COVID-19 사회적 거리두기 단계와 집단감염 규모에 따른 확진자 수 분석

시계열 분석 및 실습 8조

2013-13430 김민준
2016-18145 심태건
2016-19515 김재환

<목차>

1. 주제 소개

본 보고서는 "COVID-19 코로나 거리두기 단계와 집단감염 규모에 따른 확진자 수 분석"이라는 주제를 중심으로 작성되었다. 정부에서 코로나 확진자의 수를 통제하기 위한 정책으로 집합 금지를 포함한 '사회적 거리두기'라는 정책을 시행하고 있는데, 이는 개인 감염자들의 코로나 확산이 아닌 집단 감염의 방지를 목적으로 한 정책이다. 따라서, 이 정책이 전체 코로나 확진자의 수를 감소시키는 데에 실효성이 있는 정책인지를 판단하기 위해, 집단 감염자 수와 개인 감염자 수 각각을 시계열 모델에 적합하고, 거리두기 정책 시행에 따라 각각이 유의미한 감소를 나타내었는지를 분석하였다. 또한, 집단 감염자 수와 개인 감염자 수 사이의 상호작용을 포함한 시계열 적합 역시 시행하고, 이를 통해 현 거리두기 정책의 안정성 또한 검정하였다.

2. 데이터 수집 방법

집단감염자 수를 추정하기 위하여 질병관리청 코로나바이러스감염증-19 홈페이지의 보도자료를 크롤링하였다. 질병관리청에서는 정례보고 시 집단감염자 신규사례 및 역학조사 관련 확진자 수 변경사항을 고지하고 있다. 이런 변경사항을 수집하여 총 집단감염자 수에 대한 데이터를 확보하였다. 게시글은 보도자료 게시판의 2020년 01월 20일 부터 2021년 05월 05일의 총 1599건을 수집하였으며, 게시글의 패턴 분석을 통해 집단감염자 수에 대한 총 3490건의 역학조사 데이터를 확보하였다. 각 데이터는 특정 시점에, 특정 집단 관련해서 몇 명의 확진자가 파악되었는지로 구성된다.

이 데이터를 통해 결과적으로 '보도자료에 따른 질병관리청에서 파악한 특정 일자의 집단감염자 수' 를 파악할 수 있었다. 다만 이것은 일별로 확진 판정을 받은 인구 수에 기반한 데이터기 때문에 실제 집단감염자 수를 완벽히 반영하지 않은 데이터이며, 특히 역학조사를 통해 파악이 되지 않은 확진자 및 추후 확진이 될 수 있는 확진자가 고려되지 않은 수치이다. 그러나 비교대상으로 두었던 일자별 신규확진자 수도 마찬가지로 질병관리청 보도자료를 기반으로 한 것이므로, 데이터에 같은 제약조건이 걸려 이것이 분석에 영향을 미치지 않는 것으로 판단하였다. 구체적으로는, 신규확진자 수 또한 집단감염자 수와 마찬가지로 '질병관리청에서 특정 일자에 파악한 확진자의 수'라는 동일한 제한조건 내에서의 수치이므로, 우리의 주 목적인 '집단감염이 전체 확진자 수에 미치는 영향도'를 파악하기 위해서는 충분한 데이터라고 판단하였다.

이 과정으로 데이터를 수집한 결과, 두 가지 종류의 결측치가 발생하였다. 먼저 집단 감염자와 관련된 보도자료가 없었던 날에 대한 결측치가 있었다. 이 경우 집단 감염자의 수가 불연속적인 데이터로 나타나므로, 분석을 하기에 어려움이 있었다. 따라서, 집단 감염 보도자료가 존재하지 않았던 날과 인접한 두 날의 누적 집단 감염자 수를 선형보간함으로써 이 결측치를 해결하였다. 이 과정에 대한 설명은 Table 1과 같다.

다음으로, 하루 동안의 집단 감염자의 수가 전체 확진자 수보다 많아지는 경우가 결측치로 나타났다. 개인 감염자의 수를 전체 확진자의 수에서 집단 감염자의 수를 뺀 값으로 구하기 때

Table 1. 집단 감염자 수 관련 결측치 수정 과정

날짜	수정 전	수정 후
2020-07-10	923	923
2020-07-11	NA	954.5
2020-07-12	986	986

Table 2. 개인 감염자 수 관련 결측지 수정 과정

날짜	수정 전 감염자 수			수정 후 감염자 수		
	전체	개인	집단	전체	개인	집단
2020-07-10	45	2	43	45	2	43
2020-07-11	56	1	55	56	1	55
2020-07-12	48	62	-14	48	48	0

문에, 이 경우 개인 감염자의 수가 음수로 나타나 결측치가 된다. 이 경우, Table 2와 같이 개인 확진자의 수를 음수가 아닌 0으로 대체함으로써 결측치를 해결하였다.

3. 집단, 개인 감염자 수 분석 - 1. 거리두기 정책의 실효성 분석

집단 감염자의 수와 개인 감염자의 수 각각을 독립적인 시계열 데이터로 간주하고, 각각을 시계열 모형에 적합하고 분석해보았다. 이 과정에서 Box-Jenkins의 모형적합 방법을 이용하여, 모형 식별, 모형 추정 및 모형 진단을 차례로 진행하였다.

3.1. 모형 식별

3.1.1. 시계열도를 이용한 자료 추이 파악

개인/집단 감염자 군을 각각 plotting 한 결과는 다음과 같다.

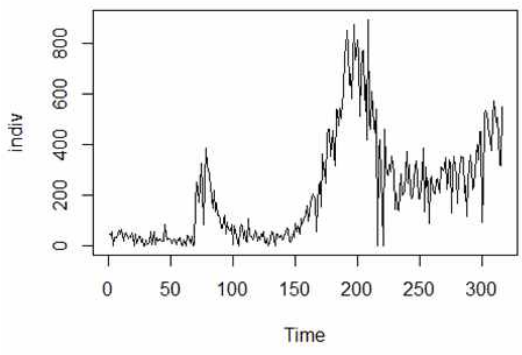

Figure 1. 개인 감염자 수 현황

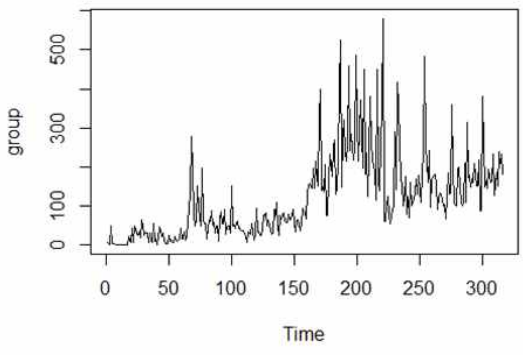

Figure 2. 집단 감염자 수 현황

각 시계열도에서 발견할 수 있는 주요한 특징을 정리해 보면 다음과 같다.

① 전체적인 데이터의 증가하는 추세성
② 이분산성
③ 데이터가 증가할수록 분산도 증가하는 경향성

이를 통해 차분이 요구되는 ARIMA 모형이나 분산의 경향이 데이터의 자기회귀적 성질을 포함하는 GARCH(이분산성) 모형을 생각해볼 수 있었다.

3.1.2. ARIMA 모형 적합의 한계

위에서 언급한 대로 모형의 적합성을 판단하기 이전에 ARIMA 모형과 GARCH(이분산성) 모형을 고려해보았다. 데이터 값이 증가할수록 분산도 증가하는 경향성이 강하게 나타나기 때문에 이분산성 모형이 더욱 적합할 것이라는 예상을 하여, 간단하게 ARIMA 모형에 대해서는 Auto.Arima를 이용하여 각 데이터를 적합시켜 확인해 보았다. 그 결과는 다음과 같다.

Table 3. Auto Arima 모형 적합 결과

Auto arima	Individual	Group
(p, d, q)	(5, 1, 3)	(0, 1, 2)

개인 감염자 모형에서는 요구하는 모수의 수가 굉장히 많음을 알 수 있고, 집단 감염자 모델에서는 차분이 필요한 MA(2)모형으로 비교적 간단하게 자료를 표현할 수 있는 모델이 결과로 출력되었다. 데이터가 정상성을 띠지 않으므로 각 결과는 1차 차분이 필요함을 요구하고 있고, 이를 시각적으로 확인하기 위해 fitting 값을 plotting 하여 실제 값과 비교하여 보았다.

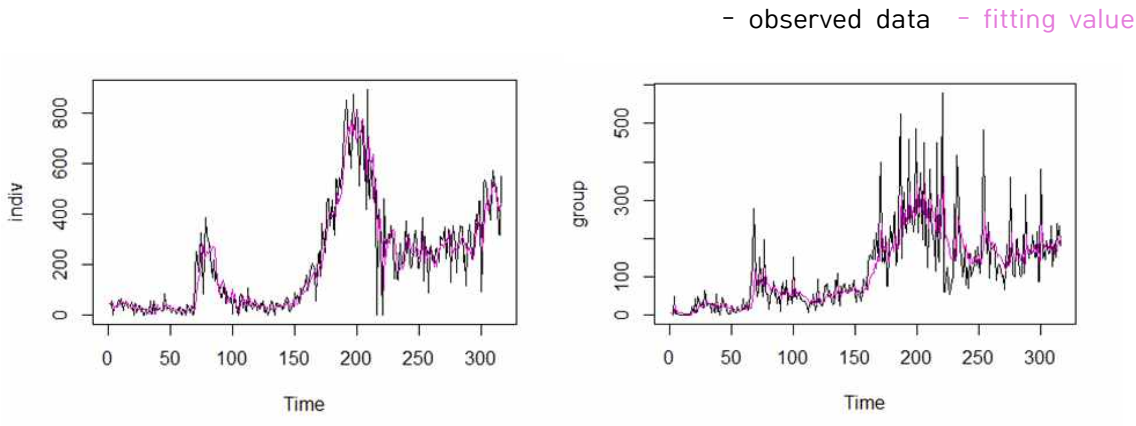

Figure 3. Arima 모형 적합 결과 - 개인 감염자

Figure 4. Arima 모형 적합 결과 - 집단 감염자

Figure 3에서 확인할 수 있듯이, 개인 감염자 모델은 비교적 추세를 잘 설명하는 모습을 보인

다. 그러나 요구되는 모수의 수가 굉장히 많고 분산의 스케일이 전 구간에 걸쳐 유사하게 나타나게 되므로 전반적으로 추세는 비슷하나 이분산성은 잘 설명하지 못하는 모습을 알 수 있다. 한편 집단 감염자 모델은 모수의 수가 적은 만큼 평균적인 추세만을 설명하는 모습을 볼 수 있다. 결론적으로 두 모델 전부 데이터를 적절히 설명하기에는 부족하다고 할 수 있다.

3.2. 모형 선택

3.2.1. AR-GARCH 모형 적합

이어서 이분산성 모형을 적합하여 보았다. 다양한 모형 중에 데이터의 전체적인 자기회귀적 성질과 함께 데이터의 값이 커질수록 분산이 커지는 종합적인 특징을 고려해 보았을 때, AR(p)-GARCH(1,1) 모형을 선택하는 것이 가장 바람직하다고 판단하였다. 따라서 위 모델의 AR order 값을 다르게 하여 3가지 모델을 적합하였고, 사용된 p는 1, 2, 3이다.

3.2.2. 정보기준 및 RMSE의 비교를 통한 모형 선택

Table 4는 각 개인/집단 감염자 별로 적합된 모형의 AIC, RMSE 값이다.
두 집단에서 모두 모수의 수가 증가할수록 AIC값이 감소하고, RMSE값이 감소하는 것을 알 수 있다. 하지만 이는 일반적으로 모수의 수가 증가할수록 나타나는 현상이므로 간결의 원리에 의해 두 집단 모두 p = 2인 모형, 즉 AR(2)-GARCH(1,1) 모형을 선택하였다.

Table 4. 개인 및 집단 감염자 수에 대한 적합 모형의 AIC 및 RMSE

Model	p	AIC	RMSE
Individual	1	3564.91	85.12
	2	3538.168	82.06
	3	3510.474	79.47
Group	1	3491.966	95.37
	2	3445.866	89.37
	3	3438.972	85.96

3.2.3. 선정된 모형 적합 결과

정보기준(AIC)과 평균제곱근오차(RMSE)와 모수의 수를 고려하여 적절히 선택된 모형은 AR(2)-GARCH(1,1) 모형이고, 이 모형의 일반적인 식은 다음과 같이 나타낼 수 있다.

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t$$

$$\epsilon_t = \zeta_t \sigma_t, \zeta_t \sim IID(0,1)$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 \epsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2$$

결론적으로 총 5가지 모수를 추정하게 되며 각 모수가 유의한지 검정이 필요하다. 각 모수에 대하여 t-test로 유의성 검정을 진행한 결과 다음과 같은 결과를 얻을 수 있었다.

Table 5. 개인 및 집단 감염자 수에 대한 AR(2)-GARCH(1, 1) 모형의 모수 유의성 검정

Model	Parameter	Estimate	Std.error	t value	Pr(>|t|)
Individual	ϕ_1	0.60684	0.06377	9.516	< 2e-16
	ϕ_2	0.35255	0.06413	5.497	3.85e-08
	α_0	91.58834	26.81244	3.416	0.000636
	α_1	0.27292	0.05745	4.750	2.03e-06
	β_1	0.76687	0.03551	21.598	< 2e-16
Group	ϕ_1	0.56048	0.06075	9.226	< 2e-16
	ϕ_2	0.47958	0.06100	7.861	3.77e-15
	α_0	80.20252	34.07154	2.354	0.0186
	α_1	0.59590	0.10290	5.791	7.00e-09
	β_1	0.58157	0.04189	13.884	< 2e-16

추정된 모든 모수가 유의수준 0.05에서 유의하다고 추정되었다. 다음으로 각 적합된 모델의 값들을 실제 값과 함께 plotting해 보았다. 그 결과는 다음과 같다.

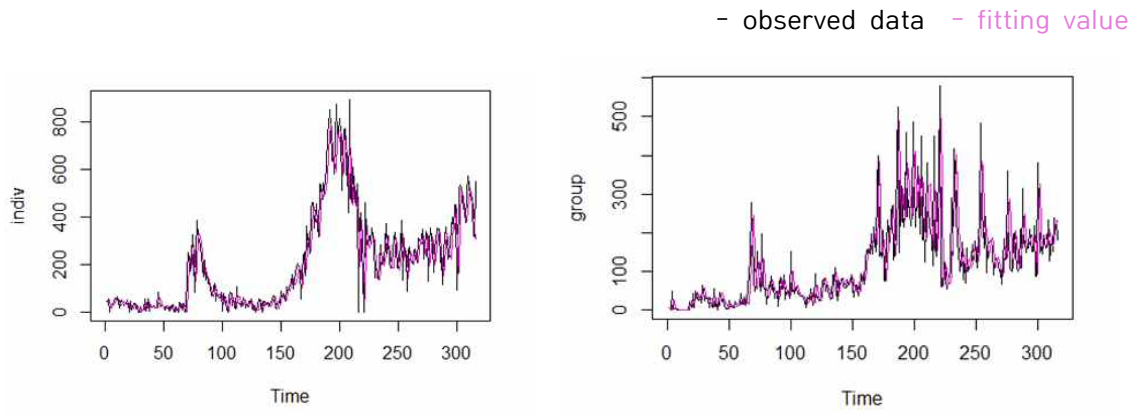

Figure 5. AR(2)-GARCH(1, 1) 모형 적합 결과 [개인 감염자 수]

Figure 6. AR(2)-GARCH(1, 1) 모형 적합 결과 [집단 감염자 수]

두 모형 모두 데이터의 전체적인 자기회귀적 성질과 함께 데이터의 값이 커질수록 분산이 커지는 경향성을 잘 반영하였다는 것을 알 수 있다.

3.3. 모형 진단

다음으로 위 모형이 적합되기 이전에 설정한 적절한 가정들을 확인하여 보고, 모수의 개수가 적절한지를 판단 진단하기 위해 잔차분석과 과적합 진단을 실시하였다. 잔차분석은 정규성 검정, 등분산 검정, 독립성 검정 순으로 진행하였고 과적합 진단은 그 후에 서술하였다.

3.3.1. 정규성 검정

먼저 정규성 검정 결과이다. 시각적으로 확인하기 위해 QQ-plot을 도시하였다.

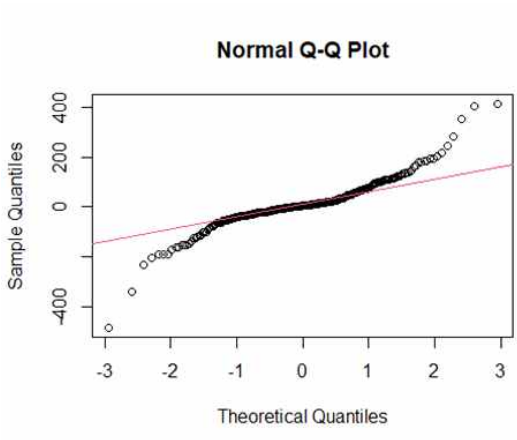

Figure 7. Q-Q Plot (ζ_t, 개인 감염자)

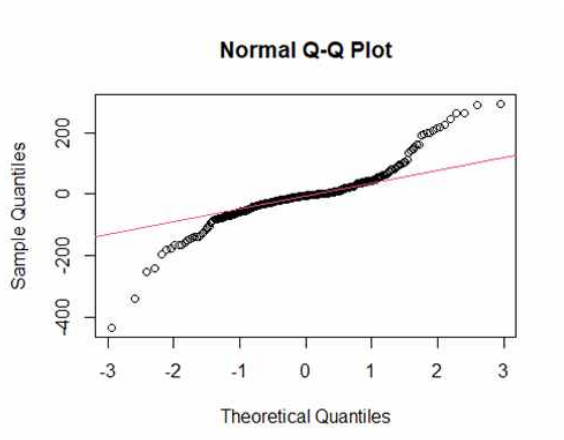

Figure 8. Q-Q Plot (ζ_t, 집단 감염자)

중앙에 위치한 값들은 정규성 경향이 잘 맞아 떨어지는 모습을 보이나, 양극단으로 갈수록 QQ-line의 경향성에서 벗어나 정규성을 잘 만족하지 않는 것으로 나타난다. 정확한 검정을 위해 Shapiro test를 실시하였다. Shapiro test는 Jarque-Bera test와 마찬가지로 정규성 검정에 이용되는 검정이며, skewness와 kurtosis를 이용하는 Jarque-Bera test와는 다르게 검정 대상이 되는 통계량이 정규분포를 따를 경우 그 순서통계량의 거동을 이용하여 정규성 가정에 대한 검정을 진행한다. 이 검정의 귀무가설과 대립가설은 다음과 같고 그 결과도 이어서 나열하였다. Table 6에서 확인할 수 있듯이, 개인 감염자와 집단 감염자에 대한 ζ_t 모두 정규분포를 따른다는 귀무가설을 기각하게 된다.

$$H_0 : \zeta_t\text{가 정규분포를 따른다.}$$
$$H_1 : \zeta_t\text{가 정규분포를 따르지 않는다.}$$

Table 6. 정규성 가정에 대한 Shapiro test 결과

구분	Individual	Group
p-value	1.315e-07	2.18e-12
Test result	귀무가설 기각	귀무가설 기각

3.3.2. 등분산 검정

이어서 등분산성 검정 결과이다. 등분산 가정을 시각적으로 확인하기 위해 잔차를 도시한 결과는 다음과 같다.

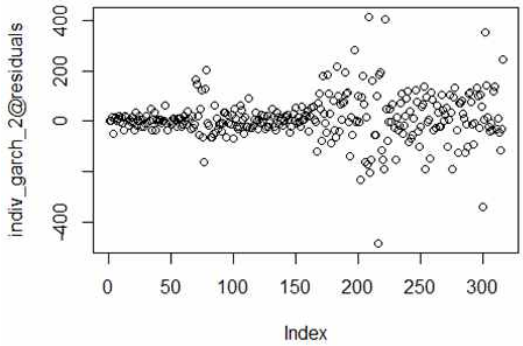

Figure 9. 개인 감염자 수 - 잔차 plot

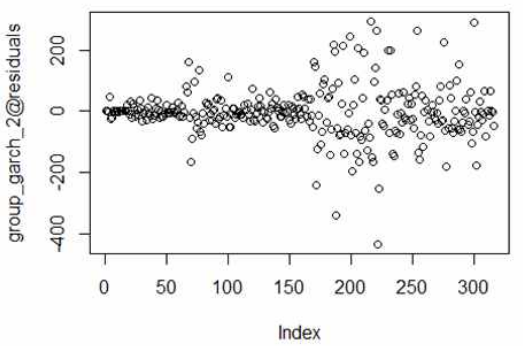

Figure 10. 집단 감염자 수 - 잔차 plot

Figure 9, 10 모두 중심축을 기준으로 잔차의 scale이 상당히 달라지는 것을 확인할 수 있다. 이로써 등분산 가정을 기각하게 된다.

3.3.3. 독립성 검정(자기상관 검정)

마지막으로 독립성 검정(자기상관 검정)에 대한 결과이다. 각 잔차에 대해 자기 상관 함수와 부분자기상관함수를 도시한 결과는 다음과 같다.

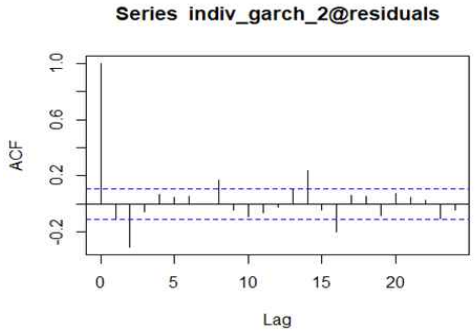

Figure 11. 개인 감염자 수 - 자기상관함수

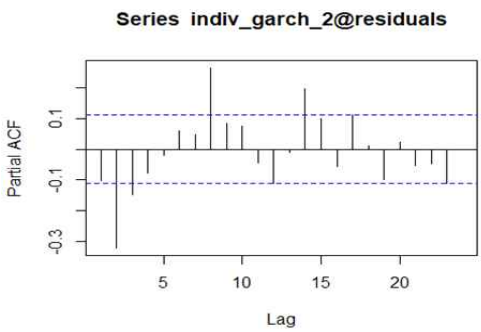

Figure 12. 개인 감염자 수 - 부분자기상관함수

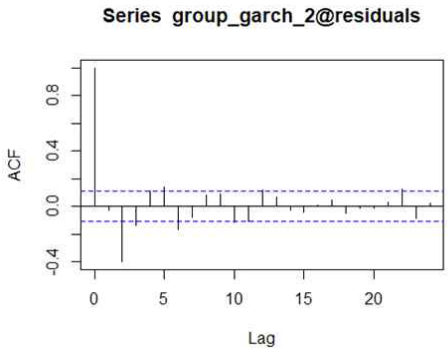

Figure 13. 집단 감염자 수 - 자기상관함수

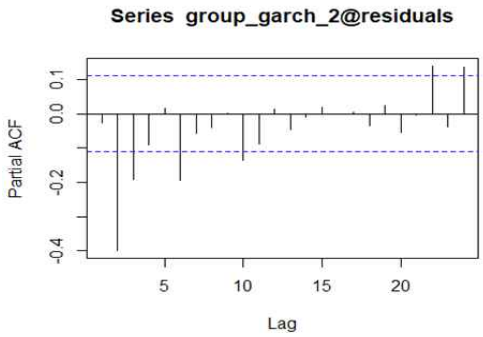

Figure 14. 집단 감염자 수 - 부분자기상관함수

그래프로 확인한 결과 각각의 경우 모두 자기 상관성을 띄지 않는 것으로 예상된다. 이를 정량적으로 검정하기 위해 Ljung-Box Test 를 실시하였다. 이 검정의 귀무가설과 대립가설은 다음과 같고, 검정의 결과 및 유의수준 역시 이어서 나열하였다.

H_0 : 오차항 사이의 자기상관이 없음

H_1 : 오차항 사이의 자기상관이 있음

Table 7. 오차항에 대한 Ljung-Box test 시행 결과

구분		Individual	Group
p-value	lag = 1	0.0653	0.6237
	lag = 3	1.392e-07	3.243e-12
	lag = 5	6.55e-07	4.849e-13
	lag = 10	2.341e-07	2.753e-14
	lag = 20	9.277e-12	1.69e-12
Test result	lag = 1	귀무가설 채택	귀무가설 채택
	otherwise	귀무가설 기각	귀무가설 기각

보편적으로 Ljung-Box test를 시행할 때, lag 값을 변화시키며 반복하여 검정을 시행하므로 위와 같이 검정을 진행하였다. 결과적으로 lag = 1인 경우를 제외한 모든 경우 귀무가설을 기각하는 결과가 도출되었다.

3.3.4. 과적합 진단

추가로 모형의 진단을 위해 과적합 진단을 실시하였다. 사용된 모수의 값을 1씩 추가하여 모수의 유의성을 검정하였으며, 적합성 비교를 위해 사용된 모형은 AR(2)-GARCH(2,1), AR(2)-GARCH(1,2)을 이용하여 최종 선택된 모형과 비교하여 보았다.

Table 8. 개인 감염자 수 - 과적합 진단

Model	Parameter	Estimate	Std.error	t value	Pr(>|t|)
AR(2)-GARCH(1,2)	ϕ_1	0.61927	0.05810	10.658	< 2e-16
	ϕ_2	0.35145	0.05917	5.939	2.86e-09
	α_0	96.95408	32.52337	2.981	0.00287
	α_1	0.10176	0.05156	1.974	0.04844
	α_2	0.25458	0.08847	2.877	0.00401
	β_1	0.70750	0.04622	15.308	< 2e-16
AR(2)-GARCH(2,1)	ϕ_1	0.6069	0.06391	9.496	< 2e-16
	ϕ_2	0.3526	0.06414	5.498	3.84e-08
	α_0	92.88	35.13	2.644	0.00819
	α_1	0.2725	0.09602	2.838	0.00453
	β_1	0.7667	0.4676	1.640	0.10107
	β_2	1.000e-08	0.3994	0.000	1.00000

Table 9. 집단 감염자 수 - 과적합 진단

Model	Parameter	Estimate	Std.error	t value	Pr(>|t|)
AR(2)-GARCH(1,2)	ϕ_1	0.57369	0.06278	9.138	< 2e-16
	ϕ_2	0.46396	0.06497	7.141	9.27e-13
	α_0	67.63230	31.71267	2.133	0.03295
	α_1	0.36231	0.13893	2.608	0.00911
	α_2	0.35007	0.18464	1.896	0.05797
	β_1	0.52180	0.04984	10.469	< 2e-16
AR(2)-GARCH(2,1)	ϕ_1	0.5603	0.0611	9.170	< 2e-16
	ϕ_2	0.4799	0.06132	7.827	5.11e-15
	α_0	83.20	62.52	1.331	0.18328
	α_1	0.5989	0.1979	3.027	0.00247
	β_1	0.5784	0.4293	1.347	0.17791
	β_2	1.000e-08	0.2907	0.000	1.00000

개인/집단 군에 대하여 모두 AR(2)-GARCH(2,1) 모형에서 추가된 모수 β_2는 유의하지 않다는 결론이 도출되었다. 또한 개인 감염 AR(2)-GARCH(1,2) 모형에서 추가된 모수는 유의하다는 결론이지만, 그 모수가 추가됨으로 인해 기존 모수 α_1의 유의확률이 0.05에 근접하게 추정된다는 문제가 발생하며, 집단 감염 AR(2)-GARCH(1,2) 모형에서는 추가된 α_2의 유의확률이 0.05를 초과하여 유의하지 않다는 결론이 도출되었다.

3.3.5. 모형 진단 결과 요약

정규성과 등분산성에 대한 검정은 기각되었고, 상관검정에 대한 결과 역시 lag = 1인 모든

경우에 유의수준이 매우 낮게 나타나 잔차들 사이의 상관관계가 존재하지 않는다는 귀무가설 역시 기각되었다. 이는 위에서 제시한 잠정 모형이 부적합하다는 것을 암시한다. 일반적으로 GARCH 모형에서 오차항의 분포가 정규분포를 따르지 않더라도 적합이 불가능한 것은 아니므로 정규성 검정에 대한 결과는 용인할 수 있지만, 자기상관 검정을 진행한 결과 잔차들끼리 독립적이지 않아 집단 감염자의 수와 개인 감염자의 수가 AR(2)-GARCH(1,1) 모형의 조건들을 잘 따른다고 보기 어렵다. 그러나 AIC, RMSE, 모수의 유의성 등을 종합적으로 고려한 결과, 다른 모형들보다 AR(2)-GARCH(1, 1)이 집단 및 개인 감염자 수를 적합할 때 더 바람직하다는 결론을 내리게 되었다. 등분산 가정에 대한 부분은 모형의 선택 단계에서 간결한 모형을 선택하기 위해 세세한 자료의 표현을 양보한 결과로 해석할 수 있고, 마지막으로 과적합 진단에 대한 결과 추가된 모수의 유의성이 대부분 없는 것 역시 기 선택된 모형이 적합한 모형이라는 근거로 작용한다.

3.4. 추가 논의

3.4.1. 변화점 탐지 분석

연구의 중점 중 하나인 개인/집단 감염에 대한 정부 정책의 주요 효과에 대한 분석을 진행하기 위하여 모델의 변화점 탐지를 진행하였다. 변화점 탐지 검정은 집단 감염자 수, 개인 감염자 수 각각에서 확진자 수의 fitted value값을 이용하여 모델의 모수가 가장 크게 변하는 지점을 찾아내는 방향으로 진행하였으며 변화점이 탐지되지 않을 때 까지 반복적으로 구간에 대한 검정을 진행하여 다음과 같은 Change point를 얻었다. 각 point에 대응되는 날짜(연/월/일)는 다음과 같다.

Table 10. 개인 및 집단 감염자 수에 대한 변화점 탐지 결과

Model	Change point	Date(Y/M/D)
Individual	168	2020-11-24
Group	65	2020-08-13
	160	2020-11-16
	168	2020-11-24

3.4.2. 정부 거리두기 정책과 변화점의 상관관계 논의

위 분석된 변화점들과 관련한 정부 정책의 효과를 분석하기 위하여 탐지된 변화점 근방에서 행해진 주요 정부 정책의 추이를 정리하여 보면 다음과 같다.

코로나 이후 주요 정부 정책

- 2020-12-08 부터 연말까지 거리두기 강화
- 2020-12-24 부터 5인이상 집합금지 시행

• 2021-01-17 까지 거리두기 확장

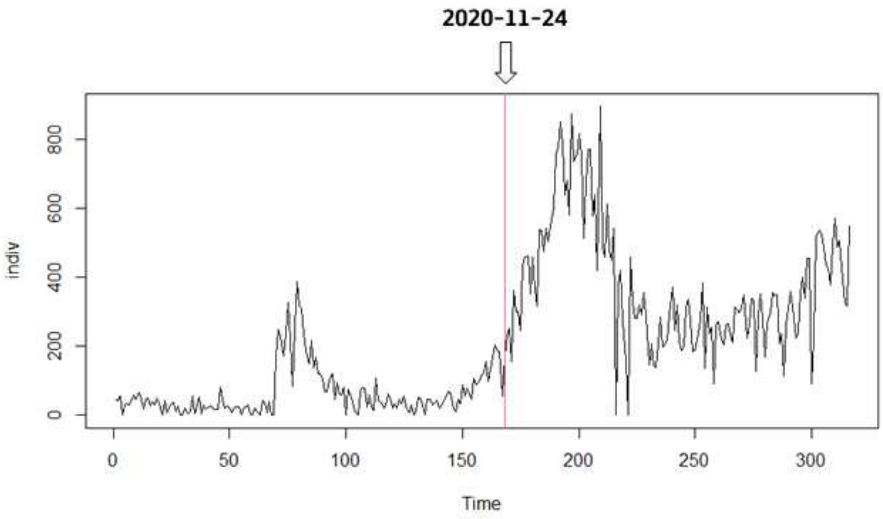

Figure 15. 개인 감염자 수에 대한 변화점 탐지 결과

• 2021-02-15 부터 거리두기 완화
• 2021-04-09 핀셋 방역 실시 / 집단 감염 위험시설 집중 방역

개인 감염군의 경우 2020년 11월에 접어들면서 감염자수가 급등하는 추세를 보인다. 이는 겨울이라는 계절적 특성이 바이러스의 전파에 강한 영향을 미친 것으로 해석된다. 실제로 탐지된 변화점 역시 2020-11-24 로써 이 근방에서 유의미한 변화가 있었음을 통계적으로 입증할 수 있다.

시계열도를 살펴보면 2020년 11월부터 확진자가 급증한 이후에 일정 기간 뒤 감염자가 감소하는 경향을 확인할 수 있다. 또한 비슷한 기간에 '5인 이상 집합금지'와 확장된 '사회적 거리두기'가 시행되면서 정부의 방역이 크게 강화되었음을 확인할 수 있는데, 이로써 개인 감염 집단이 정부 정책의 영향을 상당히 많이 받는 것으로 볼 수도 있다. 그러나 변화점으로 인식 될 정도는 아니므로 뒤에 서술할 집단 감염에 비해 상대적으로 정부 정책이 갖는 억제력이 약한 것으로 해석할 수 있다.

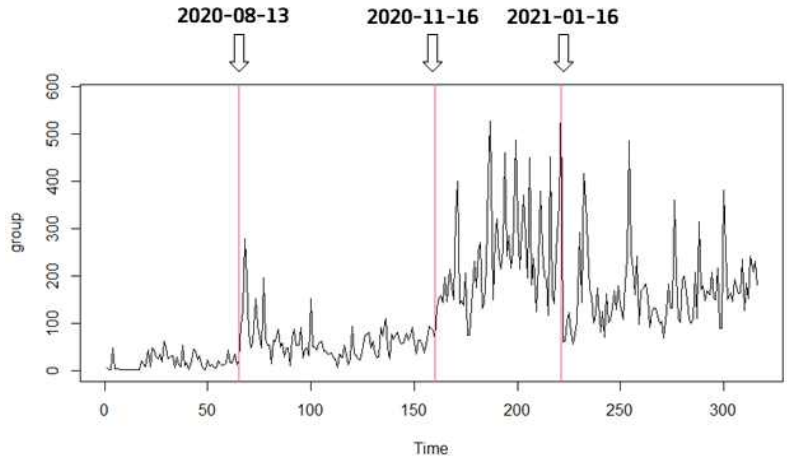

Figure 16. 집단 감염자 수에 대한 변화점 탐지 결과

집단 감염의 경우 총 3가지 변화점이 관측되었으며, 2020년 8월 13일에 급증한 집단 감염 이후 일정 기간 집단감염이 크게 발생하지 않다가 2020년 11월 16일에 또 한번 급증하는 경향을 보인다. 이는 개인 감염과 마찬가지로 겨울이라는 계절적 특성이 바이러스의 전파에 강한 영향을 준 것으로 볼 수 있으며, 비록 두 그룹으로 나누어 분석을 진행하였더라도 '동일한 바이러스에 대한 감염자 수'를 분석한 것임에는 변화가 없으므로 상당한 일관성을 갖춘 통계적 결과라고 할 수 있다.

또한 2021년 1월 16일에 또 한번의 유의미한 변화점이 감지되었는데 이는 '5인 이상 집합금지'와 확장된 '사회적 거리두기' 등의 강력한 정부의 방역 정책이 시행되었던 시기와 상당한 유사성을 보인다. 따라서 집단 감염자의 수가 정부 정책의 영향을 상당히 많이 받는 것으로 추론할 수 있다.

4. 집단 감염자 수와 개인 감염자 수 사이의 상호연관성 분석

4.1. 모형 식별

앞선 분석에서 집단 감염자 수, 개인 감염자 수가 등분산성 조건을 만족하지 않는 것으로 판단하였다. 따라서 각각을 AR(2)-GARCH(1, 1) 모형에 적합을 진행했는데, 실제로는 두 시계열이 독립적이지 않고 상호 영향을 미치는 시계열일 것이라고 예상하였다. Figure 17에서도 알 수 있듯이, 집단 감염자의 수와 개인 감염자의 수가 증가, 감소하는 추세가 유사하게 나타남을 확인할 수 있고, 감염 자체의 성격을 생각했을 때에도 집단 감염이 개인 감염을, 그리고 개인 감염이 집단 감염을 초래할 것으로 예상되기 때문이다. 따라서 이 효과를 분석하기 위해 다음과 같은 절차를 거쳤다. 먼저 개인 감염자, 집단 감염자의 수 모두 국소적으로 분산이 이분산성임을 파악했고, 변동 폭이 그 시점에서의 확진자의 수와 비례한다고 판단하여 분산 안정화 변환으로 log 변환을 선택했다. 다음으로 그레인저 검정을 이용해서 개인 감염자의 수가 집단 감염자의 수에, 그리고 집단 감염자의 수가 개인 감염자의 수에 대해 영향을 미치는 원인으로써 작용함을 Granger test을 통해 확인하였다. 결론적으로, 이 모델을 VAR 모델에 적합하였다. 모형에 대한 혼란을 덜기 위해 집단 감염자의 수, 개인 감염자의 수와 관련된 항들을 다음과 같이 정의한다. 이후 논의는 모두 분산 안정화 변환을 거친 g_t, i_t에 대해 진행하도록 한다.

- G_t: 집단 감염자의 수
- I_t: 개인 감염자의 수
- $g_t = \log G_t$, $i_t = \log I_t$: 분산 안정화 변환 이후의 집단 및 개인 감염자의 수

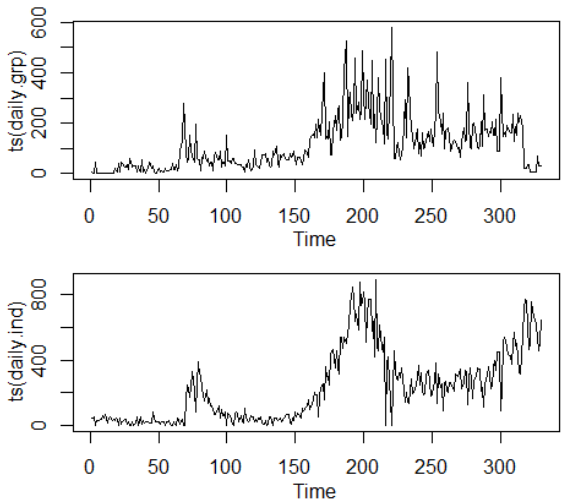

Figure 17. 시간에 따른 집단 및 개인 감염자의 수

4.2. 모형 선택

4.2.1. Granger Causality 검정을 통한 상호 연관성 파악

먼저, Granger Causality 검정을 통해 개인 감염자 수, 집단 감염자 수가 서로가 서로를 유도하는 관계에 있음을 파악하였다. 개인 감염자 수가 집단 감염자 수에 대한 원인이 되는지, 그리고 집단 감염자이 수가 개인 감염자 수에 대한 원인이 되는지를 파악하기 위해 이를 시행했는데, Granger Causality에는 두 가지 종류가 있다. Granger Causality와 Instantaneous Causality 가 그것인데, 각각에 대해 간략히 설명하면 다음과 같다.

먼저 Granger Causality는 두 시계열 A와 B가 있을 때, A가 B에 대한 원인으로 작용하는지 아닌지에 대한 검정을 위해, B를 A에 대한 정보 없이 예측했을 때와 A를 이용해서 예측했을 때의 잔차 제곱합(Sum of square of error)의 감소 비율을 이용하여 검정을 진행한다. 귀무가설이 거짓일수록 이 값이 커질 것임을 이용하여 검정을 진행한다. 즉, A가 B에 대한 원인으로 작용한다면 A에 대한 정보를 이용하여 예측을 한 경우에 잔차 제곱합이 월등히 작게 나타나, 그 비율이 커질 것임을 이용한다.

다음으로 Instantaneous Causality는 두 시계열 A와 B가 있을 때, A가 B에 대한 원인으로 작용하는지 아닌지를 조금 다른 관점에서 해석한다. A의 future value를 알고 있는 것이 B의 future value에 대한 예측에 도움이 되는지의 여부로 귀무가설인 "A가 B에 대한 원인으로 작용하지 않는다" 의 기각 여부를 결정한다.

Granger Causality test에서 귀무가설 하에서의 검정통계량의 값, 그리고 유의수준을 계산한 결과는 다음과 같다.

a) 개인 감염자의 수가 집단 감염자의 수에 대한 원인이 되는가?

H_0 : 개인 감염자의 수가 집단 감염자의 수에 대한 원인이 되지 않음

H_1 : 개인 감염자의 수가 집단 감염자의 수에 대한 원인이 됨

Table 11. Granger Causality test 결과 - 1

인과성의 종류	검정 통계량의 값	유의 수준	검정 결과
Granger Causality	106.98	<2.2e-16	귀무가설 기각
Instantaneous Causality	51.196	8.36e-13	귀무가설 기각

b) 집단 감염자의 수가 개인 감염자의 수에 대한 원인이 되는가?

H_0 : 집단 감염자의 수가 개인 감염자의 수에 대한 원인이 되지 않음

H_1 : 집단 감염자의 수가 개인 감염자의 수에 대한 원인이 됨

Table 12. Granger Causality test 결과 - 2

인과성의 종류	검정 통계량의 값	유의 수준	검정 결과
Granger Causality	73.364	<2.2e-16	귀무가설 기각
Instantaneous Causality	51.196	8.36e-13	귀무가설 기각

따라서, 개인 감염자의 수가 집단 감염자의 수에, 그리고 집단 감염자의 수가 개인 감염자의 수에 대한 원인으로 작용함을 파악하였다.

4.2.2. VAR 모형 적합

VAR (Vector Autoregression) 모형이란 AR 모형을 상호 연관성이 있는 둘 이상의 시계열로 확정한 모형이다. 그 형태 역시 AR 모형과 매우 유사하다. 본 분석에서는 두 개의 시계열(g_t와 i_t)에 대한 VAR 모형을 수립했고, 이 상황에서 VAR(p) 모형의 형태는 다음과 같다.

$$g_t = \sum_{j=1}^{p} \alpha_j g_{t-j} + \sum_{j=1}^{p} \beta_j i_{t-j} + \epsilon_t$$
$$i_t = \sum_{j=1}^{p} \gamma_j g_{t-j} + \sum_{j=1}^{p} \delta_j i_{t-j} + \eta_t$$

4.2.1에서의 Granger test는 두 시계열이 VAR(1) 모델을 따른다는 가정 하에 진행되는 것이다. 따라서 이에 상응하도록 집단 감염자와 개인 감염자의 수를 VAR(1) 모델에 적합을 하는 것이 자연스러우나, 다른 모델들과의 비교 이후 더 적합한 모형이 있다면 이에 대해 다시 모델을 수립하는 것이 합리적이므로, 이 절차에 따라 진행하였다. 이에 따라 VAR(2), VAR(3) 등의 higher order model 들에 대해서도 적합을 해본 결과, 이 모델들에서는 모수들이 유의미하지 않다는 결론이 도출되었다. 차수가 커질수록 유의미하지 않다고 판단되는 모수의 개수는 증가하였고, 따라서 VAR(2), VAR(3) 모형에 데이터를 적합하고 모수에 대한 유의성 검정을 실행한 결과만을 Table 13에 나열하였다. Table 13에서 확인할 수 있듯이, 모수의 개수가 늘어날수록 유의하지 않은 모수의 수 역시 늘어남을 확인할 수 있다. 유의수준 $\alpha = 0.05$하에서, 유의하지 않은 모수의 개수는 VAR(2) 모형의 경우 8개 중 2개, VAR(3) 모형의 경우 12개 중 6개로 나타났다.

Table 13. VAR(2), VAR(3) 모형에 대한 모수 유의성 검정

model	Parameter	Estimate	Std. Error	t value	Pr(>|t|)
VAR(2)	α_1	0.56634	0.05951	9.516	<2e-16
	β_1	0.11931	0.03952	3.019	0.00274
	α_2	0.25479	0.06044	4.216	3.24e-05
	β_2	0.03931	0.03888	1.011	0.31272
	γ_1	0.22058	0.08970	2.459	0.0144
	δ_1	0.43894	0.05956	7.369	1.44e-12
	γ_2	0.10376	0.09109	1.139	0.2555
	δ_1	0.26910	0.05859	4.593	6.27e-06
VAR(3)	α_1	0.52208	0.06170	8.462	9.49e-16
	β_1	0.11676	0.04098	2.849	0.00466
	α_2	0.18910	0.07103	2.662	0.00815
	β_2	0.03255	0.04496	0.724	0.46967
	α_3	0.13576	0.06237	2.177	0.03024
	β_3	-0.01310	0.04032	-0.325	0.74536
	γ_1	0.16929	0.09095	1.861	0.0636
	δ_1	0.35423	0.06040	5.865	1.12e-08
	γ_2	-0.0177	0.10471	-0.170	0.8654
	δ_2	0.11860	0.06628	1.789	0.0745
	γ_3	0.11216	0.09194	1.220	0.2234
	δ_3	0.29167	0.05943	4.908	1.47e-06

4.2.3. 정보기준 및 RMSE의 비교를 통한 모형 비교 및 모형 선정

VAR(n) 모델에서 n이 1, 2, 3, 4인 경우에 대해 정보기준인 AIC, BIC를 모두 계산하고, RMSE를 비교한 결과는 Table 14와 같다. 여기서 확인할 수 있듯이, higher order model에 적합할수록 AIC, BIC 및 RMSE가 감소하여 더 잘 적합된다는 결론이 도출되었으나, 앞선 분석에서 모수들이 유의하지 않다고 판단되었으므로 최종적으로 VAR(1) 모델에 적합하였다. 추가적으로, g_t와 i_t의 상관계수가 0.4368임을 파악하여, 다중 공선성 문제 역시 발생하지 않음을 확인하였다.

Table 14. VAR 모형에 대한 AIC, BIC, RMSE

Model	p	AIC	BIC	RMSE	
				ϵ_t	η_t
VAR(p)	1	1604.501	1619.698	0.4867	1.1060
	2	1555.767	1586.160	0.4569	1.0378
	3	1521.618	1567.207	0.4427	0.9619
	4	1489.025	1549.819	0.4130	0.9115

4.2.4. 선정된 모형 적합 결과

g_t, i_t를 VAR(1) 모형에 적합한 결과, 모수에 대한 유의성 검정 결과는 아래의 표와 같다.

Table 15. VAR(1) 모형에 대한 모수 유의성 검정 결과

model	Parameter	Estimate	Std. Error	t value	Pr(>|t|)
VAR(1)	α_1	0.77789	0.02574	30.217	2e-16
	β_1	0.19560	0.02284	8.565	4.31e-16
	γ_1	0.40139	0.03881	10.34	<2e-16
	δ_1	0.63657	0.03442	18.49	<2e-16

추가적으로, fitted value를 실제의 그래프와 비교해본 결과는 다음과 같이 나타났다.

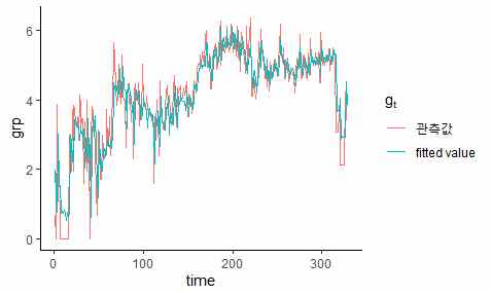

Figure 18. g_t에 대한 fitting 결과

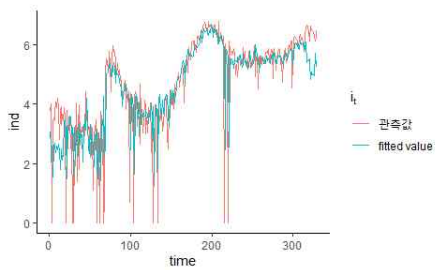

Figure 19. g_t에 대한 fitting 결과

Figure 18, 19에서 확인할 수 있듯이, fitted value가 관측된 g_t, i_t를 잘 적합함을 확인할 수 있다.

4.3. 모형 진단

이후 모형 진단에서는, g_t, i_t를 VAR(1) 모형에 적합한 이후 잔차에 대한 분석을 진행한다. Table 15에서의 모수 유의성 검정을 거쳐 수립한 최종적인 모델은 다음과 같다.

$$g_t = 0.77789 g_{t-1} + 0.19560 i_{t-1} + \epsilon_t$$
$$i_t = 0.40139 g_{t-1} + 0.63657 i_{t-1} + \eta_t$$

이후 이 모형에서 나타나는 잔차(ϵ_t, η_t)로 모형을 진단한다.

4.3.1. 정규성 검정

먼저 잔차들에 대한 정규성 검정을 시행하였다. g_t, i_t에 대해 VAR(1) 모형에서 나타난 잔차를

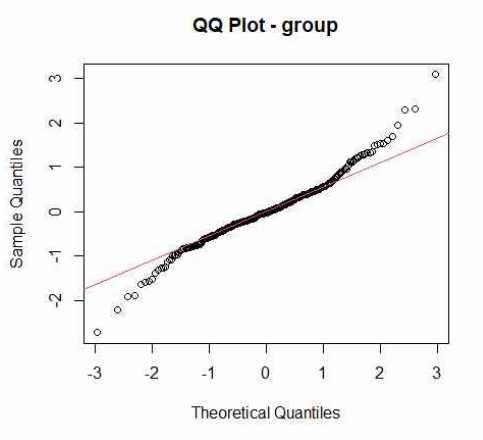

Figure 20. Q-Q Plot (ϵ_t)

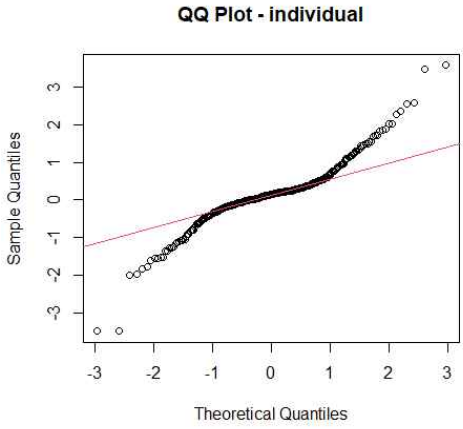

Figure 21. Q-Q Plot (η_t)

시각적으로 표현하기 위해 QQplot을 도시한 결과는 위와 같다. 두 경우 모두 잔차가 0과 가까운 부분에서는 잔차들이 정규분포를 잘 따르는 것으로 나타났으나, 그렇지 않은 부분에서는 점차 정규분포와 다른 양상을 나타낸다는 것을 확인하였다. 정량적으로 이를 검정하기 위해 Jarque-Bera test를 진행하여 잔차들이 정규분포를 따르는지를 검정하였다. 그 결과 아래와 같이 두 잔차 모두 정규분포를 따른다고 보기 어려움을 확인하였다.

H_0 : 오차항이 정규분포를 따른다.

H_1 : 오차항이 정규분포를 따르지 않는다.

Table 16. 정규성 검정을 위한 Jarque-Bera test 결과

구분	Group	Individual
statistics	70.402	945.85
p-value	5.551e-16	2e-16
Test result	귀무가설 기각	귀무가설 기각

따라서 g_t, i_t 모두에 대해서 오차항이 정규분포를 따른다는 귀무가설을 기각하게 된다.

4.3.2. 평균 및 등분산성 검정

이어서 잔차들의 평균 및 등분산성 검정을 실행하였다. 잔차들의 평균의 경우, 계산을 통해 다음과 같이 평균이 0에 가까움을 쉽게 확인할 수 있었다.

Table 17. 잔차들의 평균

	ϵ_t	η_t
mean	0.01422	0.03884

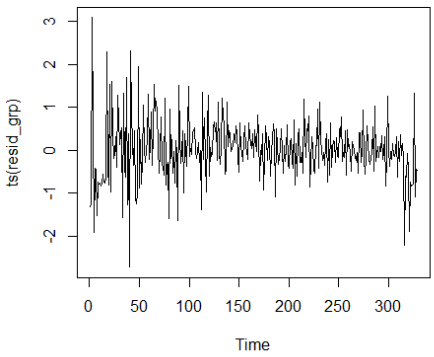

Figure 22. 잔차 그래프 - g_t

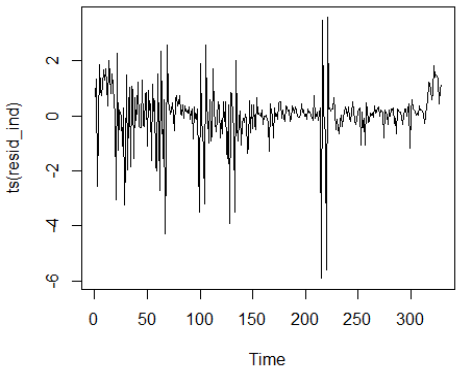

Figure 23. 잔차 그래프 - i_t

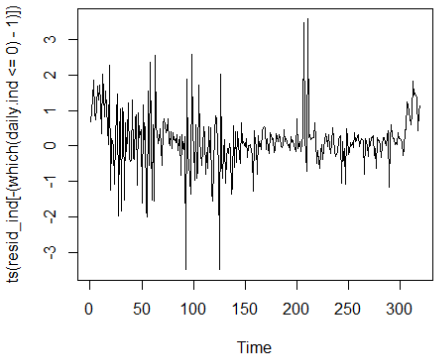

Figure 24. 잔차 그래프 - i_t (결측치 제거)

다음으로 등분산 가정을 시각적으로 확인하기 위해 잔차들의 plot을 도시하였다. g_t에 대한 잔차의 경우, 집단 감염자 수에 대한 데이터가 부족하였던 초반부를 제외하면 등분산성이라고 가정할 수 있음을 figure 22에서 확인할 수 있고, 개인 감염자의 경우는 등분산성이라고 가정하기 어려운 것으로 나타났으나, 집단 감염자의 수가 전체 감염자의 수보다 더 많아 개인 감염자의 수에 대한 결측치를 복원하였던 날짜를 제외하고 나니 등분산 가정이 비교적 잘 만족됨을 하였고, 그 결과는 Figure 24와 같다.

4.3.3. 독립성 검정(자기상관 검정)

잔차들 사이의 자기상관을 분석하기 위해, 먼저 시각적인 관찰을 위해 자기상관함수와 부분자기상관함수를 plot하고, 이후 Ljung-Box test를 이용하여 이를 정량적으로 분석하였다. 앞에서와 동일하게, lag를 총 1, 3, 5, 10, 20의 5가지로 설정하고 각각의 경우에 대한 p-value를 조사하였다. 그 결과, 자기상관함수와 부분자기상관함수의 그래프에서는 잔차들의 자기상관성이 크지 않은 것으로 나타났으나, Ljung-Box test 진행 결과 잔차들이 자기상관되어있음을 확인할 수 있었다. 그 결과는 아래와 같다.

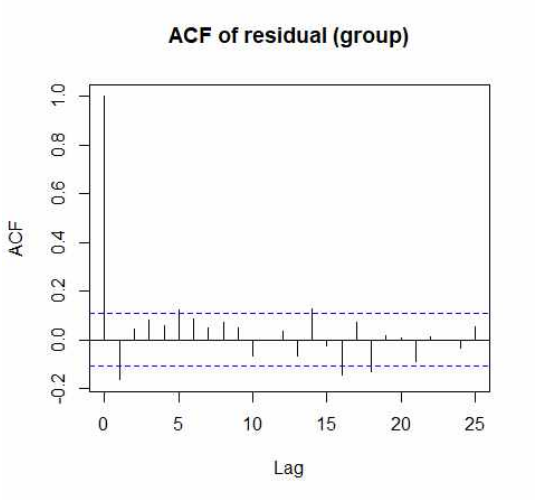

Figure 25. ϵ_t - 자기상관함수

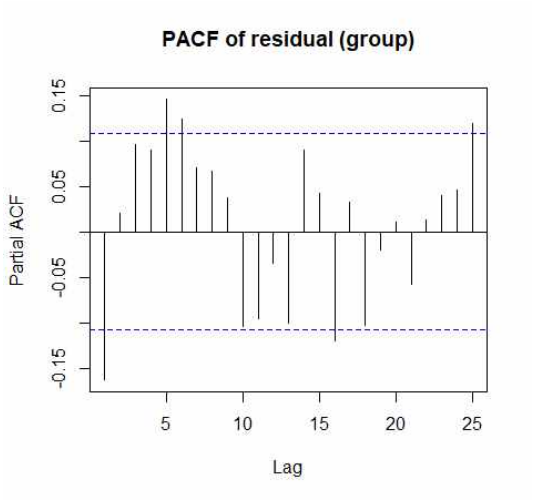

Figure 26. ϵ_t - 부분자기상관함수

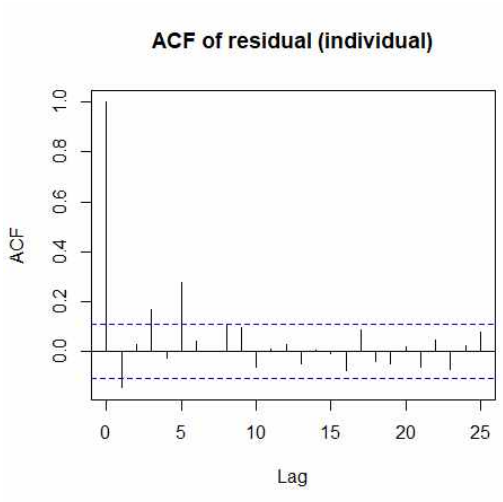

Figure 27. η_t - 자기상관함수

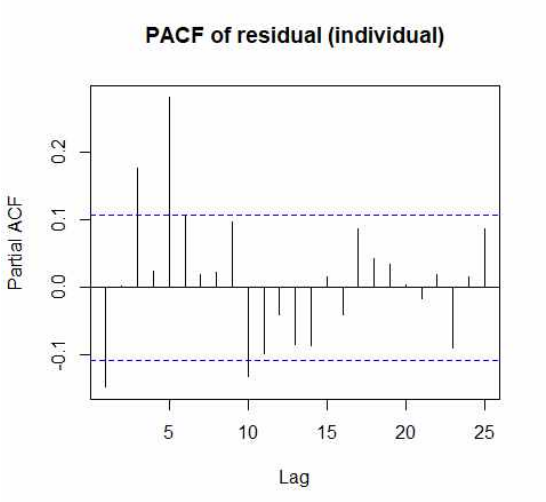

Figure 28. η_t - 부분자기상관함수

Table 18. ϵ_t, η_t에 대한 Ljung-Box test 진행 결과

		Group	Individual
P value	lag: 1	0.002959	0.007030
	lag: 3	0.007971	0.000712
	lag: 5	0.002856	3.792e-08
	lag: 10	0.004271	1.174e-07
	lag: 20	0.000453	8.993e-06
검정 결과		귀무가설 기각	귀무가설 기각

4.3.4. 모형 진단 결과 요약

g_t, i_t에 대한 VAR(1) 모형에서 잔차들이 조건을 잘 만족하지는 않았다. 등분산성 가정은 비교적 잘 만족한다고 볼 수 있었으나, 잔차들이 정규분포를 따른다고 할 수 없음을 Jarque-Bera test를 통해 확인하였고, 잔차들 사이에 자기상관이 없다고 할 수 없음 역시 Ljung-Box test를 통해 확인하였다. 하지만 Granger test를 통해 g_t와 i_t가 서로에 대한 원인이 된다는 것을 확인하였고, 여러 VAR 모형들과 AIC, BIC, RMSE, 모수의 유의성 등을 비교해본 결과 VAR(1) 모형이 가장 타당하다고 판단하여 이 모형을 최종적으로 선택하였다.

4.4. 추가 논의

4.4.1. 향후 확진자 수 경향성 예측

3 에서의 변화점 탐지 검정 결과를 이용하여, 정책의 변화 또는 계절적 특성이 없는 경우 변화점이 나타나지 않고 현재의 모델에 대한 적합이 계속 유효하다고 가정한 뒤 forecasting을 진행하였다. 그 결과 집단 감염자의 수와 개인 감염자의 수 모두 증가하지 않는 경향성을 확인하였고, 각각이 모두 감소하는 추세가 나타날 것으로 확인했다.

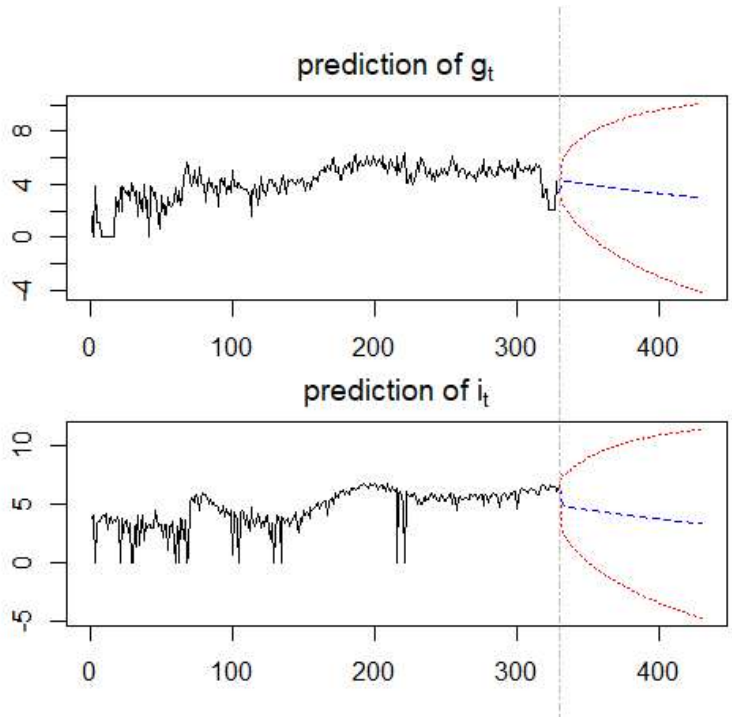

Figure 29. 향후 100일 간의 g_t, i_t forecasting 결과

5. 결론

집단 감염자의 수와 개인 감염자의 수를 각각 AR(2)-GARCH(1, 1) 모형에 적합하는 것이 가장 타당함을 여러 모형을 비교해봄으로써 확인하였고, 변화점 탐지를 통해 계절적 특성 및 사회적 거리두기 정책이 개인, 집단 감염자 수에 대한 변화점을 만들어내며, 사회적 거리두기 정책은 개인 감염자의 수보다는 집단 감염자의 수를 억제하는 데에 초점이 맞추어진 정책임을 확인할 수 있었다.

다음으로, 집단 감염자의 수와 개인 감염자의 수 사이의 연관성을 보이고, 이를 VAR(1) 모형에 적합하여 분석하였다. 이 과정을 통해 집단 감염자 수의 감소가 곧 개인 감염자 수의 감소로 이어질 것으로 추론하였고, 결론적으로 사회적 거리두기 정책이 전체적인 코로나 확진자의 감소에 대한 실효성이 있을 것으로 예상할 수 있었다.

최종적으로, 현 정책이 그대로 유지된다고 가정한 결과, forecasting을 통해 추가적인 변화점이 발생하지 않는 한 집단 감염자의 수와 개인 감염자 수 모두 완만하게 감소하는 형태가 될 것으로 예상할 수 있었다.

6. Appendix: 데이터 수집에 사용한 코드 첨부

데이터 크롤링 코드 (Python)

```
import datetime
import requests
import re
import logging
import pandas as pd
from tqdm import tqdm
from bs4 import BeautifulSoup
from dateutil.parser import parse
from collections import defaultdict


def parse_board_link(onclick: str) -> str:
    """보도자료 게시글의 html 에서 게시글 링크의 onclick attribute에서 url를 추출"""
    # Looks like: "fn_tcm_boardView('/tcmBoardView.do','3','31','5301','312', 'BDJ' );"
    # Which redirects to "http://ncov.mohw.go.kr/tcmBoardView.do?brdId=3&brdGubun=31&dataGubun=&ncvContSeq=5301&contSeq=5301&board_id=312&gubun=BDJ"
    assert "fn_tcm_boardView('/tcmBoardView.do'," in onclick
    _, brdId, brdGubun, ncvContSeq, board_id, gubun = map(lambda x: x.strip().replace("'", "").replace(" );", ""), onclick.split(','))
    if (not brdId) and (not ncvContSeq):
        logging.warning(onclick)
    return f'http://ncov.mohw.go.kr/tcmBoardView.do?brdId={brdId}&brdGubun={brdGubun}&dataGubun=&ncvContSeq={ncvContSeq}&contSeq={ncvContSeq}&board_id={board_id}&gubun={gubun}'

# 게시판 페이지 크롤링
pages = 160  # might change in the future. please check
page_responses = []
for page_index in tqdm(range(1, pages + 1)):
    url = f'http://ncov.mohw.go.kr/tcmBoardList.do?pageIndex={page_index}&brdId=&brdGubun=&board_id=140&search_item=1&search_content='
    response = requests.get(url)
    response.raise_for_status()
    page_responses.append(response)

# 게시판 페이지에서 게시글 링크 추출
board_links = []
for page_response in tqdm(page_responses):
    soup = BeautifulSoup(page_response.text)
    ta_ls = soup.select(".ta_l")
    for ta_l in ta_ls:
        onclick = ta_l.a.attrs['onclick']
        board_link = parse_board_link(onclick)
        board_links.append(board_link)
```

```
print(f"crawled {len(board_links)} pages") # should be ~ 10 * pages (current 1600)

# 추출한 게시글 링크에서 게시글 html 다운로드
board_link_responses = []
for board_link in tqdm(board_links):
    response = requests.get(board_link)
    response.raise_for_status()
    board_link_responses.append(response)

# 다운받은 html를 BeautifulSoup 객체로 변환
soups = [BeautifulSoup(response) for response in tqdm(board_link_responses)]

# 각 BeautifulSoup 객체에서 필요한 정보 파싱하는 함수 작성
def parse_date(soup):
    # 담당, 작성일(, 수정일)
    bvc_details = soup.select("span.bvc_detail")
    if len(bvc_details) not in (2, 3):
        print(bvc_details)
        assert False
    return parse(bvc_details[1].text).isoformat()

def parse_content(soup):
    bv_contents = soup.select('.bv_content')
    assert len(bv_contents) == 1
    bv_content = bv_contents[0]
    content = bv_content.text.replace('\xa0', '').replace('\r', '')  # remove whitespace characters
    return content

def parse_title(soup):
    bv_ttls = soup.select('.bv_ttl')
    assert len(bv_ttls) == 1
    bv_ttl = bv_ttls[0]
    title = bv_ttl.text.strip()
    return title

def get_title_and_related_contents(soup):
    related_rows = []
    content = parse_content(soup)
    for row in content.split('\n'):
        # 집단감염은 "(어디어디 관련)" 이라는 단어가 들어감
        if '관련)' in row or '관련하여' in row:
            related_rows.append(row.strip())
    related_contents = '\n'.join(related_rows)

    title = parse_title(soup)
    return title, related_contents

# 전체 soup들에 대해 필요한 정보를 dataframe으로 파싱
titles = []
dates = []
relateds = []
```

```
for soup in tqdm(soups):
    title, related = get_title_and_related_contents(soup)
    date = parse_date(soup)

    titles.append(title)
    dates.append(date)
    relateds.append(related)

related_contents_df = pd.DataFrame({
    'title': titles,
    'date': dates,
    'related_contents': relateds
})

# 그냥 이름 줄이려고 renaming
df = related_contents_df

# 정례브리핑 아닌 보도자료이거나 관련내용 없으면 제거
df = df[(~df.title.str.contains('보도')) & (df.related_contents != '')]

# 관련있는 보도자료 패턴 (공백제거)
title_patterns = [
    r'코로나19국내발생및예방접종현황\((\d+)월(\d+)일(?:,정례브리핑)?\)',
    r'코로나19국내발생및예방접종현황\((\d+)\.(\d+)\.(?:,정례브리핑)?\)',
    r'코로나바이러스감염증-19국내발생현황\((\d+)월(\d+)일(?:,정례브리핑)?\)',
    r'코로나바이러스감염증-19국내발생현황\((\d+)\.(\d+)\.(?:,정례브리핑)?월?\)',
    r'\((\d+)월(\d+)일(?:,정례브리핑)?\)',
    r'\((\d+)\.(\d+)\.(?:,정례브리핑)?월?\)',
    r'코로나바이러스감염증-19국내발생현황\((\d+)월(\d+)일0시\)',
    '정례브리핑',
]

def is_related(title: str) -> bool:
    """게시글 제목보고 관련있는 보도자료인지 판단하는 함수"""
    s = title.replace(' ', '')
    for pattern in title_patterns:
        if re.search(pattern, s):
            return True
    return False

# 게시글 제목이 관련이 없는 내용이면 제거
df['is_related'] = df['title'].apply(is_related)
df = df[df['is_related'] == True]

# 피와 땀으로 패턴화한 모든 확진자 데이터 regex 패턴
RELATED_CONTENTS_RE_LIST = [
    re.compile(r'(?:-|○|^)\((.+?)관련\).+[^\d](?:총|누적확진자는|누적환자는|누적확진자가)([,\d]+)명\*?\*?이(?:확진되
었)?다\.'),
    re.compile(r'(?:-|○|^)(.+?)[와과]?관련하여.+(?:총|누적확진자는|누적환자는|누적확진자가)([,\d]+)명?\*?\*?이(?:확
진되었)?다\.'),
]
```

```
def datetime_str_to_date_str(datetime_str: str) -> str:
    """'YYYY-mm-dd' 형식의 문자열을 'YYYY-mm-ddT'"""
    return datetime.datetime.fromisoformat(datetime_str).date().isoformat()

# 일자, 집단명, 총 확진자수 를 담는 데이터 그릇
new_df_rows = []

for _, row in df.iterrows():
    date = datetime_str_to_date_str(row.date)
    for line in row.related_contents.split('\n'):
        line = line.replace(' ', '')
        for regex in RELATED_CONTENTS_RE_LIST:
            match = regex.search(line)
            if match:
                cause, count_str = match.groups()
                cause = cause.replace('*', '')  # remove notes
                count = int(count_str.replace(',', ''))
                new_df_rows.append([date, cause, count])
                break


# 같은 집단끼리 (일자, 총 확진자 수) 묶음
per_cause_dict = dict()
for date, cause, count in reversed(new_df_rows):
    if cause not in per_cause_dict:
        per_cause_dict[cause] = []

    per_cause_dict[cause].append(
        (date, count)
    )

# 일자별로 집단감염 확진자 수 계산
new_df = []
per_cause = defaultdict(list)
count_per_date = {}

for date, cause, count in reversed(new_df_rows):
    per_cause[cause] = count
    count_per_date[date] = sum(per_cause.values())

# 데이터 저장
pd.DataFrame({
    '날짜': list(count_per_date.keys()),
    '집단감염자수': list(count_per_date.values()),
}).to_excel("집단감염자수.xlsx")
```

Appendix.B. 시계열 분석 코드 (R)