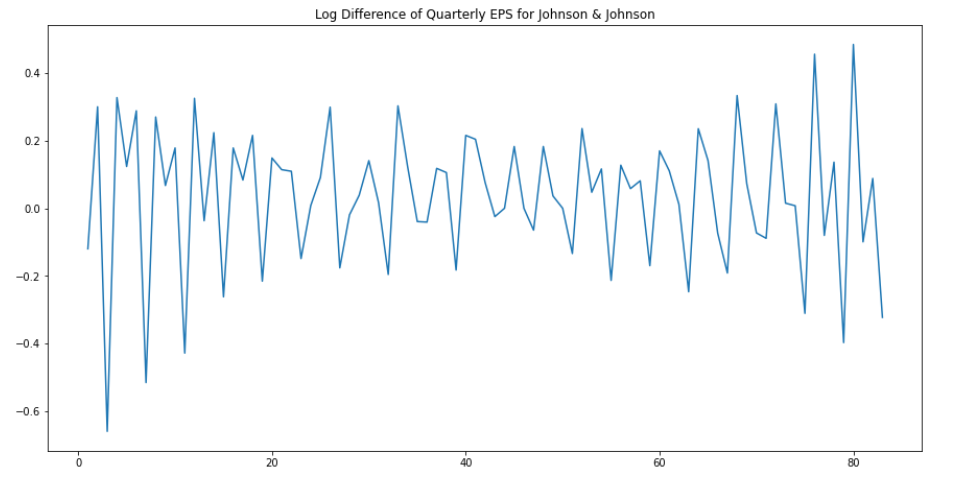
p-value가 크기 때문에 이 시계열 데이터가 정상성을 띠지 않는다는 가설을 기각할 수 없다.

정상성 확보를 위해 log difference를 취한다.

(log를 먼저 취하는 이유는 차분을 먼저 하게 되면 음수값이 발생하여 로그를 취할 수 없기 때문)

data['data'] = np.log(data['data'])  
data['data'] = data['data'].diff()  
data = data.drop(data.index[0])

새로 얻은 데이터는 아래와 같다.

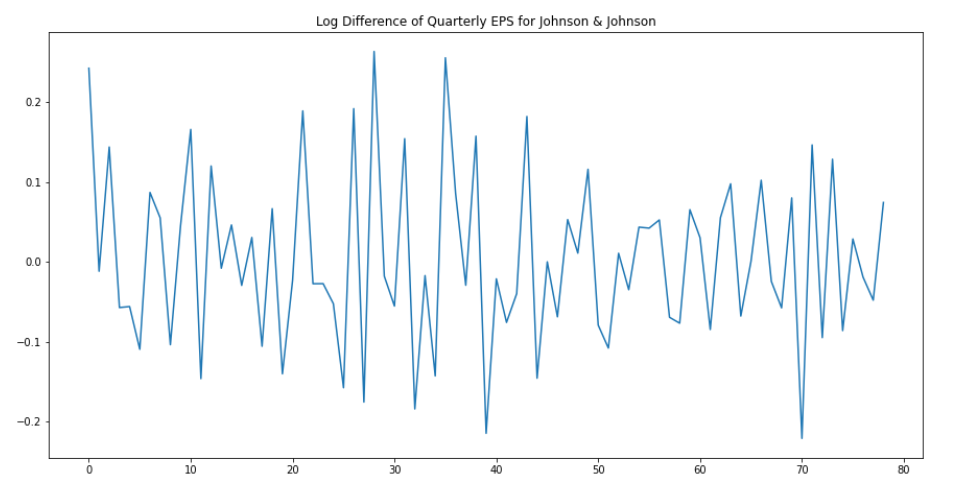


데이터가 분기 데이터이므로, 4 주기의 계절성을 가지고 있다. 따라서 4 주기의 차분을 수행한다.

# Seasonal differencing

data['data'] = data['data'].diff(4)  
data = data.drop([1, 2, 3, 4], axis=0).reset\_index(drop=True)

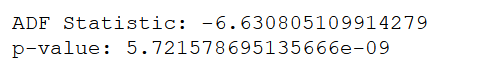
새로운 데이터를 도식화한다.



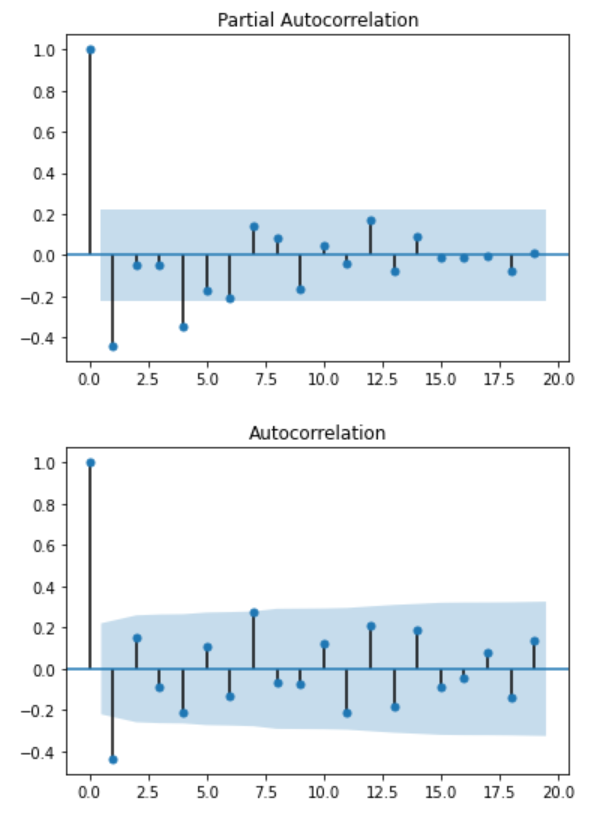
현재의 데이터는 4주기를 바탕으로 차분했으나, seasonal differencing(D)는 1이다. 왜냐하면 차분을 한 번만 수행했기 때문이다.

다시 한 번 ADF test를 수행하면 p-value가 귀무 가설을 기각할 정도로 작기 때문에

데이터가 정상성을 가짐을 확인할 수 있다.



다시 한 번 ACF와 PACF를 도식화한다.



PACF plot의 lag 1에서 peak가 보이기 때문에 AR(p=1) 모델을 고려할 수 있으며, lag 4에서 또다른 peak가 보이기 때문에 seasonal AR(P=1)을 고려할 수 있다.

ACF에서는 lag 1에서만 peak가 확인 되므로 non-seasonal MA(q=1)을 고려할 수 있다.

두 개의 그래프로부터 위와 같은 대략적 모델을 제안할 수 있으나, 더 많은 모델들을 고려하여 AIC가 가장 낮은 것을 고르는 것이 좋을 것이다. 따라서 여러 SARIMA 모델을 확인할 수 있는 함수를 생성하여 최고의 성능을 도출해 내는 parameter를 찾는다.

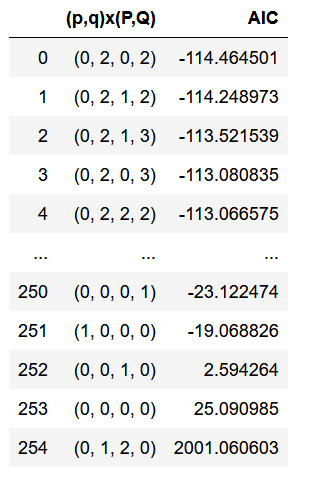
위 과정에서 seasonal과 non-seasonal 차분의 계수인 d와 D가 1인 것과 계절 주기 s가 4인 것을 확인 했으므로, p, q, P, Q의 값에 대해서만 변화시켜가며 성능을 확인해본다.

p = range(0, 4, 1)  
d = 1  
q = range(0, 4, 1)  
P = range(0, 4, 1)  
D = 1  
Q = range(0, 4, 1)  
s = 4

parameters = product(p, q, P, Q)  
parameters\_list = list(parameters)  
print(len(parameters\_list))

총 256개의 조합이 만들어질 수 있고, 생성한 함수를 통해 256개의 SARIMA 모델을 테스트 해 본다.

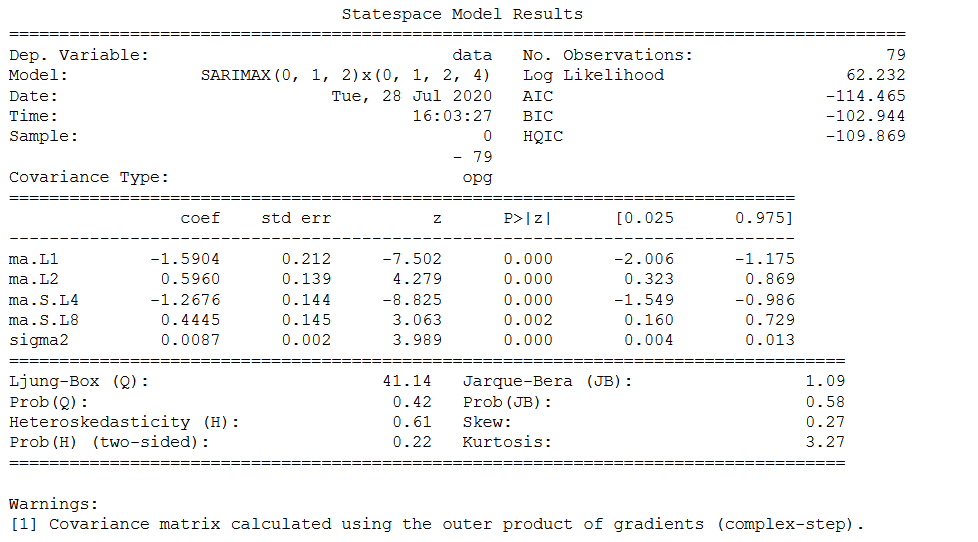
result\_df = optimize\_SARIMA(parameters\_list, 1, 1, 4, data['data'])  
result\_df



위 결과를 통해 best 모델은 SARIMA(0,1,2)(0,1,2,4)임을 알 수 있다.

이 parameter를 사용하여 summary 값을 확인한다.

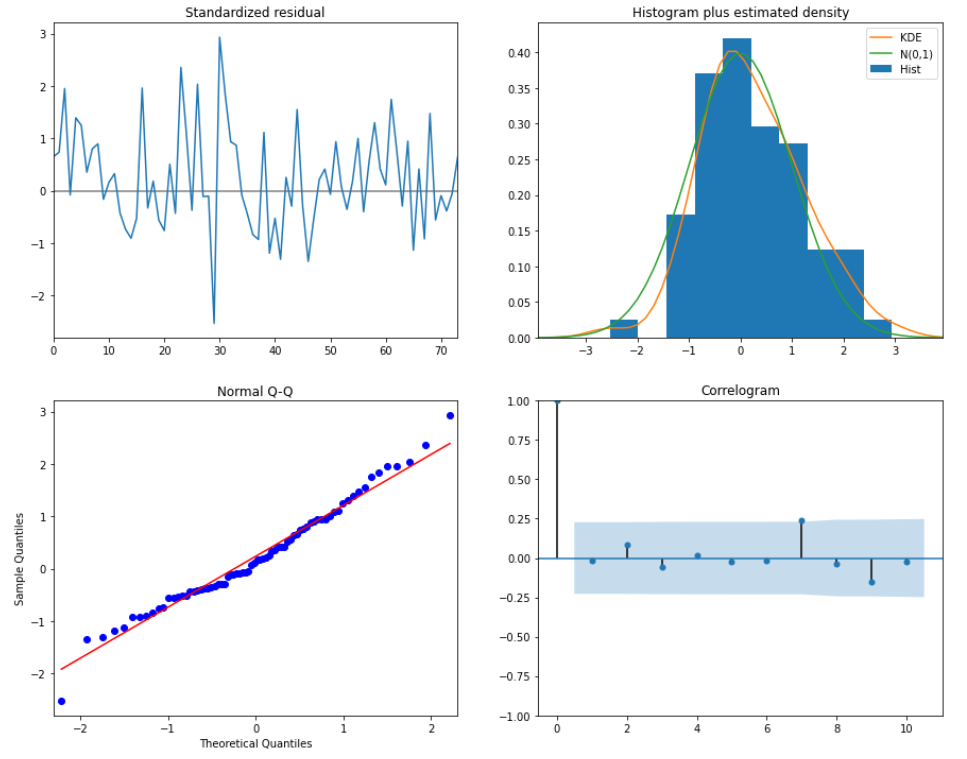
best\_model = SARIMAX(data['data'], order=(0, 1, 2), seasonal\_order=(0, 1, 2, 4)).fit(dis=-1)  
print(best\_model.summary())



이를 통해 Best 모델이 seasonal과 non-seasonal MA process를 가지고 있음을 알 수 있다.

또한 위 값들의 coefficients와 p-value를 확인 할 수 있는데, 모든 coefficients가 유의미함을 알 수 있다.

잔차를 확인하면 아래와 같다.



Normal Q-Q plot으로부터 대부분이 직선 형태를 갖고 있으며, 정규성을 만족한다고 보여진다. 또한 Correlogram을 통해 잔차 간 자기상관성이 없어 백색잡음이라고 볼 수 있다.

데이터를 예측하고 결과를 도식화 해본다.

*(Train, Test 나누어서 돌린 것처럼 해야하면, 예측값은 끄트머리 몇 개만 plotting)*

