한국의 코로나 추이 변화에 영향을 미친 변수에 관한 연구

발제자: 박지수, 유병민, 최민서

본 연구는 2020년 1월 코로나-19의 발병 이후 한국의 일일 확진자 수에 영향을 준 변수를 찾기 위해 다변량 시계열 분석을 진행한 것이다. 독립변수로는 레그값, 일일 재난문자 발송량, 일일 코로나 검사량, 일일 전국 대중교통 정거장 이용량, 일일 전국 고속도로 이용량, 일일 마스크 공급량이 사용되었다. 또한, 전파와 확진 사이의 시간 차이를 고려하여 종속변수와 독립변수 간에는 7일의 간격을 두었다. 선형 회귀분석 결과 일일 재난문자 발송량과 일일 코로나 검사량, 일일 전국 대중교통 정거장 이용량이 유의한 변수인 것으로 나타났으며, 예상과 달리 일일 재난문자 발송량은 일주일 뒤 일일 확진자 수와 양의 상관관계를 가졌다. 이 같은 결과를 확정하고 모델의 적합성을 검증하기 위해 다중공산성 검정, 스케일링, 이분산성 검정, Hausman Test, 정상성 검정, 정규성 검정, 선형성 검정 등이 시행됐다. 결론적으로 코로나 확산 추이를 막는데 기여한 변수는 일일 코로나 검사량뿐인 것으로 드러났다. 다만 코로나 사태가 현재 진행 중이기에 데이터가 부족한 점, 선행 연구의 부족으로 독립변수 선정에 있어 이론적 정당성이 다소 떨어지는 점은 아쉬움으로 남는다.

0. 서론

2019년 12월 12일 중국 후베이성 우한시에서 코로나바이러스가 최초 보고[1]된 뒤 한 달만인 2020년 1월 20일, 한국에서 첫 확진자가 발생하였다. 이후 약 4달이 지난 지금까지도 바이러스는 종식되지 않고 있으며, 확진자 수는 증감을 거듭하고 있다. 전문가들과 정부 관계자들은 코로나 이전의 일상으로 돌아갈 수 없다고 말한다.[2] 사태가 이렇게까지 흐르니 '정부는 뭘 했을까?'라는 의문이 자연스럽게 들기 마련이다. 우리는 정부가 아주 많은 것을 했다는 것을, 관계자들이 피땀 흘려 노력했다는 것을 안다. 피상적으로만. 그래서 우리는 어떤 정책들이 코로나 바이러스의 확산에 있어 변화를 가져왔는지 가려내고 싶었다. 비판과 찬사는 실증적 근거를 바탕으로 해야 하기 때문이다. 본 연구는 여러 가지 변수를 설정하고, 일일 확진자 수에 이러한 변수들이 어떤 영향을 미쳤는가에 관해 다변량 시계열 회귀분석을 수행한 것이다. 본 보고서는 특히 모델의 타당성을 검증하는 데 많은 기술을 할애하고 있으니 시계열 분석을 진행하려는 다른 예비연구자들에게 유용하게 쓰일 수 있을 것이라고 기대해본다. 본 보고서는 다음과 같은 순서로 진행된다. <1. 데이터>에서는 변수의 설정과 그 이유, 사용한 데이터에 대한 설명 및 각 변수에 따라 수립된 가설을 다룬다. <2. 모델 및 결과>와 <3. 검증>에서는 시계열 분석의 결과와 모델의 강건성(robustness)을 검증하기 위한 여러 가지 테스트의 수행, 그리고 모델의 수정 과정을 다룬다. <4. 결론>에서는 검증을 통과한 모델의 계수해석과 함의 및 한계를 도출한다.

---

1) 다음을 참조하라. https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)30183-5/fulltext

2) 권다희, "강경화 美 방송 인터뷰 "코로나 이전으로 돌아갈 수 없어"", <머니투데이>, 2020. 05. 22.

1. 데이터

종속변수: $dependent_{t+7}$

확진자 수의 추이 분석을 위해 종속변수로는 위키피디아에 공개된 일일 확진자 수를 사용하였다.[3] 특히, 코로나 바이러스에 노출된 후 발병하기까지 잠복기가 평균 5.1일[4]이라는 점을 고려하여 독립변수에 비해 종속변수의 시계열을 7일 후로 미뤘다. 가령, 3월 4일 독립변수들의 추이가 3월 11일 확진자 수의 추이에 영향을 주었는지를 분석하는 것이다.

독립변수

① 레그값: dep_{t+6}

확진자 수의 추이는 그 전날 확진자 수의 추이와 밀접한 관련이 있을 것이라는 예상에서 레그를 추가하였다. 또한, 데이터의 추이에 트렌드가 있다면 차분(difference)을 통해 이를 제거하려는 목적에서도 모델에 레그를 추가하는 것은 타당성과 편리성을 갖추게끔 한다. 이때 종속변수가 독립변수에 비해 7일 후의 시계열 데이터를 다루므로 레그 변수에는 다른 독립변수에 비해 6일 후의 시계열 데이터가 위치한다. 즉, 다른 독립변수들이 3월 14일의 데이터를 다룬다면 레그 변수에는 3월 20일의 일일 확진자 데이터가 위치하는 것이다.

가설
H1: 일주일 후의 확진자 수는 그 전날의 확진자 수에 영향을 받았을 것이다.

② 일일 재난문자 발송량: $message_t$

메르스 사태의 가장 큰 교훈은 재난 시 정보를 국가가 독점하고 통제하는 것이 질병 확산 방지에 역효과를 낳는다는 것이다.[5] 이는 메르스 사태 당시 메르스의 가장 큰 전염 원인이 '의료쇼핑'이었다는 점에서 잘 나타난다.[6] 즉, 정부와 병원에서 확진자가 방문한 병원명을 공개하지 않아 환자들이 그 병원에서 진료를 보고, 다시 다른 병원으로 옮겨 가 바이러스를 전파하는 역할을 했다는 것이다. 만약 메르스 당시의 교훈을 정부가 잘 학습했다면 정부는 그때와 달리 코로나 국면 초기부터 적극적으로 발병 장소 및 이들이 다녀간 병원 등을 공개했을 것이다. 이와 같은 의지는 실제로 2020년 1월 20일 한국에서 첫 확진자가 발생한 뒤 3일 만인 2020년 1월 23일 행정안전부가 안전안내문자를 시작한 데서 관측된다. 따라서 본 연구는 정부가 재난관련 정보를 적극적으로 공개하는 것이 질병의 확산 추이에 영향을 주는 것으로 보고, 전국 재난문자 일일 발송량을 독립변수 중 하나로 정했다. 데이터의 출처는 국민재난안전포털에서 제공하는 재난문자 데이터이다.[7]

---

3) 다음을 참조하라. https://en.wikipedia.org/wiki/COVID-19_pandemic_in_South_Korea
4) 다음을 참조하라. https://pubmed.ncbi.nlm.nih.gov/32150748/
5) 다음을 참조하라. 보건복지부, 『2015 메르스 백서』, 보건복지부, 2016.07.
6) 손종관, ""의료쇼핑이 메르스 화 키웠다."", <Medical Observer>, 2015.06.18
7) 다음을 참조하라. http://www.safekorea.go.kr/idsiSFK/neo/sfk/cs/sfc/dis/disasterMsgList.jsp?

가설
H2: 일일 재난문자가 발송량이 많을수록 일주일 후의 확진자 수는 감소할 것이다.

③ 일일 코로나 검사량: $testdaily_t$

한국의 우수한 코로나 대응력은 코로나 발병 초기부터 발 빠르게 검사 수를 늘려간 데 있다.[8] 전염병 발생 초기에 진단이 활발하다면 확진자를 선별, 격리할 수 있고 그만큼 전염병의 확산을 막거나 늦출 수 있기 때문이다. 이에 본 연구는 일일 코로나 검사량이 7일 후의 확진자 추이에 영향을 미쳤을 거라고 예상했고, 독립변수 중 하나로 선정하였다. 일일 코로나 검사량 데이터는 Our World in Data에서 제공하는 코로나 일일 검사량을 이용하였다.[9]

가설
H3: 일일 코로나 검사량이 많을수록 일주일 뒤의 확진자 수는 감소할 것이다.

④ 일일 서울시 버스 이용량: bus_t

사회적 거리두기 캠페인이 효과가 있다면 대중교통 이용량과 차량 통행량 모두 감소했을 것이다. 그리고 교통량의 감소는 일일 확진자 수에 영향을 미쳤을 것이다. 이에 본 연구에서는 대중교통 이용량으로 서울시 열린데이터 광장에서 제공하는 서울시 버스노선별, 정류장별 승하차 인원 정보 데이터를 이용하였다.[10] 그러나 종속변수와 다른 독립변수들의 표본이 전국 단위인 데 비해 버스 이용량의 표본은 서울시에 한정되므로 변수들의 모집단이 서로 다른 결과를 낳게 되었다. 따라서 일일 서울시 버스 이용량이 독립변수로서 적절하지 않다고 판단, 후술할 전국 버스 정류장 데이터로 대체하였다.

⑤ 일일 전국 대중교통 정거장 이용량 : $Transit_t$

상술한 바와 같이 서울시 버스 이용량은 일일 전국 대중교통 정거장 이용량으로 대체되었다. 데이터는 구글에서 제공하는 『코로나 19 지역사회 이동성 보고서』[11]의 대중교통 정거장 데이터를 이용하였다. 또한, 데이터는 단위 값으로 주어지지 않고 기준 일을 정해 일마다 얼마나 변화했는지 퍼센티지(%)로 제공됐으므로, 기준 일의 단위 값을 100으로 선정한 후 일일 변화율을 반영하여 단위 값으로 변환하였다.[12]

---

menuSeq=679
8) Chad Terhune et al. “Special Report: How Korea trounced U.S. in race to test people for coronavirus”, Reuters, 2020. 03.19.
9) 다음을 참조하라. https://ourworldindata.org/coronavirus-testing
10) 다음을 참조하라. https://data.seoul.go.kr/dataList/OA-12912/S/1/datasetView.do
11) 다음을 참조하라. https://www.google.com/covid19/mobility/data_documentation.html?hl=ko
12) 구글 데이터의 구체적인 취합 방법 및 구성에 대해서는 다음을 참조하라. https://support.google.com/covid19-mobility/answer/9824897?hl=ko&ref_topic=9822927

가설
H4: 일일 전국 대중교통 정거장 이용량이 많을수록 일주일 후의 확진자 수는 증가할 것이다.

⑥ 일일 전국 고속도로 이용량 : $traffic_t$

대중교통 이용량에 더해 전국 고속도로 이용량도 변수로 추가되었다. 전국 고속도로 이용량은 한국도로공사의 고속도로 공공데이터 포털에서 제공하는 일일 전국교통량 데이터를 이용하였다.[13)]

가설
H5: 일일 전국 고속도로 이용량이 많을수록 일주일 후의 확진자 수는 증가할 것이다.

⑦ 일일 마스크 공급량: $mask_t$

정부는 코로나 발병 초기부터 마스크 착용을 권장하였다. 이로 인해 한때 마스크 품귀현상이 벌어지면서, 정부가 직접 마스크 공급업체들과 계약을 맺고 공적 마스크라는 형태로 시장에 공급하는 『마스크 긴급수급 조정조치』가 2월 26일부로 시행되었다. 이에 본 연구는 일일 마스크 공급량이 일일 확진자 수에 영향을 미쳤을 거라 예상하고 이를 독립변수로 추가하였다. 데이터는 식품의약품안전처에서 제공하는 『마스크 공적판매 수급상황』을 이용하였다.[14)] 단, 정부가 데이터를 집계한 2월 9일부터 3월 11일까지는 데이터의 공시가 마스크 생산량을 대상으로 하지만 3월 12일부터는 오직 공적마스크 공급량만이 발표되어 데이터의 범주가 달라지는 어려움이 있었다. 이에 공적마스크 공급량과 마스크 생산량에 대한 데이터가 함께 제공된 2월 28일부터 3월 11일까지를 비교해본 결과, 마스크 생산량이 공급량에 비해 평균 50% 더 많다는 추계에 도달했다. 또한, 정부의 마스크 긴급수급 조정조치에 따라 각 공급업체의 생산량 중 50%가 공적마스크로 공급된다는 점을 고려했을 때 50%라는 평균값은 타당하다고 판단됐다. 나아가 3월 3일부로 수정된 긴급수급 조정조치에서 정부의 마스크 구입량이 생산량의 50%에서 80%로 상향 조정됐지만, 3월 3일부터 3월 11일까지 데이터를 비교해봤을 때 공적 마스크 공급량이 생산량에 비해 평균 60% 더 많았으므로 50%의 추계를 그대로 유지하였다. 즉, 데이터에 사용된 일일 마스크 공급량은 정부에서 발표한 공적 마스크 공급량 단위 값에 2배를 곱한 값이다.

가설
H6: 일일 마스크 공급량이 많을수록 일주일 뒤의 확진자 수는 감소할 것이다.

2. 모델 및 결과

---

13) 다음을 참조하라. http://data.ex.co.kr/portal/fdwn/view?type=TCS&num=31&requestfrom=dataset

14) 다음을 참조하라. https://www.mfds.go.kr/brd/m_99/list.do

위 데이터를 종합한 모델은 다음과 같다.

$$dependent_{t+7} = \beta_0 + \beta_1 dep_{t+6} + \beta_2 message_t + \beta_3 traffic_t + \beta_4 Transit_t + \beta_5 mask_t + \beta_6 testdaily_t + \epsilon_t$$

회귀분석의 결과는 [표 1]에 제시되어 있다.

[표 1]에 따르면 유의한 변수는 오직 일일 코로나 검사량(test_daily)과 일일 재난문자 발송량(message) 뿐이다. 두 변수는 각각 1% 수준에서, 5% 수준에서 유의하다. 또한, 일일 코로나 검사량의 계수는 예상대로 음의 부호를 나타냈지만 일일 재난문자 발송량의 계수는 예상과 달리 양의 부호를 나타냈다. 즉, 일일 코로나 검사량이 증가하면 일주일 뒤의 확진자 수가 감소하지만, 일일 재난문자 발송량이 증가했을 때에는 일주일 뒤의 확진자 수도 증가한다는 것이다. 이처럼 쉽게 해석하기 어려운 결과와 높은 조건수 (모델 하단에 제시되어 있음)'는 모델의 정합성을 검증하게끔 했다.

[표 1] OLS 결과

OLS Regression Results

Dep. Variable:	dependent_t7	R-squared:	0.519
Model:	OLS	Adj. R-squared:	0.436
Method:	Least Squares	F-statistic:	6.285
Date:	Wed, 27 May 2020	Prob (F-statistic):	0.000149
Time:	18:57:02	Log-Likelihood:	-271.43
No. Observations:	42	AIC:	556.9
Df Residuals:	35	BIC:	569.0
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	118.1296	648.766	0.182	0.857	-1198.935	1435.194
dep_t6	0.1039	0.233	0.446	0.659	-0.370	0.577
message	1.8762	0.866	2.166	0.037	0.118	3.635
traffic	2.362e-05	6.27e-05	0.377	0.709	-0.000	0.000
Transit	0.2411	8.360	0.029	0.977	-16.731	17.213
mask	-0.0418	0.084	-0.499	0.621	-0.212	0.128
test_daily	-0.0340	0.009	-3.723	0.001	-0.053	-0.015

Omnibus:	16.320	Durbin-Watson:	1.222
Prob(Omnibus):	0.000	Jarque-Bera (JB):	20.593
Skew:	1.244	Prob(JB):	3.38e-05
Kurtosis:	5.361	Cond. No.	1.58e+08

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.58e+08. This might indicate that there are
strong multicollinearity or other numerical problems.

3. 검증

[그림 1] VIF 검정

	VIF Factor	features
0	612.8	Intercept
1	4.1	dep_t6
2	6.8	message
3	2.5	test_daily
4	5.3	Transit
5	2.9	traffic
6	2.0	mask

[표 2] 조건수 조정 후 OLS

OLS Regression Results

Dep. Variable:	dependent_t7	R-squared:	0.519
Model:	OLS	Adj. R-squared:	0.436
Method:	Least Squares	F-statistic:	6.285
Date:	Sun, 31 May 2020	Prob (F-statistic):	0.000149
Time:	15:01:44	Log-Likelihood:	-44.242
No. Observations:	42	AIC:	102.5
Df Residuals:	35	BIC:	114.6
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.469e-17	0.117	2.96e-16	1.000	-0.238	0.238
dep_t6	0.1055	0.237	0.446	0.659	-0.375	0.586
message	0.6619	0.306	2.166	0.037	0.042	1.282
traffic	0.0754	0.200	0.377	0.709	-0.331	0.481
Transit	0.0078	0.270	0.029	0.977	-0.540	0.556
mask	-0.0837	0.168	-0.499	0.621	-0.425	0.257
test_daily	-0.6897	0.185	-3.723	0.001	-1.066	-0.314

Omnibus:	16.320	Durbin-Watson:	1.222
Prob(Omnibus):	0.000	Jarque-Bera (JB):	20.593
Skew:	1.244	Prob(JB):	3.38e-05
Kurtosis:	5.361	Cond. No.	6.42

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

①VIF 검정

우선, 높은 조건수를 낳게 하는 원인 중 하나인 다중공산성을 검증하기 위해 VIF 검정을 시행하였다. VIF 검정은 각 변수의 VIF 요소값이 10이 넘어가면 그 변수에 다중공산성이 있는 것으로 해석한다. [그림 1]에서 보이는 것과 같이 상수항을 제외한 모든 변수의 VIF 값은 10을 넘어가지 않았다. 따라서 다중공산성 문제는 아닌 것으로 결론 내렸다.

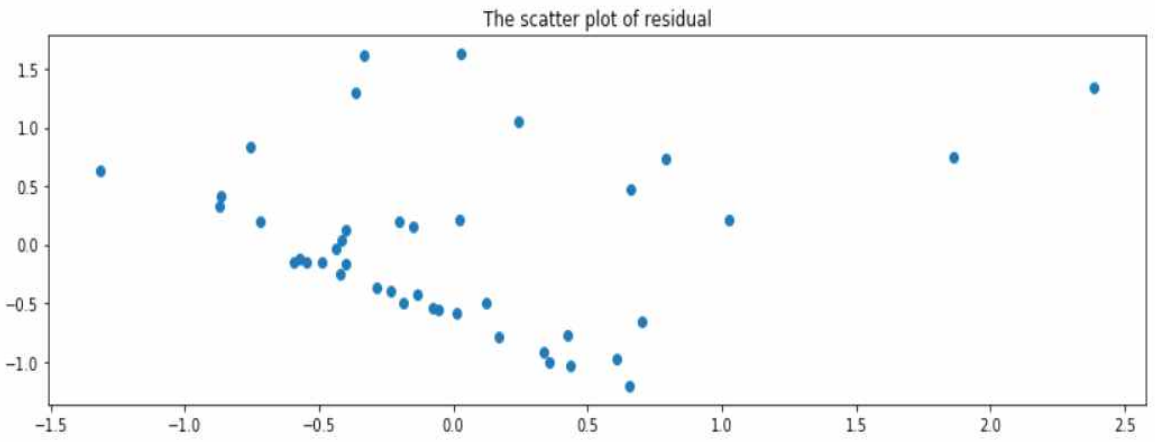
② Standard Scaling

다중공산성에 문제가 없으므로 조건수가 높은 이유는 데이터의 수치적 문제일 가능성이 높아졌다. 실제로 각 데이터의 단위가 일일 확진자는 최대 1000명이지만 일일 검사량은 확진자가 많이 나온 시기 거의 매일같이 10000건 단위가 넘었으므로, 데이터의 규모를 조정할 필요가 있었다. 이에 데이터 전처리 방법으로 Standard Scaling을 선택하여 조정한 후 이를 바탕으로 회귀분석을 진행하였다. 그 결과는 [표 2]에 제시되어 있다. 조건수는 기존 모델에서의 1.58e+08에서 6.42로 크게 감소했으며 변수들의 유의성도 달라지지 않았다.

③ 이분산성 검정

다음으로는 OLS의 기본 가정 중 하나인 이분산성을 검증하였다. 이분산성을 검증하는 방법으로는 대표적으로 잔차의 산점도(scatter plot)가 특정 추세(Trend) 없이 랜덤하게 그려지는지 확인하는 방법과 Breusch-Pagan Lagrange Multiplier test가 있다. 먼저, [그림 2]에 나타난 잔차의 산점도는 랜덤하게 분포한 것으로 보인다. 이를 더욱 정확히 확인하기 위해 시행된 Breusch-Pagan Lagrange Multiplier test의 결과가 [그림 3]에 나타나 있다. Breusch-Pagan Lagrange Multiplier test는 귀무가설이 '이분산성이 존재하지 않는다.'이므로, P값이 유의한 본 모델의 결과 이분산성이 존재하는 것으로 결정됐다. 따라서 이분산성을 조정한 HC1 type의 Robust Standard Error를 사용하여 회귀분석을 진행하였다. 그 결과는 [표 3]에 나타나 있다.

[그림 2] 잔차의 산점도



[그림 3]

```
#이분산성을 검증하는 방법시도: Breusch-Pagan Lagrange Multiplier test for heteroscedasticity
#해석: p값이 significant하므로 이분산성이 존재한다.
_, pval, __, f_pval = statsmodels.stats.diagnostic.het_breuschpagan
(residual, std_scale[['dep_t6', 'message', 'traffic', 'Transit', 'mask', 'test_daily']])
pval, f_pval
```

```
(0.016361119711329117, 0.018683803694130044)
```

[표 3] 이분산성이 조정된 OLS

	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
Intercept	0.00000	0.11727	0.00000	1.00000	-0.23808	0.23808
dep_t6	0.10553	0.23200	0.45487	0.65201	-0.36545	0.57652
message	0.66193	0.30697	2.15635	0.03801	0.03875	1.28511
traffic	0.07536	0.18241	0.41313	0.68203	-0.29496	0.44567
Transit	0.00779	0.25046	0.03109	0.97537	-0.50067	0.51624
mask	-0.08371	0.15706	-0.53296	0.59743	-0.40257	0.23515
test_daily	-0.68975	0.25872	-2.66598	0.01154	-1.21498	-0.16451

④ 내생성(endogeneity) 검정

다음으로 오차항과 독립변수들의 독립성을 검정하는 내생성 검정이 수행되었다. 검정 방법은 Hausman Test를 이용하였다. 만약 독립변수 중 내생성이 있는 변수가 존재한다면 2SLS를 통해 해당 변수의 내생성을 제거해 준 모델을 사용해야 할 것이다. Hausman Test의 검정 결과는 [표 4]에 제시되어 있다.

[표 4] Hausman Test 결과

Table 1 - Hausman Tests

	dep_t6	Transit	traffic	mask
dep_t6	0.32***	0.06	0.26	0.27
	(0.10)	(0.22)	(0.22)	(0.22)
message	0.44***	1.08***	1.37***	0.76***
	(0.16)	(0.33)	(0.36)	(0.27)
test_daily	-0.62***	-0.46**	-1.04***	-1.07***
	(0.18)	(0.19)	(0.20)	(0.21)
Transit	0.01	0.69*	-0.18	-0.15
	(0.26)	(0.36)	(0.24)	(0.24)
traffic	0.10	0.18	1.76***	0.13
	(0.21)	(0.19)	(0.57)	(0.19)
mask	-0.10	-0.11	-0.18	0.91**
	(0.18)	(0.17)	(0.16)	(0.38)
residual_t6	-0.22			
	(0.17)			
residual_Transit		-0.98**		
		(0.39)		
residual_traffic			-1.63***	
			(0.53)	
residual_mask				-1.09***
				(0.37)
R-squared	0.52	0.59	0.62	0.62
No. observations	41	41	41	41

Standard errors in parentheses.
* p<.1, ** p<.05, ***p<.01

Hausman Test는 내생성이 의심되는 각 변수에 대해 수행된다. 상술했다시피 일일 재난문자 발송량($message_t$)과 일일 코로나 검사량($testdaily_t$)을 제외한 모든 변수가 유의하지 않다는 결과가 나왔으므로 두 변수를 제외한 모든 독립변수를 대상으로 Hausman Test가 수행됐다. 이때 '내생성이 없다'는 것이 귀무가설이므로 해당 변수의 잔차가 유의하면 내생성이 있는 것으로 해석한다. [표 4]에서 보이는 바와 같이, 본 연구의 모델에서 내생성이 없는 것은 하루 전 코로나 확진자 수(dep_{t+6})뿐이었다. 따라서 일일 전국 대중교통 정거장 이용량($Transit_t$), 일일 전국 고속도로 이용량($traffic_t$), 일일 마스크 공급량($mask_t$)을 대상으로 2SLS 모델이 사용됐다. 이때, 2SLS 모델의 데이터는 위에서 제시된 것과 마찬가지로 스케일링 된 것이며, 이분산 또한 HC1으로 조정된 것이다. 그 결과는 [표 5]에 제시되어 있다.

[표 5] 2SLS 결과

Table 2 - 2SLS results

	(1)	(2)
dep_t6	-0.01	-0.06
	(0.18)	(0.22)
message	0.89***	0.89**
	(0.32)	(0.34)
test_daily	-0.55*	-0.50*
	(0.29)	(0.25)
predict_Transit	0.24	0.35*
	(0.25)	(0.20)
predict_traffic	0.18	
	(0.20)	
predict_mask	-0.23	-0.18
	(0.16)	(0.13)
Intercept	0.00	0.00
	(0.11)	(0.11)
R-squared	0.57	0.56
No. observations	41	41

Standard errors in parentheses.
* p<.1, ** p<.05, ***p<.01

[표 5]의 (1)은 내생성이 있는 세 변수에 2SLS가 적용된 것이다. 세 변수 모두 2SLS 상에서도 유의하지 않음에는 변화가 없다. 반면 (2)는 일일 전국 고속도로 이용량($predictedtraffic_t$)이 제거된 모델이다. 그 결과는 일일 전국 대중교통 정거장 이용량이 10% 유의수준에서 유의하다는 것을 말해준다. 본 연구는 최종 모델로 2SLS를 적용한 (2) 모델을 택했다. 정리하자면 다음과 같다.

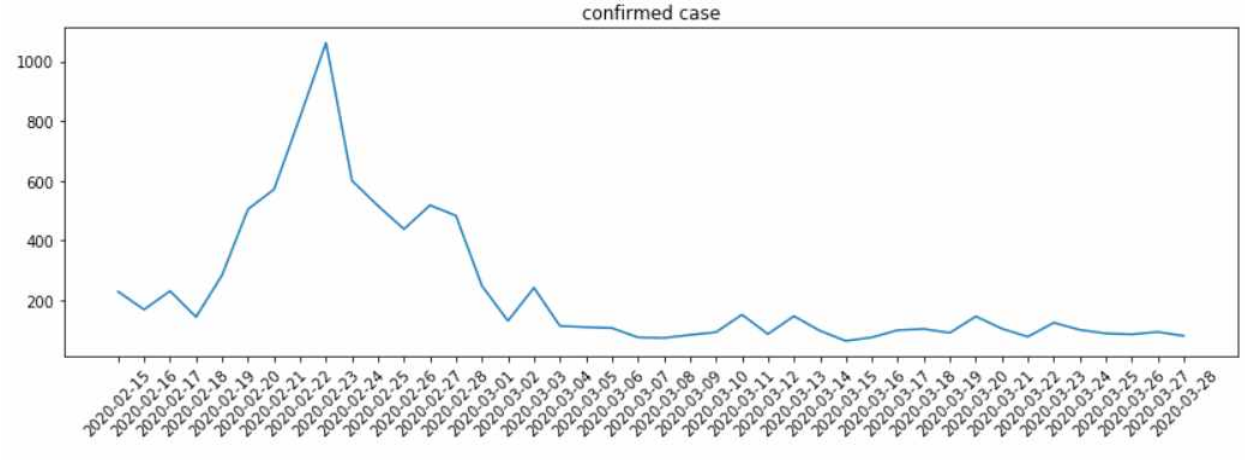
1) 일일 재난문자 발송량은 일주일 뒤의 확진자 수와 1% 유의수준에서 양의 상관관계를 갖는다.
2) 일일 코로나 검사량은 일주일 뒤의 확진자 수와 10% 유의수준에서 음의 상관관계를

갖는다.
3) 일일 전국 대중교통 정거장 이용량은 일주일 뒤의 확진자 수와 10% 유의수준에서 양의 상관관계를 갖는다.

④ 정상성 (Stationarity) 검정

다음으로 데이터의 정상성을 검증하기 위해 일일 확진자 수를 플롯팅하고, 날짜를 중간에서 나눠 일일 확진자 수에 대한 왼쪽 그룹과 오른쪽 그룹의 평균 및 표준 편차를 비교하였다. 만약 두 그룹의 평균 및 표준 편차가 비슷하다면 데이터의 정상성이 확보될 것이다. 그러나 2월 말부터 3월 초 신천지 사태로 폭증한 일일 확진자 수와 그 이후 소강상태로 진입한 사정 때문에 두 그룹의 이 같은 수치에 많은 차이가 있었다. 이는 그림 [4-1]. [4-2] 에 나타나 있다. 정상성 확보를 위한 보다 정확한 검증은 Augmented Dickey-Fuller Test를 수행하는 것이다. ADF Test의 귀무가설은 '데이터가 정상적이지(stationary)않다'는 것이다. 테스트 결과 p값이 0.000000으로 강하게 귀무가설을 기각하므로, 본 모델의 데이터는 정상적(stationary)이라고 할 수 있다. 이는 그림 [5]에 나타나 있다.

[그림 4-1] 일일 확진자 수 플롯팅



[그림 4-2] 양 그룹의 평균과 표준오차

```
confirmed case
mean of left group, right group: 361.4761904761905, 98.80952380952381
std of left group, right group: 260.8573593072695, 24.46552482089654
```

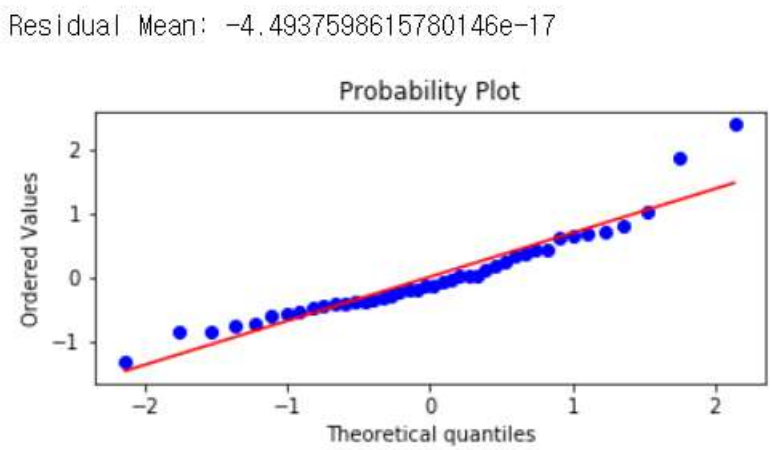
[그림 5] Augmented Dicky-Fuller Test 결과

```
ADF Statistic: -10.249241
p-value: 0.000000
Critical Values:
        1%: -3.646
        5%: -2.954
        10%: -2.616
--------
```

⑤ 정규성(Normality) 검정

다음으로 잔차의 정규성 검증하기 위해 QQ plot을 사용하였다. 만약 잔차가 정규성 가정을 만족한다면 Quantile 값이 직선 주변에 위치할 것이다. [그림 6]은 실제로 잔차가 이 같은 분포를 만족한다는 사실을 보여준다. 또한, 잔차의 평균값이 매우 작다는 점도 주어진 데이터가 정규성 가정을 만족한다는 사실을 뒷받침한다.

[그림 6] QQ Plot



⑥ 선형성(Linearity) 검정

마지막으로 주어진 데이터가 선형성을 만족하는지 검정했다. 선형성이 만족 되기 위해서는 정규성, 비(非) 자기상관성, 작은 조건수 세 조건이 만족 되어야 한다. 정규성과 작은 조건수는 위 검증과정에서 보인 바 있다. 비(非) 자기상관성은 Hausman Test 결과 dep_{t+6}가 외생적이라는 사실을 통해서 자동으로 확보된다. 왜냐하면 자기상관성이 존재하는 경우 dep_{t+6}의 오차항 $dependent_{t+7}$의 오차항 사이에 상관성이 있어야 하는데, 모델에서 dep_{t+6} 자체가 $dependent_{t+7}$의 오차항에 대해 외생적이므로 $Cov(\epsilon_{t+6},\epsilon_{t+7})=0$이 되기 때문이다. 따라서 비(非) 자기상관성까지 모두 만족 되므로 선형성 또한 만족 된다.

4. 결론

위에서 검증한 바와 같이, 본 연구의 모델은 선형회귀 분석의 선형성, 정규성, 비(非)자기상관성 등을 만족하므로 타당한 모델이라고 할 수 있다. 따라서 계수해석과 가설검정이 의미를 갖는다. 최종적인 모델의 OLS 결과는 [표 5]의 (2)에 제시된 것이다. 이에 따르면 일일 재난문자 발송량과 일일 검사량, 일일 전국 대중교통 정거장 이용량만이 유의한 변수이므로, 다음 가설들은 모두 채택되지 않았다.

H1: 일주일 후의 확진자 수는 그 전날의 확진자 수에 영향을 받았을 것이다.

H5: 일일 전국 고속도로 이용량이 많을수록 일주일 후의 확진자 수는 증가할 것이다.

H6: 일일 마스크 공급량이 많을수록 일주일 뒤의 확진자 수는 감소할 것이다.

특히, 예상과 달리 일일 재난문자 발송량과 일주일 후 확진자 수는 양의 관계를 갖는 것으로 나타났다. 즉, 다음 가설은 기각된다.

H2: 일일 재난문자 발송량이 많을수록 일주일 후의 확진자 수는 감소할 것이다.

이는 일일 재난문자 발송량은 매일의 확진자 수를 반영하기 때문이라고 해석할 수 있다. 즉, 일일 재난문자가 많이 발송되고 있다는 것은 그만큼 확진자 수가 증가하고 있다는 것이고, 그에 따라 바이러스의 전파도 활발히 이루어지고 있다는 것을 의미할 가능성이 높다. 이는 다시 일주일 뒤 확진자 수의 증가로 이어져 결국 일일 재난문자 발송량과 일주일 뒤의 확진자 수가 양의 상관관계를 갖게 됐다고 해석할 수 있다.

결론적으로 코로나 확산 추세를 저지하는 유일한 변수는 일일 검사량을 늘리는 것으로 나타났다. 이러한 결과는 다수의 전문가가 초기 검사량을 획기적으로 증가시킨 것을 한국형 방역의 성공 요인으로 꼽은 것과 일치한다. 즉, 일일 검사량의 증가는 확진자를 선별·격리하게끔 해서 새로운 확진자가 더 이상 나오지 않도록 억지하는 효과가 있는 것이다.

본 연구의 한계는 무엇보다 관측치의 수가 42개로 너무나 작았다는 데 있다. 이는 데이터가 충분히 쌓일 만큼 코로나 발병 이후 시간이 오래 지나지 않았기 때문이다. 더군다나 변수마다 데이터가 집계된 시기가 달라 상대적으로 풍족한 데이터가 있는 변수들은 데이터 중 일부가 잘려나갈 수밖에 없었다. 또한, 코로나와 같이 전염성이 매우 강한 바이러스에 대한 선행 연구가 상대적으로 많지 않은 것도 어려움 중 하나였다. 그나마 성격이 비슷한 메르스나 사스에 대한 선행연구를 참조하여 독립변수를 선정했지만, 이론적 정당성이 탄탄한 것은 아니다. 따라서 데이터가 충분히 쌓일 만큼 시간이 지나고 코로나에 대한 실증분석 연구가 활발하게 진행되면 더 나은 분석을 도출해낼 수 있을 것이라 기대해본다.

<참고 문헌>

Chaolin Huang et al, “Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China”, *The Lancet,* Vol.395, (2020): 497-506

Chad Terhune et al. “Special Report: How Korea trounced U.S. in race to test people for coronavirus”, Reuters, 2020. 03.19.

“Community Mobility Reports 고객센터”, 「*Google*」, 2020.05.13., https://support.google.com/covid19-mobility/answer/9824897?hl=ko&ref_topic=9822927

“COVID-19 pandemic in South Korea”, 「*Wikipedia*」, 2020.05.14., https://en.wikipedia.org/wiki/COVID-19_pandemic_in_South_Korea

Stephen A Lauer et al, “The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application”, *Ann Intern Med,* 172(9), (2020): 577-582.

“Statics and Research: Coronavirus (COVID-19) Testing”, 「*Our World in Data*」, 2020.05.21., https://ourworldindata.org/coronavirus-testing

“고속도로 전국 교통량”, 「한국도로공사 고속도로 공공데이터 포털」, 2020.05.12., http://data.ex.co.kr/portal/fdwn/view?type=TCS&num=31&requestfrom=dataset

권다희, “강경화 美 방송 인터뷰 ”코로나 이전으로 돌아갈 수 없어“”, <머니투데이>, 2020. 05. 22.

보건복지부, 『2015 메르스 백서』, 보건복지부, 2016.07.

“보도자료”, 「식품의약품안전처」, 2020.05.14., https://www.mfds.go.kr/brd/m_99/list.do

“서울시 버스노선별 정류장별 승하차 인원 정보”, 「서울 열린데이터 광장」, 2020.05.20., https://data.seoul.go.kr/dataList/OA-12912/S/1/datasetView.do

손종관, ““의료쇼핑이 메르스 화 키웠다.””, <Medical Observer>, 2015.06.18

“재난문자”, 「국민재난안전포털」, 2020.05.14., http://www.safekorea.go.kr/idsiSFK/neo/sfk/cs/sfc/dis/disasterMsgList.jsp?menuSeq=679

“코로나19 지역사회 이동성 보고서”, 「*Google*」, 2020.05.13.,
https://www.google.com/covid19/mobility/data_documentation.html?hl=ko