The Korean Journal of Applied Statistics
2021, Vol. 34, No. 2, 239–254

DOI: https://doi.org/10.5351/KJAS.2021.34.2.239
Print ISSN1225-066X / Online ISSN 2383-5818

Time series analysis for Korean COVID-19 confirmed cases: HAR-TP-T model approach

SeongMin Yu[a], Eunju Hwang[1,a]

[a]Department of Applied Statistics, Gachon University

Abstract

This paper studies time series analysis with estimation and forecasting for Korean COVID-19 confirmed cases, based on the approach of a heterogeneous autoregressive (HAR) model with two-piece t (TP-T) distributed errors. We consider HAR-TP-T time series models and suggest a step-by-step method to estimate HAR coefficients as well as TP-T distribution parameters. In our proposed step-by-step estimation, the ordinary least squares method is utilized to estimate the HAR coefficients while the maximum likelihood estimation (MLE) method is adopted to estimate the TP-T error parameters. A simulation study on the step-by-step method is conducted and it shows a good performance. For the empirical analysis on the Korean COVID-19 confirmed cases, estimates in the HAR-TP-T models of order $p = 2, 3, 4$ are computed along with a couple of selected lags, which include the optimal lags chosen by minimizing the mean squares errors of the models. The estimation results by our proposed method and the solely MLE are compared with some criteria rules. Our proposed step-by-step method outperforms the MLE in two aspects: mean squares error of the HAR model and mean squares difference between the TP-T residuals and their densities. Moreover, forecasting for the Korean COVID-19 confirmed cases is discussed with the optimally selected HAR-TP-T model. Mean absolute percentage error of one-step ahead out-of-sample forecasts is evaluated as 0.0953% in the proposed model. We conclude that our proposed HAR-TP-T time series model with optimally selected lags and its step-by-step estimation provide an accurate forecasting performance for the Korean COVID-19 confirmed cases.

Keywords: COVID-19, heterogeneous autoregressive model, two-piece t distribution, Least squares estimate

1. 서론

2020년 전 세계적으로 확산되고 있는 COVID-19 감염병 대유행은 사망과도 같은 돌이킬 수 없는 치명적 사태를 초래할 뿐만 아니라, 경제 및 문화 활동 등 지역 사회를 마비시키는 국가적 재난 상황을 일으킬 수도 있다. 인구밀도가 높은 한국에서, 외부로부터 전파되는 감염병에 대한 위험도는 매우 크다고 볼 수 있다. 지난 1월 21일, COVID-19 바이러스 확진 판정을 받은 첫 환자가 발생한 이후, 질병관리본부에서 바이러스의 유입과 전파를 차단하고자 노력하고 있다. 지속적인 방역 활동에도 불구하고, 2월에 집단감염이 발생하며 개학 연기

This research was supported by Research Fund of Gachon University (GCU-202003640001).
[1] Corresponding author: Department of Applied Statistics, Gachon University, 1342 Seongnamdaero, Sujeong-gu, Seongnam-si, Gyeonggi-do 13120, South Korea. E-mail: ehwang@gachon.ac.kr

Published 30 April 2021 / journal homepage: http://kjas.or.kr

등의 사회적 파장을 야기시켰고, 8월에 또 한 번의 집단감염 사태가 발생한 이후 산발적인 감염 사례가 속출하고 있다. 감염병에 대한 인식과 대비책이 마련되지 않으면 예기치 않은 감염 폭증이 발생할 경우, 의료체계 붕괴 및 경제적 파장 등 걷잡을 수 없는 사회적 재난 상황을 피할 수 없을 것이다. 본 연구에서는, 연구 개시 시점인 2020년 10월까지의 한국내의 COVID-19 확진자 수에 대한 시계열 자료 분석을 수행하고자 한다. 일일 누적 확진자 수의 자료변환(transformation)을 통한 정상화된(stationary) 시계열 자료를 모형화(modeling)하고, 모수 추정 및 분포 추론 방법을 제시함으로서, 한국 COVID-19 확진자 수에 대한 시계열 모형 제안 및 예측 성능을 제공하는 것을 연구의 목적으로 한다.

최근 COVID-19 감염병 대유행에 관련하여 많은 학자들이 각 나라에서의 COVID-19 확진자 수 및 감염병 관련 랜덤 현상에 관한 확률통계적 이론과 실험 연구를 꾸준히 수행하고 있다. 특히 Benvenuto 등 (2020)과 Ceylan (2020)는 COVID-19 분석 및 예측을 위해 ARIMA 모형을 적용하였고, Kirbas 등 (2020)과 Ribeiro 등 (2020)은 ARIMA모형을 포함하여 다양한 시계열 모형을 이용하여 비교 분석하였다. 또한 Maleki 등 (2020)은 자료변환을 한 정상 시계열 COVID-19 누적 확진자 수에 대해 비대칭성(asymmetry)과 두꺼운 꼬리(heavy tail)를 갖는 확률분포, 즉 two-piece scale mixture of normal (TP-SMN) 분포의 오차과정을 채택하여 AR 모형에 적용한, AR-TP-SMN 모형으로 분석 예측하였다. Arellano-Valle 등 (2005)과 Ghasami 등 (2020)의 TP-SMN 분포는 정규성 가정을 완화하기 위한 방법으로, 대칭 또는 비대칭 분포와, 얕은 또는 두꺼운 꼬리를 가진 분포에 유연하게 적용할 수 있으며, 이는 이상치로부터 로버스트하다. Maleki 등 (2020)의 COVID-19 예측 분석은 AR 모형에 TP-SMN 분포를 고려하여 AR 시계열 모형의 계수들과 오차과정의 모수들을 동시에 최대우도추정법(maximum likelihood estimation; MLE)을 이용하여 추정하였다. Ghasami 등 (2020)에 따르면, TP-SMN 분포 계열 중에서 two-piece t (TP-T) 분포의 오차과정을 갖는 AR-TP-T 모형이 MLE 추정 방법을 이용했을때, 가장 성능이 좋음을 검증하였다. 이러한 이유로 Maleki 등 (2020)도 또한 AR-TP-T 모형을 선택하여 2020년 4월 30일까지의 전세계 COVID-19 누적 확진자 수와 완치자 수에 대한 시계열 분석 및 예측 연구를 수행하였다.

본 연구에서는 한국에서의 COVID-19 확진자 수 분석 및 예측을 위하여 이질적 자기회귀(heterogeneous autoregressive; HAR) 모형을 선택하고, 오차과정으로는 TP-T 분포를 채택한다. HAR 모형은 금융변동성의 장기기억성(long-memory) 특성을 분석 예측하기 위해 Corsi (2009)에 의해 제안되었다. Corsi (2009)는 Müller 등 (1993)에서 보인 이질적 시장가설(heterogeneous market hypothesis)과 Andersen과 Bollerslev (1998)의 실현변동성을 이용하여, 일별, 주별, 월별의 3가지 기간의 차이에 따른 이질성을 반영하는 실현변동성-이질적 자기회귀모형(HAR of realized volatility; HAR-RV)을 미래변동성 예측 모형으로 제안하였다. 그러나, 고정된 시간간격에서의 이동평균을 회귀변수(regressor)로 채택하고 있는 HAR(3) 모형은 이론적으로는 AR(22) 모형이다. HAR(3) 모형에서의 일별, 주별, 월별 대신에 이동평균 기간을 다양하게 고려해 볼 수 있고, 또한, 그것들을 또 다른 모수들로 간주해 볼 수도 있다. HAR 모형에 대해서는 다음 절에서 다시 자세히 다루기로 한다.

본 연구는 2020년 1월 21일부터 2020년 10월 23일까지 국내 일일 누적 확진자 수를 분석하고자, HAR 모형으로 모형화(modeling)하여, 모형 선택 및 추정, 예측 문제를 실증적 데이터 분석과 함께 다룬다. 2020년 10월 23일까지 두 차례의 폭발적 집단감염 사태가 발생하였고, 이로 인한 COVID-19 시계열 데이터는 정상성(stationarity) 특성을 띠고 있지는 않다. 이러한 이유로 로그차분을 두번 시도하며, 누적합 검정(CUSUM test)을 시행하여 로그차분한 데이터의 평균 변화점이 있음을 탐지하였다. 변화점 탐지 시점 이후의 정상(stationary) 데이터를 모형 추정 및 예측을 위해 사용한다. 기존 HAR 모형의 일별, 주별, 월별 기간 대신에, 바이러스 잠복기간으로 알려져 있는 14일을 기준으로 HAR 모형을 설정한다. 그러나, 추가적으로 다른 기간 및 다른 차수에 대해서도 시도한다. 적합되어진 시계열 모형 추정의 오차를 최소화하도록 최적의 이동평균의 시간간격을 선택하여 비교한다. Maleki 등 (2020)의 논문에서는 전체 모수들을 MLE로 추정하였으나, 본 연구에서는 모형의 오차를 최소화하여 예측의 정확성을 높이는 것을 연구 목적으로 하기에, 통상적 최소제곱추정법(ordi-

nary least squares estimation; OLSE)으로 시계열 모형의 계수를 먼저 추정한다. 잔차를 구하여 two-piece (TP) 분포에 필요한 모수를 추정한 후, t-분포의 모수는 최대우도추정법(maximum likelihood estimation; MLE)을 이용하는 단계적 추정 방법을 제안한다. Maleki 등 (2020)은 AR-TP-T 모형에서 단지 MLE 방법만을 사용하여 분석하였는데, 이것과 본 연구의 추정 방법을 추정 결과를 제시함으로 함께 비교한다. 제안된 단계별 추정 방법은 두 가지 측면에서 기존의 것보다 더 우수함을 입증할 수 있다. 두 가지 오차, 즉, HAR 모형의 평균제곱오차(mean squares error; MSE)와, 잔차와 이의 밀도함수의 차이를 제곱하여 평균을 구한 오차 (mean squares difference between relative frequency and density; MSD), 이 두 측면에서, 본 연구의 제안 방법이 더 적합함을 보인다. 또한 더 나아가 제안된 추정 방법을 활용한 한국 코로나 확진자 수에 대한 예측(out-of-sample forecasting)을 수행하여, 예측정확도의 한 측도로서 mean absolute percentage error (MAPE)를 계산하였고, 매우 작은 0.0953%의 오차를 가지는 예측정확도의 결과를 얻었다. 본 연구에서 제안하고 있는 시계열 모형과 추정 방법이 보다 정확한 예측 성능을 제공하여 한국내 COVID-19 확산 방지 및 정부의 대응 정책 수립에 기여하리라 생각한다.

이후 논문의 구성은 다음과 같다. 2절에서는 HAR 모형에 대해 간략히 소개하며, 3절에서는 일반적인 two-piece (TP) 분포에 대해 설명한다. 4절에서는 본 연구의 주제인 HAR-TP-T 모형에서의 추정 방법을 제안하며, 5절에서는 간단한 모의실험 결과를 제시한다. 6절에서는 한국 COVID-19 누적 확진자 수 실증데이터 분석을 수행하며, 7절에서는 요약 및 결론을 기술한다.

2. 이질적 자기회귀모형(Heterogeneous autoregressive model)

금융 시장을 설명하는 이론 중 하나는 이질적인 시장 가설이 있다. 이 이론은 투자 정보를 받아들이는 투자자들의 반응이 각각 다르다는 것을 전제로 한다. 주식 시장의 이질적 시장 가설은 Müller 등 (1993)에 의해 제안된 가설이며, 딜러 및 주식 시장의 트레이더는 주식 시장에서 거래를 하기 위해 단시간에 거래되는 주식 시장의 상황을 주시한다. 즉 시간적 의미에서 거의 분 단위로 주식 시장의 상황과 주식 거래의 빈도수를 파악하여 거래를 시도한다. 반면에 중 · 장기적으로 주식을 자산 형태로 보유하려고 하는 투자자는 정부의 정책 변경 등에 민감하고 시중 은행 같은 경우에는 주식 거래의 빈도수에 그다지 민감하게 반응하지 않는다. 따라서 시장 참가자는 주가 변동성과 함께 중 · 장기적으로 발생하는 사건에 주목한다. 이와 같이 시장 참가자는 투자 의사 결정을 할 때 시간적 이질성이 있다는 것을 전제로 경제 전체에 의해 구성되는 시장을 이질적 시장이라 한다.

Müller 등 (1993)은 시장의 동태적 변화를 파악하기 위해 이질적 시장 가설을 이용하여 주식 시장의 변동성을 분석해야 한다고 주장한다. 이질적 시장 가설에 관한 선행 연구는 Müller 등 (1993)의 HARCH (heterogeneous autoregressive conditional heteroscedastic) 모형, Corsi (2009)의 HAR-RV (heterogeneous autoregressive realized volatility) 모형 등이 있다. Müller 등의 이질적 시장 가설과 HARCH 모형에서 영감을 받아 Corsi (2009)는 HAR-RV 모형을 제안하였고, 이 모형은 일차, 주차, 월차 이동평균을 회귀변수로 이용한 선형 자기회귀모형으로 이론적으로는 차수 22의 자기회귀모형과 같다. 그러나 장기기억(long memory) 등, 변동성에 나타나는 특성을 잘 반영하고 있으며, 실현변동성에 대한 좋은 예측 성능으로 높이 평가 받고 있다. 이에 대한 응용으로서 비대칭성(asymmetry), 점프모형(jump diffusion model), 무한차수의 HAR 모형(infinite-order HAR) 등이 최근 활발하게 연구되어 오고 있다. Corsi (2009)의 HAR-RV 모형은 다음과 같이 나타낼 수 있다.

$$\mathrm{RV}_t = \alpha_0 + \alpha_d \mathrm{RV}_{t-1} + \alpha_w \mathrm{RV}_{t-5:t-1} + \alpha_m \mathrm{RV}_{t-22:t-1} + \epsilon_t,$$

여기서 $\mathrm{RV}_{t-5:t-1}$와 $\mathrm{RV}_{t-22:t-1}$는 과거의 주간, 월간 평균 실현변동성이며, 모수 $\alpha_d, \alpha_w, \alpha_m$은 각각 일일, 일주일, 일개월에 대한 계수를 뜻한다.

본 연구에서는 한국내에서의 발생하고 있는 COVID-19 확진자 수를 분석하기 위하여, 기존의 Maleki 등 (2020)의 선형 AR 모형 대신에 감염 바이러스의 잠복기를 고려하여 잠복기간 동안의 평균을 회귀변수로 고려한 HAR 모형을 채택하고자 한다. 더 나아가, 제안하고 있는 HAR 모형의 오차를 최소화하는 이동평균 시간간격 또한 추정하여, 잠복기간을 이동평균 기간으로 선택한 경우와 비교 분석하고자 한다. 기존의 금융 실현변동성 모형인 HAR 모형을 선택하되, 최근 폭발적으로 발생하는 COVID-19 감염 사례를 실증적으로 분석하기 위하여 비대칭성 및 두꺼운 꼬리 확률분포까지도 포함할 수 있는 오차과정을 선택하여 분석한다. 이러한 이유로 다음 절에서는 일반적으로 사용하는 정규분포가 아닌 비대칭성 two-piece scale mixture of normal (TP-SMN) 분포를 서술한다.

3. Two-piece scale mixture of normal (TP-SMN) 분포

비가우시안 백색잡음을 가진 비대칭 특성이 금융 자산수익, 통신, 사회학, 생물학 등 많은 영역에서 발견되고 있다. 분산의 비대칭성과 이상치 문제를 연구하기 위해 많은 연구자들은 비정규 시계열 모형을 적용하였다. 예를 들면, Bondon (2009)의 epsilon-skew-normal (ESN)에 기반한 AR 모형(AR-ESN)이나 Ghasami 등 (2019)의 generalized hyperbolic distribution (GH)을 따르는 AR 모형을 통한 모수 ML 추정, skewed AR 모형에 대한 베이지안 연구 등이 있다.

Scale mixtures of normal (SMN) 계열의 분포는 대칭 자료에 대해서 이상값에 영향을 받지 않는 통계적 추론을 하는 데 유용하게 이용되는 가벼운/두꺼운 꼬리 형태를 가지는 대칭 분포 집단이다. 이 분포는 비대칭 변수의 비정형 데이터를 다루는 시계열 모형을 사용하는 것에 적합하지 않다. 따라서, 비대칭성 시계열 자료에 로버스트한 추론을 하기 위하여 비대칭 계열의 분포를 AR 모형과 같은 시계열 모형 연구에 접목할 필요가 있다.

SMN 계열의 skewed 모형은 Branco과 Dey (2001)가 비대칭 데이터에 대해서 로버스트 추론을 하기 위해 사용했다. 그들은 scale-mixture skewed normal 분포의 mixtures를 SMN 계열에 조정하고 확장했다. SMN 계열의 skewed 모형을 구성하는 또 하나의 방법으로는 two-piece (TP) 분포를 이용하는 방법이 있다. TP 분포는 분리된 영역(왼쪽과 오른쪽의 절반 분포)의 2개의 혼합 성분으로 나타낼 수 있다. Two-piece normal (TP-N) 분포는 모양은 같지만, 분산이 다른 두 정규분포의 반쪽을 결합하는 분포이며, 최근 금융 저널에도 다시 발견되고 있다. 이 방식으로 기존의 SMN 계열은 로버스트하며 비대칭성을 다루는 TP-SMN 계열 분포로 확장된다. Ghasami 등 (2020)은 대칭 SMN 계열과 함께 TP-N, two-piece t (TP-T), two-piece slash (TP-SL), two-piece contaminated normal (TP-CN) 분포의 two-piece scale mixture of normal (TP-SMN) 계열을 다루고 있다.

TP-SMN 계열 분포는 다음과 같은 확률밀도함수(probability density function; pdf) 식을 가진다.

$$g(y|\mu,\sigma,\gamma) = \frac{2}{\sigma[a(\gamma)+b(\gamma)]}\left[f\left(\frac{y-\mu}{\sigma b(\gamma)}\right)I_{(-\infty,\mu]}(y) + f\left(\frac{y-\mu}{\sigma a(\gamma)}\right)I_{(\mu,+\infty)}(y)\right],$$

여기서 $f(\cdot)$은 실수 공간에서의 0을 중심으로 하는 표준 대칭 밀도함수이며, 실수 μ는 위치 모수, $\sigma > 0$은 척도 모수, $a(\gamma)$, $b(\gamma) > 0$은 왜도 모수이며, $I_A(\cdot)$는 A에서 1, A^c에서 0을 나타내는 지시함수이다. 표준화된 SMN 밀도 함수의 경우, $a(\gamma) = 1-\gamma$, $b(\gamma) = \gamma$를 고려한다.

이러한 비대칭 TP-SMN 계열 분포의 기초가 되는 잘 알려진 SMN은 Andrews과 Mallows (1974)가 소개한 것으로, 다음과 같은 pdf 식을 가진다. X가 SMN(μ,σ,ν) 분포일 때, 이것의 pdf는

$$f_{\text{SMN}}(x|\mu,\sigma,\gamma) = \int_0^\infty \phi\left(x|\mu,u^{-1}\sigma^2\right)dH(u|\nu), \quad \in R$$

이고, 이 때, 이 확률변수는 $X = \mu+\sigma U^{-1/2}W$로 표현 될 수 있다. 여기서, $\phi(\cdot\,|\mu,\sigma^2)$은 $N(\mu,\sigma^2)$ 분포의 밀도함수

를 나타내고, $H(\cdot|\nu)$는 모수 ν의 스칼라나 벡터로 지수화할 수 있는 척도혼합 확률변수 U의 누적분포함수이며, W는 U와 독립인 표준정규 확률변수이다.

TP-SMN 계열은 앞서 언급한 바와 같이, 비대칭 가벼운 꼬리 TP-N (또는 epsilon skew-normal, ESN), 비대칭 두꺼운 꼬리 TP-T, TP-SL, TP-CN 등이 있다. 각각에 대한 $f_{\text{SMN}}(\cdot)$, U의 분포 등은 Ghasami 등 (2020) 논문의 표 1을 참고한다. Ghasami 등 (2020)는 이러한 TP-SMN 분포의 오차과정을 이용한 AR 시계열 모형의 실제 사례를 분석하였고, Maleki 등 (2020)는 COVID-19 확진자 및 완치자 수에 대한 AR-TP-SMN 시계열 예측을 논하였다.

본 연구에서는 특히 Ghasami 등 (2020)의 AR-TP-SMN 모형 연구 비교에서 수행한 오차과정 분포들 중에서 가장 좋은 적합성을 보여준 TP-T 오차과정을 택하여 한국 코로나 확진자 수 데이터를 분석하고자 한다. TP-T 분포의 pdf는 다음과 같다. 실수 y에 대해서,

$$g(y|\mu,\sigma_1,\sigma_2,\nu) = 2\frac{\sigma_1}{\sigma_1+\sigma_2} f_T(y|\mu,\sigma_1,\nu) I_{(-\infty,\mu]}(y) + 2\frac{\sigma_2}{\sigma_1+\sigma_2} f_T(y|\mu,\sigma_2,\nu) I_{(\mu,\infty)}(y), \tag{3.1}$$

$$f_T(y|\mu,\sigma_i,\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}\sigma_i}\left(1+\frac{1}{\nu}\left(\frac{y-\mu}{\sigma_i}\right)^2\right)^{-\frac{\nu+1}{2}}, \quad i=1,2.$$

위의 식에서 척도 모수 σ와 왜도 모수 γ는 $\sigma = \sigma_1+\sigma_2$와 $\gamma = \sigma_2/(\sigma_1+\sigma_2)$로 표현된다.

다음 장에서는 본 연구에서 제안하고자 하는 TP-T 오차과정을 가지는 HAR 모형을 소개하고, 이 모형을 한국 COVID-19 확진자 수에 적용하여 모수를 추정하는 방법을 제안한다. 오차를 줄이기 위하여 기존의 OLSE와 MLE, 두 추정 방법을 보완 개선함으로서 실제 데이타에 더 적합하도록 시계열 모형 추정 방법을 제안한다.

4. Heterogeneous autoregressive model with two-piece t distributed errors (HAR-TP-T) 모형

본 연구의 주요 내용으로서, 위의 두 절에서 언급한 HAR 모형과 TP-T 분포의 오차과정을 따르는 시계열 모형인 HAR-TP-T 모형을 소개하며 이에 대한 추정 방법을 제안하고자 한다. HAR-TP-T 모형을 기반으로 하여 한국내에서의 COVID-19 일일 누적 확진자 수에 대해 분석한다. HAR(p)-TP-T 모형은 다음과 같다.

$$X_t = \phi_1 X_{t-1}^{(1)} + \phi_2 X_{t-1}^{(2)} + \cdots + \phi_p X_{t-1}^{(p)} + Z_t, \tag{4.1}$$

$$X_{t-1}^{(i)} = \frac{1}{h_i}(X_{t-1} + X_{t-2} + \cdots + X_{t-h_i}), \quad i=1,2,\ldots,p.$$

단, $1 = h_1 < h_2 < \cdots < h_p$ 양의 정수이며, 오차과정 $\{Z_t\}$는 비대칭성과 두꺼운 꼬리 확률분포를 고려하기 위하여 채택한, 식 (3.1)의 TP-T 분포를 따르는 확률변수들이다.

본 연구에서 제안하는 방법은 다음과 같다.

첫째, ordinary least squares estimation (OLSE) 방법으로 HAR 모형의 계수를 추정한다. 모형 (4.1)의 행렬형식은 $X_t = \phi^\top \mathbf{Y}_{t-1} + Z_t$로 표현되며, $\phi = (\phi_1,\ldots,\phi_p)^\top$, $\mathbf{Y}_{t-1} = (X_{t-1}^{(1)},\ldots,X_{t-1}^{(p)})^\top$, 계수 벡터 ϕ에 대한 최소제곱추정량(OLSE)은 다음과 같이 $\hat{\phi}$으로 표현된다.

$$\hat{\phi} = \left(\sum_{t=1}^{n} \mathbf{Y}_{t-1}\mathbf{Y}_{t-1}^\top\right)^{-1}\left(\sum_{t=1}^{n} \mathbf{Y}_{t-1}X_t\right).$$

6절의 그래프에서 보듯이 한국 COVID-19 확진자수 데이터에 대한 최소제곱추정량으로 계산된 잔차의 분포는 다소 비대칭성을 나타내고 있으며 또한 정규분포보다는 두꺼운 꼬리를 가지는 분포에 더 가깝다고 할 수 있다. 잔차들 $\{\hat{Z}_t\}$을 이용하여, 식 (3.1)의 TP-T 분포의 모수들을 차례대로 추정한다.

둘째, OLSE 잔차들(residuals) $\{\hat{Z}_t := X_t - \hat{\phi}^\top \mathbf{Y}_{t-1},\ t = 1, \ldots, n\}$를 구한 후, 잔차들의 표본평균 $\hat{\mu}$과

$$\hat{\mu} = \frac{1}{n}\sum_{t=1}^{n} \hat{Z}_t.$$

두 표본표준편차 $\hat{\sigma}_1, \hat{\sigma}_2$, 즉, 표본평균 $\hat{\mu}$ 보다 작은 쪽의 표본표준편차 $\hat{\sigma}_1$, 표본평균보다 큰 쪽의 표본표준편차 $\hat{\sigma}_2$를 각각 계산한다:

$$\hat{\sigma}_1^2 = \frac{1}{n}\sum_{t=1}^{n} \left(\hat{Z}_t - \hat{\mu}\right)^2 I_{(-\infty,\hat{\mu}]}\left(\hat{Z}_t\right), \qquad \hat{\sigma}_2^2 = \frac{1}{n}\sum_{t=1}^{n} \left(\hat{Z}_t - \hat{\mu}\right)^2 I_{(\hat{\mu},\infty)}\left(\hat{Z}_t\right).$$

셋째, 모수 ν에 대한 우도함수 $\mathcal{L}(\nu|\mathcal{X}) = \prod_{t=1}^{n} g(\hat{Z}_t|\hat{\mu}, \hat{\sigma}_1, \hat{\sigma}_2, \nu)$가 최대가 되는, 또는 이의 로그우도함수가 최대가 되는 최대우도추정값을 찾는다. 단, $\mathcal{X}$는 관측된 데이터에 대한 모든 정보이며, 함수 $g(\cdot\,|\hat{\mu}, \hat{\sigma}_1, \hat{\sigma}_2, \nu)$는 식 (3.1)에서 볼 수 있다. 알고리즘 및 실험을 통해 최대우도추정값 $\hat{\nu}$를 찾는다:

$$\hat{\nu} = \arg\max_{\nu} \mathcal{L}(\nu|\mathcal{X}) = \arg\max_{\nu} \sum_{t=1}^{n} \log g\left(\hat{Z}_t|\hat{\mu}, \hat{\sigma}_1, \hat{\sigma}_2, \nu\right).$$

단계별 추정 방법의 결과를 검증하기 위해 다음 절에서 간단한 모의실험을 시행한다. HAR-TP-T 모형에서 OLS 추정량의 일치성을 모의실험을 통해 입증하며, 또한 단계별로 이루어지는 TP-T 모수 추정도 매우 좋은 성능을 가짐을 확인한다. 6절의 COVID-19 실증적 데이터 분석에서는, 한국 COVID-19 확진자수 분석을 위해 여러가지 차수의 HAR-TP-T 모형을 채택하고, 각 모형에서의 모수 추정값을 계산하여 기존 방법의 결과와 비교한다. 즉, Maleki 등 (2020)의 논문에서처럼 모든 모수들을 MLE 방법으로 추정하는 경우의 추정결과를, 이 연구에서 소개한 방법과 비교 분석한다. Maleki 등 (2020)의 연구는 선형 AR 모형에 TP-T 오차과정을 고려하였고, 모형의 계수까지 포함하는 모수 전체를 MLE 방법으로 추정하였다. 그러나 본 연구에서는 HAR 모형을 다루고 있으며, OLSE 추정의 장점인 오차를 최소화하는 방법을 이용하여 첫단계로서의 HAR 모형의 계수를 추정한 후, 두 번째 단계로서의 잔차들의 평균과 표준편차 추정 후, 마지막 단계로 TP-T 분포의 주요 모수인 ν만을 MLE로 추정한다는 점에서 차별성이 있다. 또한 제안하고 있는 단계별 추정 방법이 위에 언급하고 있는 MLE 방법에서의 성능과 비교하여 더 우수함을 입증한다.

5. 모의 실험

본 연구에서 제안하고 있는 HAR-TP-T 시계열 모형에서의 단계별 추정 방법 검증을 위해 간단한 모의 실험을 수행하여 그 결과를 확인하고자 한다. HAR(3) 모형에서 계수 $(\phi_1, \phi_2, \phi_3) = (0.3, -0.2, 0.1)$와, TP-T 분포의 모수 $(\mu, \sigma_1, \sigma_2, \nu) = (0, 0.2, 0.3, 3.5)$를 선택하고 표본크기 $n = 200, 400, 600$에 대한 모의실험을 500회 반복하여 수행한다. 모수 추정값 각각의 표본평균과 표준오차 및 추정 모형에 관한 오차인 MSE를 계산한다.

모의 실험에서의 모수 추정값 결과는 표본크기 n이 커질수록 추정값의 평균이 실제 모수값에 더 가까운 값으로 계산되었고, 표준오차 및 MSE 또한 n이 커질수록 작아지는 것을 확인할 수 있다. 실험 결과는 Table 1에서 볼 수 있으며, HAR 모형 추정계수를 이용한 잔차의 TP-T 분포와 확률밀도함수 추정 그래프는 Figure 1에서 확인할 수 있다. 비대칭을 의미하는 σ_1, σ_2가 각각 0.2,0.3이므로 잔차들이 양의 값, 즉 0을 중심으로 오른쪽으로 다소 많은 분포를 띠고 있음을 볼 수 있다. 다음 절에서는, 제안하고 있는 모형과 추정 방법을 한국 COVID-19 데이터에 적용한 실제 사례 분석을 제시한다.

Table 1: Simulation results of estimates (sample mean and standard error (se)) in HAR(3) models with $(h_1, h_2, h_3) = (1, 7, 14)$; replication number = 500

	HAR(3)-TP-T model with $(1, 7, 14)$		
	$n = 200$	$n = 400$	$n = 600$
$\phi_1 = 0.3$	0.297 (0.074)	0.303 (0.051)	0.301 (0.042)
$\phi_2 = -0.2$	−0.219 (0.212)	−0.212 (0.146)	−0.202 (0.123)
$\phi_3 = 0.1$	0.567 (0.254)	0.565 (0.184)	0.552 (0.149)
$\mu = 0$	0.052 (0.006)	0.052 (0.005)	0.051 (0.004)
$\sigma_1 = 0.2$	0.232 (0.025)	0.232 (0.019)	0.232 (0.017)
$\sigma_2 = 0.3$	0.309 (0.031)	0.307 (0.024)	0.306 (0.021)
$\nu = 3.5$	3.461 (0.194)	3.485 (0.162)	3.507 (0.138)
MSE	0.139	0.136	0.125

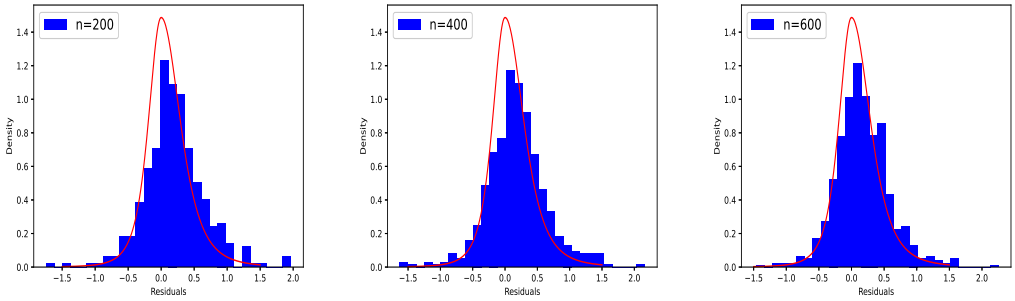

Figure 1: *Simulation results of distributions for residuals by OLSE-TP-T fitting in HAR(3)-TP-T model with* $(h_1, h_2, h_3) = (1, 7, 14)$*; sample size* $n = 200, 400, 600$.

6. 한국 COVID-19 누적 확진자 수 실증 분석

본 절에서는, 앞에서 제안한 HAR-TP-T 모형을 한국 COVID-19 확진자 수에 적용하여 모수들을 추정하고자 한다. 2020년 1월 21일부터 2020년 10월 23일까지의 누적 확진자 수 데이터를 Figure 2에 제시하였다. Figure 2(왼쪽)에서 보듯이 누적 확진자 수는 두 차례의 폭발적 확산을 잘 나타내며 증가하고 있다. 이 중에서 2020년 1월 21일–2020년 9월 17일(in-sample period)의 누적 확진자 수를 모형의 추정을 위해 이용하였고, 2020년 9월 18일–2020년 10월 23일 데이터를 our-of-sample 예측을 위해 사용하였다. 일일 누적 확진자 수 데이터를 정상화(stationary)하기 위한 자료변환을 하였고, 누적합 검정(CUSUM test)를 이용하여 분포의 변화시점을 탐지(change point detection)하였다. 즉, 일일 누적 확진자 수 데이터를 로그차분한 후 누적합 검정 통계량(CUSUM test statistic) 값을 계산하여 평균이 달라지는 시점을 탐지하였고, 변화시점 탐지 결과 46번째 시점에서 평균변화가 가장 크게 탐지되었다. 한번 로그차분한 데이터의 변화시점 이후의 그래프는 Figure 2의 두 번째와 같으며, 다시 한번 로그차분한 그래프는 Figure 2의 세 번째와 같다. 실제 분석을 위해서 변화시점 이후의 두번 로그차분한 데이터, 즉, Figure 2 세 번째 (오른쪽) 2020.03.07 이후의 정상(stationary) 데이터를 예측 모형에 사용한다.

본 연구에서는 차수 $p = 2, 3, 4$의 HAR 모형을 적용하였다. 먼저, $p = 2$일 때, 바이러스 잠복기 2주를 고려한 $h_2 = 14$를 채택하였고, $p = 3$일 때, 일주일 기간을 추가하여 $h_2 = 7$, $h_3 = 14$를 선택하였고, $p = 4$일 때는 3주 기간을 뜻하는 $h_4 = 21$을 추가하였다. 이 세 모형의 MSE를 계산하고, 각각 같은 차수에 대해, 오차가 더 작은 HAR 모형을 찾고자 $(h_1, h_2, \ldots, h_p)$를 모수로 간주하여 (단, $h_1 = 1$), MSE가 최소가 되게 하는 HAR(p)

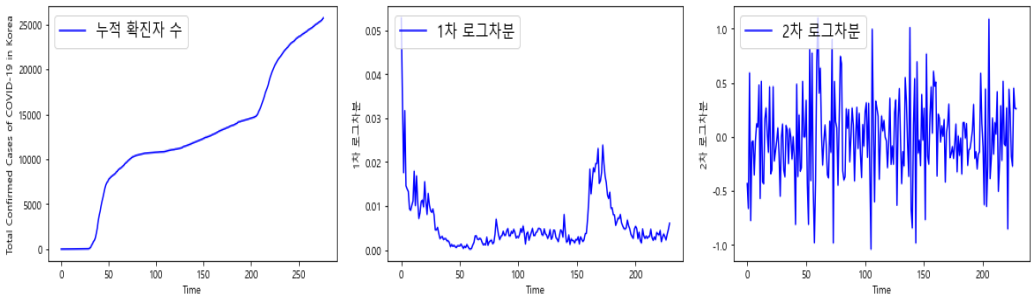

Figure 2: *Total confirmed cases of COVID-19 in Korea from 21-Jan to 23-Oct of 2020 (left), First difference of Log after change point from 07-Mar to 23-Oct of 2020 (center), Second difference of Log after change point from 08-Mar to 23-Oct of 2020 (right).*

Table 2: MSEs compared for six different HAR(p) fittied models for Korean COVID-19

p	$(h_1, h_2, \ldots, h_p)$	MSE	$(h_1, h_2, \ldots, h_p)$	MSE
2	(1, 14)	0.1469740	(1, 5)*	0.1412910
3	(1, 7, 14)	0.1450136	(1, 2, 5)*	0.1396784
4	(1, 7, 14, 21)	0.1472110	(1, 2, 5, 8)*	0.1393637**

Mark * indicates lags with the minimized MSE of given order p and mark ** for the minimum MSE among all cases.

에서의 $(h_1, h_2, \ldots, h_p)$를 추정하였다. 총 여섯 가지 모형의 MSE는 Table 2에서 볼 수 있다.

위의 여섯 가지 HAR(p) 모형에서의 계수 추정 및 잔차에 대한 모수 추정에 대해 기존의 MLE-TP-T 추정과 본 연구에서 제안하고 있는 단계별 OLS-TP-T 추정 방법에 대해 비교 분석한다. 첫째로 HAR 모형의 계수들을 OLSE로 추정한 후 잔차를 TP-T 분포로 가정하여 앞의 4절에서의 추정 방법을 이용하여 계산한 추정값들을 비교한다. 또한, Maleki 등 (2020)의 논문에서처럼 계수 포함한 모든 모수들을 MLE로 추정하는 방법과 비교 분석한다.

추정법을 비교하기 위하여 다음과 같은 기준을 이용한다. 특히 본 연구의 초점은 오차과정을 최소화하기 위함이 목적이므로 모형의 오차를 두 측면에서 계산한다. 이를 위하여 본 연구에서의 오차명을 MSE과 MSD로 구별한다: (1) 모형에 대한 오차 제곱의 평균, MSE; (2) 잔차의 분포를 추정한 후 이에 대한 오차 측도로서의 MSD, 즉 잔차 상대도수와 추정된 확률밀도함수의 차이를 제곱하여 평균 계산. 이 두 용어에 대한 정확한 수식 표현은 아래 기술한다.

MSE = 오차 제곱들의 평균, i.e., mean squares differences between X_t and $\hat{\phi}^\top \mathbf{Y}_t$,

$$\text{MSE} = \frac{1}{n}\sum_{t=1}^{n} \hat{Z}_t^2 = \frac{1}{n}\sum_{t=1}^{n} \left(X_t - \hat{\phi}^\top \mathbf{Y}_t\right)^2$$

MSD = 잔차들의 상대도수와 추정확률밀도함수의 차이에 대한 제곱들의 평균(mean squares differences between relative frequency of residuals $\hat{Z}_t$ and the estimated density),

$$\text{MSD} = \frac{1}{n}\sum_{t=1}^{n} \left(\hat{f}\left(\hat{Z}_t\right) - f_{\hat{\theta}}\left(\hat{Z}_t\right)\right)^2,$$

단, $\hat{f}(z) = (1/m)\sum_{i=1}^{m} I_{(x_i - dx, x_i]}(z)$, $\min\{\hat{Z}_t\} = x_0 < x_1 < x_2 < \cdots < x_m = \max\{\hat{Z}_t\}$, $dx = (1/m)(\max\{\hat{Z}_t\} - \min\{\hat{Z}_t\})$, 그리고 $f_{\hat{\theta}}(z)$는 추정된 모수 $\hat{\theta}$를 가지는 TP-T 확률밀도함수이다.

Table 3: Estimation results of HAR(2) models with $(h_1, h_2) = (1, 14), (1, 5)$ for Korean COVID-19

	HAR(2) model with $(1, 14)$		HAR(2) model with $(1, 5)$	
	OLSE	MLE	OLSE	MLE
ϕ_1	−0.164	0.279	−0.117	−0.299
ϕ_2	−0.126	−0.009	−0.508	−0.299
μ	−0.005	−0.005	−0.009	−0.008
σ_1	0.246	0.193	0.236	0.195
σ_2	0.250	0.192	0.246	0.186
ν	3.051	1.925	3.035	2.040
LL	−84.46	−92.66	−86.59	−95.71
AIC	170.93	187.33	175.20	193.43
BIC	174.11	190.51	178.43	196.67
MSE	0.147	0.149	0.141*	0.145
MSD	0.0143*	0.0218	0.0156	0.0289

Mark * indicates the smallest error.

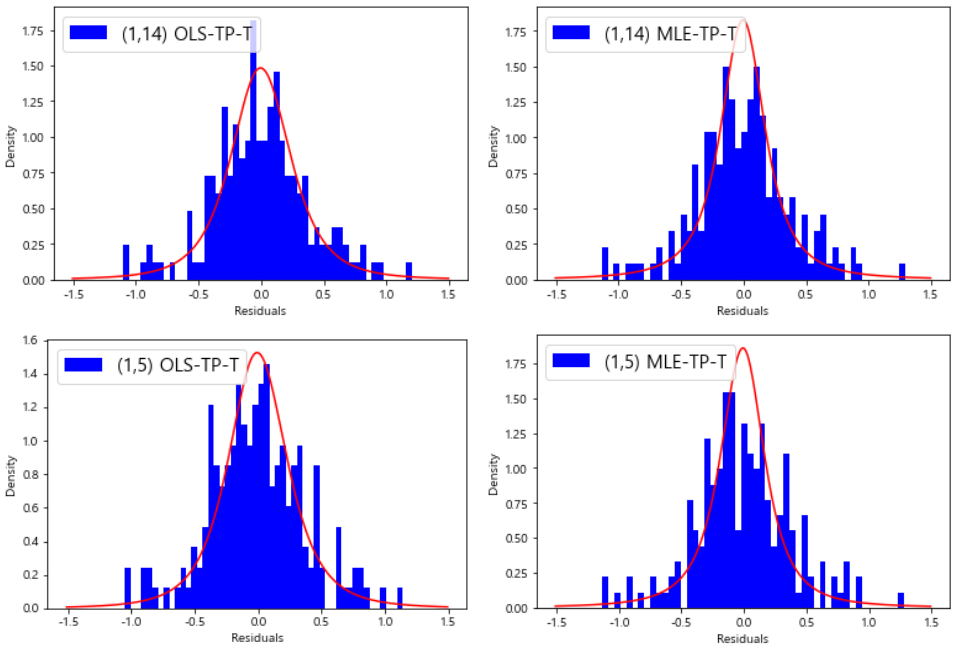

Figure 3: *Distributions for residuals by OLSE-TP-T fitting and MLE-TP-T fitting, in HAR(2) model with* $(h_1, h_2) = (1, 14)$ *and* $(1, 5)$.

이를 포함하여, LL, AIC, BIC의 다섯가지 기준으로 추정법을 비교한다: LL = 로그우도함수(LogLikelihood function), AIC = 아카이케 정보 기준(Akaike's information criterion), BIC = 베이지안 정보 기준(Bayesian information criterion)

$$\text{AIC} = -2\text{LL} + 2k, \qquad \text{BIC} = -2\text{LL} + k\log(n).$$

단, k는 모수들의 개수이다.

Table 4: Estimation results of HAR(3) models with $(h_1, h_2, h_3) = (1, 7, 14), (1, 2, 5)$ for Korean COVID-19

	HAR(3) model with (1, 7, 14)		HAR(2) model with (1, 2, 5)	
	OLSE	MLE	OLSE	MLE
ϕ_1	−0.131	−0.349	−0.048	−0.100
ϕ_2	−0.589	−0.110	−0.227	−0.325
ϕ_3	0.392	0.109	−0.351	−0.100
μ	−0.004	−0.005	−0.009	−0.008
σ_1	0.239	0.195	0.228	0.201
σ_2	0.248	0.194	0.253	0.175
ν	3.025	1.931	3.030	2.001
LL	−84.25	−92.88	−85.84	−94.77
AIC	170.50	187.78	173.70	191.55
BIC	173.69	190.96	176.93	194.78
MSE	0.145	0.152	0.140*	0.141
MSD	0.0214	0.0253	0.0184*	0.0265

Mark * indicates the smallest error.

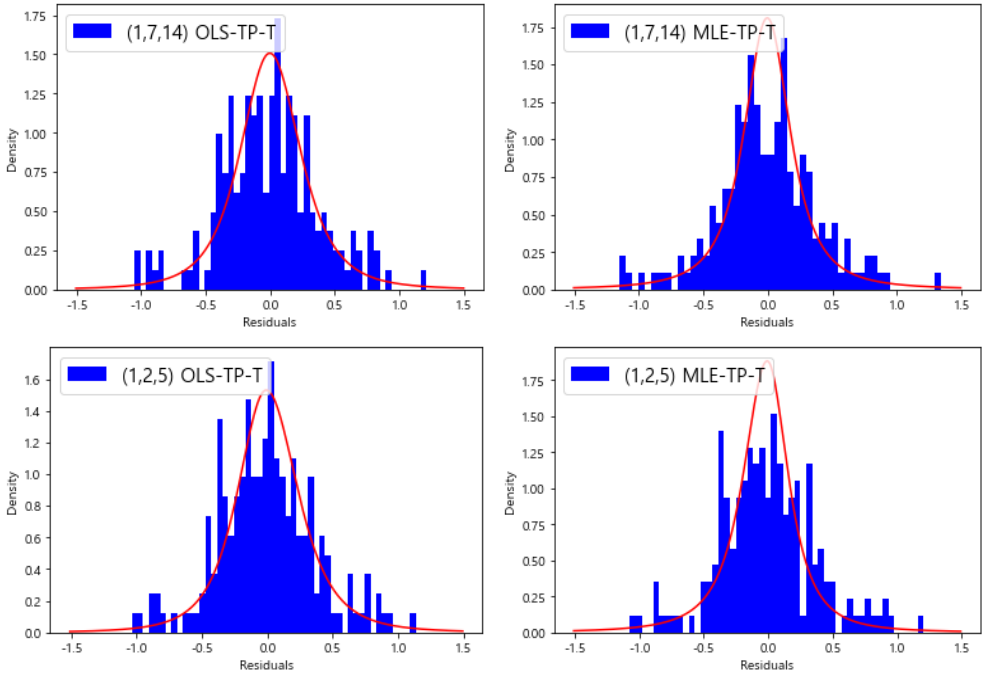

Figure 4: *Distributions for residuals by OLSE-TP-T fitting and MLE-TP-T fitting in HAR(3) model with* $(h_1, h_2, h_3) = (1, 7, 14)$ *and* $(1, 2, 5)$.

Table 3–5에서의 마지막 다섯 행은 이 다섯 가지 기준에 대한 계산 결과이다. Table 2에서 보듯이 제시하고 있는 여섯 모형들 중에서 HAR(4)-TP-T with $(1, 2, 5, 8)$ 모형에서의 MSE가 가장 작다. 그리고 이 모형에서의 두 추정 방법 중 OLSE-TP-T 방법이 MSD가 더 작은 값을 나타내고 있다. 그러므로 이 모형을 선택하여 본 연구에서의 결론으로서 예측 성능을 확인하고자 한다.

Table 5: Estimation results of HAR(4) models with $(h_1, h_2, h_3, h_4) = (1, 7, 14, 21), (1, 2, 5, 8)$ for Korean COVID-19

	HAR(4) with $(1, 7, 14, 21)$		HAR(4) with $(1, 2, 5, 8)$	
	OLSE	MLE	OLSE	MLE
ϕ_1	−0.131	−0.250	−0.041	−0.100
ϕ_2	−0.560	−0.278	−0.252	−0.375
ϕ_3	0.577	0.888	−0.643	−0.300
ϕ_4	−0.326	−0.889	0.495	0.250
μ	−0.005	−0.005	−0.003	−0.003
σ_1	0.247	0.189	0.227	0.188
σ_2	0.250	0.198	0.254	0.188
ν	3.063	1.888	3.038	2.000
LL	−81.09	−88.54	−84.15	−92.36
AIC	164.19	179.08	170.31	186.72
BIC	167.34	182.27	173.53	189.94
MSE	0.147	0.150	0.139*	0.141
MSD	0.0221	0.0215	0.0207*	0.0299

Mark * indicates the smallest error.

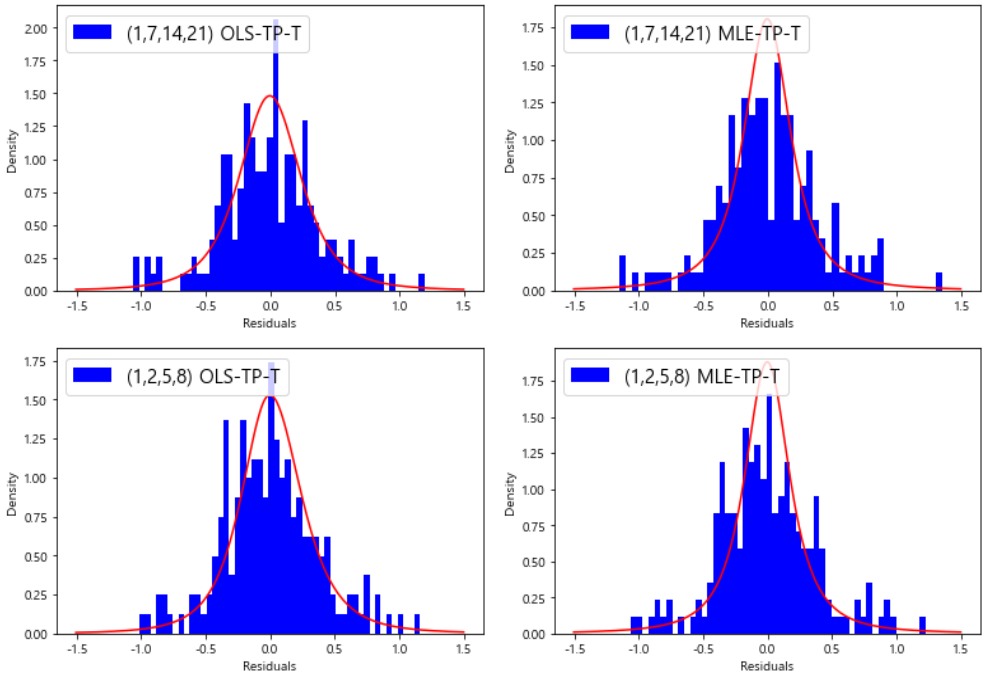

Figure 5: *Distributions for residuals by OLSE-TP-T fitting and MLE-TP-T fitting in HAR(4) model with* $(h_1, h_2, h_3, h_4) = (1, 7, 14, 21)$ *and* $(1, 2, 5, 8)$.

한국 COVID-19 일일 확진자 수 예측을 목적으로, 앞서 기술한 HAR(4)-TP-T with (1,2,5,8) 모형에서의 OLSE-TP-T 결과를 활용한다. 2020년 9월 18일부터 2020년 10월 23일까지의 실제 데이터와 비교하는 예측(out-of-sample forecasting)을 수행하여 예측 성능 결과를 살펴보고자 한다. HAR(4)-TP-T with $(1, 2, 5, 8)$ 모

Table 6: Forecast and 95% Prediction Interval (P.I) using HAR(4)-TP-T model with $(1, 2, 5, 8)$, by the OLSE

Date	Real value	Forecast	Lower of P.I.	Upper of P.I.
2020.09.18	22783	22800	22732	22959
2020.09.19	22893	22906	22843	23005
2020.09.20	22975	23005	22947	23106
2020.09.21	23045	23063	23023	23165
2020.09.22	23106	23121	23082	23188
2020.09.23	23216	23175	23149	23240
2020.09.24	23341	23321	23273	23419
⋮	⋮	⋮	⋮	⋮
2020.10.19	25275	25283	25251	25394
2020.10.20	25333	25354	25310	25466
2020.10.21	25424	25400	25365	25472
2020.10.22	25543	25502	25467	25591
2020.10.23	25698	25644	25592	25738

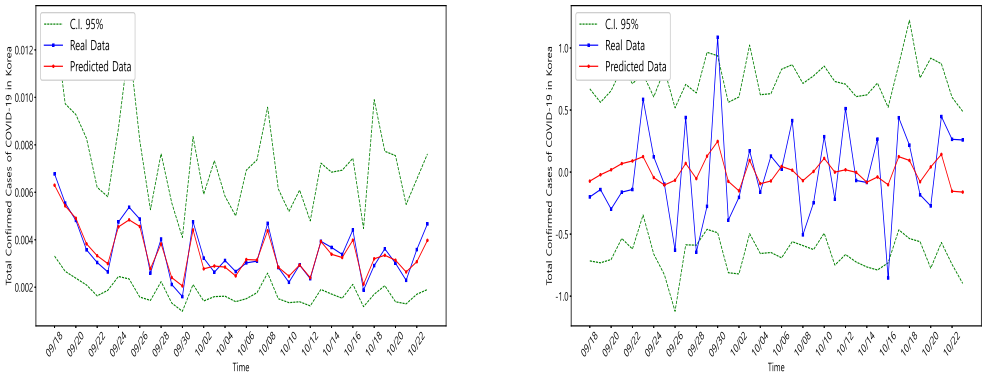

Figure 6: *Forecasting the first difference of log from 18-Sep to 23-Oct of 2020 in Korea (left), Forecasting the second difference of log from 19-Sep to 23-Oct of 2020 in Korea (right).*

형으로 적합시켜서 계산된 예측값과 95% 예측구간은 Figure 6과 Figure 7에서 볼 수있다. Figure 6의 왼쪽은 코로나 확진자 수 데이터의 한번 로그차분한 데이터의 실제값과 예측값, 예측구간을 나타내며, 오른쪽 그림은 두번 로그차분한 데이터의 실제값과 예측값, 예측구간을 의미한다. 95% 예측구간은, 두번 로그차분한 데이터에 대한 HAR(4)-TP-T 모형의 OLSE 잔차의 분포를 실험을 반복하여 생성하여 계산하였다. Figure 7은 실제 한국 COVID-19 일일 누적 확진자 수와 예측값 및 예측구간을 보여주고 있다. 예측값과 예측구간 모두 실제값과 크게 차이가 없음을 확인할 수 있다. 특히 주목할 것은, 예측구간의 상한값이 더 크게 나타난 이유는 TP-T 오차과정 추정에서 $\hat{\sigma}_1, \hat{\sigma}_2$의 값이 각각 0.227, 0.254로서 $\hat{\sigma}_2 > \hat{\sigma}_1$이기 때문이다. 1-step 예측의 정확도를 파악하기 위해 손실함수는 다음과 같이 mean absolute percentege error (MAPE)를 사용한다.

$$\text{MAPE} = \frac{100}{m}\left(\sum_{n=N-m}^{N-1}\frac{|\hat{C}_{n+1|n} - C_{n+1}|}{C_{n+1}}\right),$$

여기서 C_{n+1}은 $n+1$시점의 일일 누적 확진자 수이고, $\hat{C}_{n+1|n}$은 시점 n까지의 자료에 근거한 C_{n+1}의 예측값이다. $(C_{n+1}, \hat{C}_{n+1|n})$의 실제값과 예측값은 Table 6에서 그 일부를 볼 수 있다. 2020년 9월 18일부터 2020년 10월 23

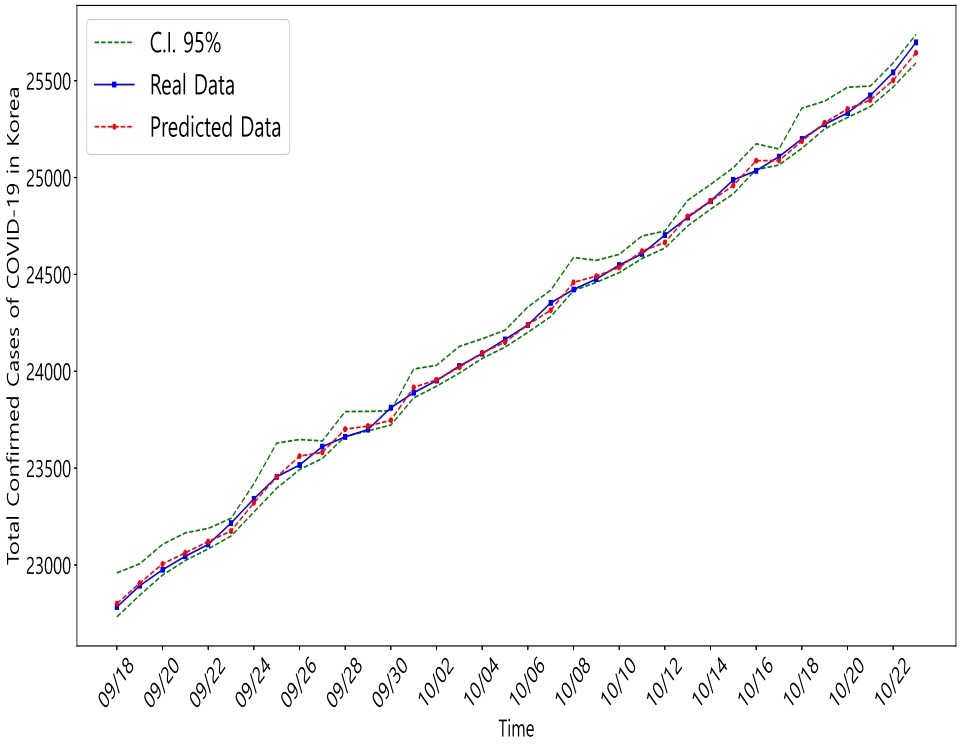

Figure 7: *Forecasting of total confirmed cases of COVID-19 with HAR(4)-TP-T model in Korea from 18-Sep to 23-Oct of 2020.*

일까지의 $m = 36$일간의 MAPE는 0.0953%로 계산되었다. 이는 본 연구에서의 추정 방법을 사용한 예측은 오차가 매우 작음을 나타낸다고 할 수 있다.

7. 결론

전 세계적으로 확산되고 있는 COVID-19 감염병에 대한 학계에서의 다각도의 연구는 매우 시급하며, 통계학 분야에서의 그 역할도 중요하다고 할 수 있다. 시간이 지남에 따라, 진정되기 보다는 산발적으로 발생하는 바이러스 확산은 보다 면밀한 학계에서의 노력이 필요하다. 특히 시계열 자료로 축적되는 COVID-19 확진자, 사망자 및 완치자 수의 데이터는 시계열 분석 연구에 매우 중요한 자료라고 할 수 있다. 본 연구 논문에서는 한국 코로나 확진자 수에 대한 시계열 분석 및 예측의 필요성을 실감하며 예측의 정확성에 초점을 맞추어 시계열 모형과 추정 방법을 제안한다. 또한 이를 기반으로 확진자 수에 대한 예측을 수행하여, 제안하고 있는 모형의 예측정확도를 계산하여 우수한 예측 성능을 제시한다.

본 연구에서는 2020년 1월 21일 이후부터 2020년 10월 23일까지 한국 COVID-19 일일 누적 확진자 수 데이터에 대한 시계열 분석을 수행하였다. 데이터 정상화 방법으로 누적합 검정법 및 로그차분 자료변환을 하여 정상시계열 데이터인 2020년 3월7일부터의 데이터를 추정 및 예측 모형으로 사용하였고, 이 중 2020년 9월 18일 이후의 데이터는 out-of-sample 예측을 위해 사용하였다. 통계방법론적 시계열 분석을 위해, 2개의

혼합된 t-분포(TP-T)의 오차과정을 따르는 이질적 자기회귀 (HAR) 모형을 선택한다. HAR-TP-T 모형에서의 차수 $p = 2, 3, 4$에 대해, 먼저 바이러스 잠복기간으로 잘 알려져 있는 14일을 기준으로 HAR 모형을 설정한다. 그러나, 적합되어진 시계열 모형 추정의 오차를 최소화하도록 이동평균의 기간을 선택하여 비교한다. 기존의 알려진 AR-TP-T 모형에서의 추정 방법인 MLE만을 사용하여 추정하기 보다는, 본 연구에서는 HAR 계수 추정을 위해서는 먼저 오차를 줄이고자 OLSE 방법을 사용하고, 다음 단계로 잔차들을 이용한 TP의 모수들을 추정한 후, 마지막 단계로 t-분포의 모수를 추정하는 단계별 추정 방법을 제안한다. 이러한 추정은 두 가지 오차 측면에서 기존의 방법보다 우수하다고 할 수 있다. 즉, HAR 모형의 평균제곱오차와, 잔차분포에 대한 밀도함수 추정의 평균제곱오차, 두 측면에서 모두, 본 연구에서의 제안 방법이 기존 방법보다 더 작은 오차 결과를 나타내며 더 우수함이 입증되었다. 나아가, 최적화 추정 방법을 활용한 한국 COVID-19 확진자 수 예측(out-of-sample forecasting) 결과(예측값, 예측구간)을 계산하였고, 예측정확도의 한 측도로서 MAPE를 계산하여 0.0953%의 매우 작은 오차값을 얻었다. 본 연구에서 제안한 시계열 분석 및 추정 방법은 보다 정확한 한국 코로나 확진자 수 예측 성능을 제공한다고 할 수 있으며, 본 연구를 통한 시계열 모형과 추정 결과는 정확한 예측 이론으로 미래를 계획하고자 하는 한국 정부의 코로나 확산 방지 및 대응 방안 수립에 기여하리라 생각한다.

본 연구에서는 확진자 수 데이터 분석만을 다루었으나, 추후 연구에서는 다변량 시계열 모형을 이용한 확진자 수, 사망자 수 등의 이변량 또는 다변량의 데이터 분석 및 예측 연구를 수행하고자 한다. 기존의 HAR 모형의 일별, 주별, 월별의 시간간격의 제한을 넘어서, 모형의 오차를 줄이는 것을 목적으로 예측의 정확성에 초점을 맞추어, 이동평균 시간간격을 모수로 간주하여 추정하고, 오차과정에 대한 분포를 본 연구에서처럼 비대칭 및 두꺼운 꼬리 분포로 선택하여 다변량 시계열 분석을 해 볼 수 있다. 다시 말해서 서로 상관성을 갖고 있는 확진자 수, 사망자 수, 완치자 수 등, 또는 코로나 확산의 원인을 제공하는 다양한 변수들을 포함한 다변량 시계열 모형에 관한 연구는 향후 가치있고 흥미로운 연구 주제가 될 것이다.

References

Andersen, T. G. and Bollerslev, T. (1998). Answering the Skeptics: YES, Standard Volatility Models do Provide Accurate Forecasts, *International Economic Review*, **39**, 885–905.

Andrew, D. R. and Mallows, C. L. (1974). Scale mixture of normal distribution. *Journal of the Royal Statistical Society: Series B*, **36**, 99–102.

Arellano-Valle, R. B., Gómez, H., and Quintana, F. A. (2005). Statistical inference for a general class of asymmetric distributions, *Journal of Statistical Planning and Inference*, **128**, 427–443.

Benvenuto, D., Giovanetti, M., Vassallo, L., Angeletti, S., and Cicozzi, A. (2020). Application of the ARMA model on the COVID-2019 epidemic dataset, *Data in Brief,* **29**, 105340.

Bondon, P. (2009). Estimation of autoregressive models with epsilon-skew-normal innovations, *Journal of Multivariate Analysis*, **100**, 1761–1776.

Branco, M. D. and Dey, D. K. (2001). A general class of multivariate skew-elliptical distributions, *Journal of Multivariate Analysis* , **79**, 99–113.

Ceylan, Z. (2020). Estimation of COVID-19 prevalence in Italy, Spain, and France, *Science of the Total Environment*, **729**, 138817.

Corsi, F. (2009). A simple approximate long-memory model of realized volatility, *Journal of Financial Econometrics*, **7**, 174–196.

Ghasami, S., Khodadadi, M., and Maleki, M. (2019). Autoregressive processes with generalized hyperbolic innovations, *Communications in Statistics - Simulation and Computation*, **49**, 3080–3092.

Ghasami, S., Maleki, M., and Khodadadi, Z. (2020). Leptokurtic and platykurtic class of robust symmetrical and asymmetrical time series models, *Journal of Computational and Applied Mathematics*, **376**, 1–12.

Kirbas, I., Sözen, A., Tuncer, A. D., and Kazancıoglu, F. (2020). Comparative analysis and forecasting of COVID-19 cases in various European countries with ARIMA, NARNN and LSTM approaches, *Chaos, Solitons & Fractals*, **138**, 110015.

Maleki, M., Mahmoudi, M., Wraith, D., and Pho, K. (2020). Time series modelling to forecast the confirmed and recovered cases of COVID-19. *Travel Medicine and Infectious Disease*, **37**, 101742.

Müller, U. A., Dacorogna, M. M., Dave, R. D., Pictet, O. V., Olsen, R. B., and Ward, J. R. (1993). Fractals and intrinsic time: a challenge to econometricians. *in Proceedings of the 39th International AEA Conference on Real Time Econometrics*, Luxembourg, October 1993.

Ribeiro, M. H. D. M., da Silva, R. G., Mariani, V. C., and Coelho, L. D. S. (2020). Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil, *Chaos, Solitons & Fractals*, **135**, 109853.

Received November 17, 2020; Revised December 23, 2020; Accepted January 5, 2021

한국 COVID-19 확진자 수에 대한 시계열 분석: HAR-TP-T 모형 접근법

유성민[a], 황은주[1,a]

[a]가천대학교 응용통계학과

요 약

이 논문에서는, 2개의 혼합된 t-분포(TP-T)의 오차과정을 따르는 이질적 자기회귀 (HAR) 모형을 이용하여, 한국 코로나 (COVID-19) 확진자 수 데이터에 대한 시계열 분석, 즉 추정과 예측에 대하여 연구한다. HAR-TP-T 시계열 모형을 고려하여 HAR 모형의 계수 뿐 아니라 TP-T 오차과정의 모수를 추정하고자 단계별 추정법을 제안한다. 본 연구에서 제안하고 있는 단계별 추정법은, HAR 계수 추정을 위해서는 통상적 최소제곱추정법을 채택하고, TP-T 모수 추정을 위해서는 최대우도추정법을 이용한다. 단계별 추정법에 대한 모의실험을 수행하여, 성능이 우수함을 입증한다. 한국 코로나 확진자 수에 대한 실증적 데이터 분석에서, HAR 모형에서의 차수 $p = 2, 3, 4$에 대해, 모형의 평균제곱오차가 최소가 되도록 하는 최적화 시간간격(optimal lag)을 포함하여, 여러가지 시간간격을 고려한 HAR-TP-T 모형의 모수 추정값을 계산한다. 제안된 단계별 추정방법과 기존의 MLE만의 방법을, 추정 결과를 제시함으로 함께 비교한다. 본 연구에서 제안하고 있는 추정은 두 가지의 오차 측면, 즉 HAR 모형의 평균제곱오차와 잔차분포에 대한 밀도함수 추정의 평균제곱오차, 두 측면에서 모두 우수함을 입증하였다. 나아가, 추정 결과를 활용한 코로나 확진자 수 예측을 수행하였고, 예측정확도의 한 측도로서 mean absolute percentage error (MAPE)를 계산하여 0.0953%의 매우 작은 오차값을 얻었다. 본 연구에서 선택한 최적화 시간간격을 고려한 HAR-TP-T 시계열 모형 및 단계별 추정 방법은, 정확한 한국 코로나 확진자 수 예측 성능을 제공한다고 할 수 있다.

주요용어: COVID-19, 이질적 자기회귀모형, Two-piece t-분포, 최소제곱추정법

본 연구는 가천대학교 교내연구과제 (GCU-202003640001) 지원을 받아 수행되었음.

[1]교신저자: (13120) 경기도 성남시 수정구 성남대로 1342, 가천대학교 응용통계학과. E-mail: ehwang@gachon.ac.kr