종단자료에 대한 unit root test의 이해와 세계 기온 자료에의 적용

2016-11797
수리과학부 한기현

1. 서론

본 연구는 시계열 자료에서 unit root의 존재 여부를 확인하는 unit root test의 이론적인 배경을 서술한다. 시계열 자료 중 동일한 설명변수에 대해 중복된 관측을 다루는 종단 자료(패널 자료, longitudinal data, panel data) 분석에 초점을 맞추어 이 자료들의 모형 설정과 이 모형들에 적합된 unit root test의 방법들을 탐구한다. 또한, 종단 자료의 시계열 분석에서 중요한 연구 중 하나인 K. Im과 M. Pesaran, Y. Shin의 연구[1]를 심도 있게 서술한다. 나아가, 본 연구는 세계 각지에 위치한 기상 관측소에서 추출된 1981년부터 2020년까지의 월평균 일 최고기온 자료를 시계열 특성을 가지는 종단 자료라고 가정하여 이 자료에 실제로 unit root test를 적용하고 기온 자료의 unit root의 존재성과 stationarity를 분석한다. 그 결과로 월별 기온 변화에는 unit root가 존재하지 않고 기온 변화가 stationary함을 시사한다.

2. 이론적 배경

2.1. Dickey-Fuller test와 Augmented Dickey-Fuller test

시계열 자료 y_t에 unit root가 존재한다는 것은 그 시계열이 stationary하지 못하다는 것을 의미하며, 이는 random walk처럼 시간 t의 변화에 따라 분산 $Var(y_t)$과 특정 시간차 l에 대하여 $Cov(y_t, y_{t-l})$의 값이 변화함을 뜻한다. 시계열 자료에 적용되는 unit root test의 기원은 D. Dickey와 W. Fuller의 연구에서 최초로 등장한 Dickey-Fuller Test(DF test)와 Augmented Dickey-Fuller test (ADF test)이다.

먼저, DF test에서 가정하는 모형은 다음과 같다.[2] Autocorrelation을 가지는 시계열 자료 y_t와 백색 소음(white noise)의 특성을 보이는 오차 항(error term) u_t에 대하여 다음 모형을 만족한다.

$$y_t = \alpha + \beta t + \rho y_{t-1} + u_t$$

이 모형은 $\Delta y_t = y_t - y_{t-1}$를 이용하여 나타내면 $\Delta y_t = \alpha + \beta t + (\rho - 1) y_{t-1} + u_t$와 동일한 식이다. DF test에는 다음과 같은 3가지 형태가 존재한다. (1)과 (2)는 귀무가설 $H_0 : \rho = 1$과 대립가설 $H_1 : |\rho| < 1$에 대한 검정, (3)은 귀무가설 $H_0 : \rho = 1$ and $\beta = 0$, $H_1 : not\ H_0$에 대한 검정이다.

---

1) Im, K.S.,Pesaran, M.H., Shin, Y., 2003. Testing for unit roots in heterogeneous panels, Journal of Econometrics, Volume 115, Issue 1, 53-74,

2) Dickey, D.A., Fuller, W.A., 1979. Distribution of the estimators for autoregressive time series with a unit root. Journal of the American Statistical Association 74, 427-431.

(1) $\alpha=\beta=0$일 때, $y_t=\rho y_{t-1}+u_t$ ($AR(1)$ 모형)

(2) $\beta=0$일 때, $y_t=\alpha+\rho y_{t-1}+u_t$ (drift가 있는 $AR(1)$ 모형)

(3) 일반적일 때, $y_t=\alpha+\beta t+\rho y_{t-1}+u_t$ (drift와 선형 deterministic trend가 있는 $AR(1)$ 모형)

$AR(1)$ 모형에 대한 검정이었던 DF test와 유사하게, ADF test는 $AR(p+1)$ 모형에 대한 검정이며 식으로 나타내면 다음과 같다.

$$y_t=\alpha+\beta t+\rho y_{t-1}+\theta_1\Delta y_{t-1}+\theta_2\Delta y_{t-2}+\cdots+\theta_p\Delta y_{t-p}+u_t$$

$$\Delta y_t=\alpha+\beta t+(\rho-1)y_{t-1}+\theta_1\Delta y_{t-1}+\theta_2\Delta y_{t-2}+\cdots+\theta_p\Delta y_{t-p}+u_t$$

ADF test 또한 3가지 형태가 존재한다. DF test와 마찬가지로 (1)과 (2)는 귀무가설 $H_0:\rho=1$과 대립가설 $H_1:|\rho|<1$에 대한 검정, (3)은 귀무가설 $H_0:\rho=1$ and $\beta=0$, $H_1:not\ H_0$에 대한 검정이다. 이 검정을 $ADF(p)$ 검정으로 나타내기로 한다. 이에 따라 DF test는 $ADF(0)$ 검정으로 쓰인다.

(1) $\alpha=\beta=0$일 때 $y_t=\rho y_{t-1}+\theta_1\Delta y_{t-1}+\theta_2\Delta y_{t-2}+\cdots+\theta_p\Delta y_{t-p}+u_t$ ($AR(p+1)$ 모형)

(2) $\beta=0$일 때 $y_t=\alpha+\rho y_{t-1}+\theta_1\Delta y_{t-1}+\theta_2\Delta y_{t-2}+\cdots+\theta_p\Delta y_{t-p}+u_t$ (drift가 있는 $AR(p+1)$ 모형)

(3) 일반적일 때, $y_t=\alpha+\beta t+\rho y_{t-1}+\theta_1\Delta y_{t-1}+\theta_2\Delta y_{t-2}+\cdots+\theta_p\Delta y_{t-p}+u_t$ (drift와 선형 deterministic trend가 있는 $AR(p+1)$ 모형)

DF와 ADF test에는 최대가능도비 검정(maximum likelihood ratio test)를 이용하여 도출된 t 검정통계량이 존재한다.

$$t=\frac{\hat{\rho}-1}{\sqrt{\hat{\sigma}^2}}$$ (단, $\hat{\sigma}^2$는 ordinary least square를 이용한 추정치)

이 통계량은 최소제곱법으로 계산되었기 때문에 잘 알려진 t 분포와 유사한 식 형태를 가지지만 시계열 자료의 autocorrelation의 존재성에 의해 t 분포와 완벽히 동일하지는 않다. 또한 이 통계량의 분포는 y_t의 총 길이인 T에 대하여 T가 충분히 커질 때 Browinan motion $W(t)$로 나타낸 분포로 수렴한다.[3] 본 보고서에서는 (1)과 (2) 형태의 검정에 대한 극한분포 ζ와 η만 기술한다.

(1)일 때 $\zeta=\dfrac{\frac{1}{2}\{W(1)^2-1\}}{\int_0^1 W(r)^2dr}$, (2)일 때 $\eta=\dfrac{\frac{1}{2}\{W(1)^2-1\}-W(1)\int_0^1 W(r)dr}{\int_0^1 W(r)^2dr-[\int_0^1 W(r)dr]^2}$

각 검정에서 t 검정통계량들의 분포는 Brownian motion에 의하여 제시되어 분위수를 비롯한 분포의 일부 특성이 명확히 계산되지 못하므로 3가지 형태에 따라 Monte Carlo 시뮬레이션을 활용하여 유의수준 0.01, 0.05에 대한 critical value들만이 표로 제시된다.

---

3) Hamilton, J.D., 1994. Time Series Analysis. Princeton University Press, Princeton. pp. 489-501

2.2. 종단 자료에 대한 Unit Root Test

종단 자료는 N개의 단위 중 단위 i ($i=1,2,\cdots,N$)에 대하여 특정한 공통된 변수 $y_{i,t}$를 시간 t의 흐름 ($t=0,1,\cdots,T_i$) 에 따라 기록한 자료들이다. 이 단위는 각 국가나 각 기업 또는 개인이 해당될 수 있으며 공통된 변수로는 사회, 경제, 문화적인 다양한 변수가 대입될 수 있다. 종단 자료 분석은 특정 시계열 변수가 여러 단위들에 걸쳐 어떠한 특성을 가지고 있는지를 보여준다.

종단 자료는 각 단위마다 자료가 존재하는 0부터 T_i까지의 시간 범위가 동일하면 balanced, 그렇지 않으면 unbalanced라 말한다. 또한, 단위마다 변수가 독립적인지 또는 각 단위 사이에 상관관계가 존재하는지에 따라 각각 독립 종단 자료와 cross-sectional 종단 자료로 분류한다.

일반적으로 종단 자료의 모형은 ADF test에서와 유사하게 다음과 같이 식으로 표현된다. 이 식에서 $d_{i,t}$는 선형성이나 절편의 존재성과 같은 deterministic trend를 의미한다.

$$\Delta y_{i,t} = d_{i,t} + (\rho_i - 1) y_{i,t-1} + \sum_{j=1}^{p_i} \theta_{i,j} \Delta y_{i,t-j} + u_{i,t}$$

그 예시로 A. Levin과 C. Lin는 balanced이고 독립인 종단 자료에 대한 unit root test(LL test)를 제시한다.[4] LL test는 ρ_i들이 모두 ρ로 일치한다는($\rho_1 = \rho_2 = \cdots = \rho_N = \rho$) 사실과 $u_{i,t}$들이 독립이며 동일한 정규분포를 따른다는 사실을 가정한다. 이 검정은 귀무가설 $H_0 : \rho = 1$와 대립가설 $H_1 : |\rho| < 1$에 대한 결과를 보여준다. 그들은 ADF에서처럼 $d_{i,t}$를 활용하여 모형의 drift와 deterministic trend의 유무를 변화시켜가며($\varnothing$, α_i, $\alpha_i + \beta_i t$ 등) 생성한 여러 모형을 다룬다. 이 연구는 중심극한정리를 이용하여 특정 가정 하에서 $N \to \infty$과 $T \to \infty$이면서 $N/T \to 0$일 때 각 모형의 $\hat{\rho}$와 검정통계량인 t_ρ 통계량이 정규분포로 수렴한다는 것을 증명한다. 그리고, Monte Carlo 시뮬레이션을 통해 유한한 N과 T에 대하여 critical value들을 유의수준에 따라 표로 소개한다.

2.3. IPS test에 대한 이해

K. Im, M. Peseran. Y. Shin은 서로 독립이고 오차 항 $u_{i,t}$가 i에 따라 이종적인(heterogenous) 분산 σ_i^2를 가지는 모형에 대한 unit root test(IPS test)를 실시한다.[5] 이들은 LLC test와 달리 모든 ρ_i가 일치한다고 가정하지 않고 귀무가설 $H_0 : \rho_1 = 1, \rho_2 = 1, \cdots \rho_N = 1$, 대립가설을 H_1: $\rho_1, \rho_2, \cdots, \rho_N$ 중 1이 아닌 것 존재로 설정한다.

먼저 이 연구는 $y_{i,t}$들이 각 i에 대하여 DF test의 (2)를 만족한다고 가정한다. $\rho_i - 1 = \gamma_i$로 쓰면 그 식은 다음과 같다.

$$\Delta y_{i,t} = \alpha_i + \gamma_i y_{i,t-1} + u_{i,t}, \quad i = 1, \cdots, N, \quad t = 1, \cdots, T$$

이 식을 T차원 벡터 $\Delta y_i = (\Delta y_{i1}, \cdots, \Delta y_{iT})'$, $\tau_T = (1, 1, \cdots, 1)'$, $y_{i,-1} = (y_{i,0}, y_{i,1}, \cdots, y_{i,T-1})'$,

---

4) Levin, A., Lin, C.F., 1993. Unit root tests in panel data: asymptotic and finite-sample properties.
5) Im, K.S.,Pesaran, M.H., Shin, Y., 2003. Testing for unit roots in heterogeneous panels, Journal of Econometrics, Volume 115, Issue 1, 53-74,

$u_i = (u_{i,1}, \cdots, u_{i,T})'$ 로 다시 서술하면 다음과 같은 식이 등장한다.

$$\Delta y_i = \alpha_i \tau_T + \gamma_i y_{i,-1} + u_i, \ i = 1, \cdots, N, \ t = 0, \cdots, T$$

서로 다른 i에 대하여 독립임을 가정하기 때문에 전체 자료 $y_{i,t}$에 대한 최소제곱법의 결과로 독립적인 γ_i의 추정량 $\widehat{\gamma_{iT}}$와 t 통계량 t_{iT}를 얻는다. $T \times T$ 행렬 $M_\tau = I_T - \frac{1}{T}\tau_T \tau_T'$에 대하여 $\widehat{\gamma_{iT}}$와 t_{iT}는 다음과 같이 쓰여진다.

$$\widehat{\gamma_{iT}} = \frac{\Delta y_i' M_\tau y_{i,-1}}{y_{i,-1}' M_\tau y_{i,-1}}$$

$$t_{iT} = \frac{\widehat{\gamma_{iT}} (y_{i,-1}' M_\tau y_{i,-1})^{1/2}}{\sqrt{\widehat{\sigma_{iT}^2}}}$$

이 때, $\widehat{\sigma_{iT}^2} = \frac{\Delta y_i' M_{X_i} y_{i,-1}}{T-2}$(단, $X_i = (\tau_T, y_{i,-1})$이고 $M_{X_i} = I_T - X_i (X_i' X_i)^{-1} X_i'$)으로 최소제곱법에 의해 결정된 MSE(Mean Squared Error)의 값이다. 데이터가 balanced이므로 T가 i에 대하여 모두 동일하고, 따라서 각 t_{iT}는 H_0 하에서 동일한 분포를 가진다. IPS test의 $t-bar$ 통계량은 이 점을 이용하기 위하여 다음과 같이 정의된다.

$$t-bar_{NT} = \frac{1}{N} \sum_{i=1}^{N} t_{iT}$$

t_{iT}가 T가 증가함에 따라 앞서 DF test의 (2)의 t 검정통계량의 극한분포 η_i로 수렴한다는 사실이 알려져 있다. 또한, $T > 5$일 때 t_{iT}의 2차 적률이 존재한다는 것은 Cauchy-Schwarz 부등식을 이용하여 증명된다. 따라서 $t-bar$ 통계량을 표준화하여 다음과 같은 Z_{tbar}을 얻으면, 중심극한정리에 의해 Z_{tbar}은 N의 증가에 따라 표준정규분포로 수렴한다,

$$Z_{tbar} = \sqrt{N} \frac{t-bar_{NT} - E(t_{iT})}{\sqrt{var(t_{iT})}} \Rightarrow N(0,1)$$

그러나 이 연구는 실제 계산 결과 N이 작을 때에도 Z_{tbar}의 분위수와 표준정규분포의 분위수가 큰 차이가 없음을 언급한다. 즉, N이 작을 때에도 Z_{tbar}가 IPS test의 unit root test로서 효과를 보장한다는 것을 의미한다.

나아가, $ADF(p)$ 검정에서와 같이 서로 다른 i에 대하여 $y_{i,t}$가 $AR(p_i + 1)$에 적합된 모형에 대해서도 $t_{iT}(p_i, \theta_i)$는 T가 증가함에 따라 η_i로 분포수렴한다. 동일한 방법으로 만들어진 Z_{tbar}은 $N, T \to \infty$이고 어떤 음 아닌 실수 k에 대하여 $N/T \to k$일 때 표준정규분포로 수렴함이 증명된다.

이와 유사하게 이 연구는 $t-bar_{NT}$의 표준화를 $\gamma_i = 0$ 조건부 하에서 진행한 W_{tbar}도 T와 N이 커질 때 표준정규분포로 수렴함을 제안한다.

$$W_{tbar} = \sqrt{N}\frac{t-bar_{NT} - E(t_{iT}(p_i,\theta_i)|\gamma_i = 0)}{\sqrt{var(t_{iT}(p_i,\theta_i)|\gamma_i = 0)}} \Rightarrow N(0,1)$$

실제 자료에 적용할 때 최적의 p_i의 값은 AIC(Akaike Information Criterion)이나 SIC(Schwarz Bayesian Information Criterion) 등으로 구해진다. 일반적인 종단 자료 ARMA 시계열 모형에 적용하였을 때, 앞서 LL test는 $p_i \geq 1$인 $ADF(p_i)$ 검정에서 검정력이 매우 작기 때문에 실제로 자료에 unit root가 없음에도 불구하고 귀무가설을 채택하는 제2종 오류가 빈번히 발생할 수 있다는 단점을 지니는데, IPS test의 W_{tbar}은 검정력이 크기 때문에 이 문제점을 보완할 수 있다.

2.4. Cross-sectional 종단 자료에 대한 Unit root test

M. Pesaran의 연구는 cross-sectional 종단 자료 시계열 모형에 대한 unit root test 방법을 제시한다.[6] 모형의 오차 항 $u_{i,t}$가 시간에 대한 성분 f_t와 각 단위 i에 대한 성분 g_i의 곱으로 표현된다는 가정이 적용된다. 이 때, $\epsilon_{i,t}$는 모든 i와 t에 대하여 독립이다.

$$u_{i,t} = g_i f_t + \epsilon_{i,t}$$

f_t항이 $\overline{y_t} = \frac{1}{N}\sum_{i=1}^{N} y_{i,t}$들의 일차결합으로 표현된다는 가정 하에 이 모형은 다음과 같은 형태로 변형된다.

$$\Delta y_{i,t} = d_{i,t} + (\rho_i - 1)y_{i,t-1} + \sum_{j=1}^{p_i}\theta_{i,j}\Delta y_{i,t-j} + c_i\overline{y_{t-1}} + \sum_{j=0}^{p_i}\Psi_{i,j}\Delta\overline{y_{t-j}} + \epsilon_{i,t}$$

이 모형에서 $t_{\rho,i}$가 유사하게 얻어지지만, 이들의 i에 대한 평균 $C = \frac{1}{N}\sum_{i=1}^{N} t_{\rho,i}$는 앞서 IPS test에서처럼 정규분포를 따르지 않아 이 통계량의 critical value는 시뮬레이션을 통해 얻어진다.

3. 연구 결과

3.1. 자료 소개

본 연구는 1981년 1월부터 2020년 12월까지 세계 각 지역의 매일 최고기온의 월평균 수치 자료를 분석한다. 이 자료는 세계 기상관측소의 기온 및 기상 자료를 보관하고 온라인으로 제공하는 National Centers for Environmental Information에서 획득되었다. 웹사이트에 저장된 기온 자료는 화씨온도로 기록되어있다. (https://www.ncdc.noaa.gov/data-access)

이 자료들은 기상관측소들이 각 단위에 해당하며 시계열 특성을 가지는 종단자료로 해석될 수 있다. 세계의 기상관측소 중 대부분은 1981년 1월부터 2020년 12월까지 40년의 기간 동안 월별 기온 자료의 5% 이상이 결측값(missing value)으로 남아 있다. 480개의 월 중 결측값이 5% 이내인 기상관측소는 그리 많지 않았으며 이 기상관측소 중 7곳을 찾아내어 본 연구에 월별 일 최고기온자료를 이용하였다. 본 연구에서 분석한 기상관측소들의 이름과 위치, 결측값의 개수와 비율은

6) Pesaran, M.H. (2007), A simple panel unit root test in the presence of cross-section dependence. J. Appl. Econ., 22: 265-312.

표 1이 제시하는 바와 같다.

표 1 기상관측소 명단 (순번/이름/위치/위도/경도/결측값 개수/결측값의 비율)

순번(i)	기상관측소 이름	위치	위도($^\circ$)	경도($^\circ$)	결측값	비율(%)
1	BAKERSFIELD AIRPORT	미국 캘리포니아	35.43	-119.05	0	0
2	BARROW AIRPORT	미국 알래스카	71.28	-156.78	0	0
3	CHUUK WEATHER SERVICE OFFICE AIRPORT	미크로네시아 연방	7.46	151.85	9	1.88
4	CONDOBOLIN AG RESEARCH STATION	호주	-33.06	147.22	7	1.46
5	KOKPEKTY	카자흐스탄	48.75	82.36	5	1.04
6	SPRINGFIELD WEATHER SERVICE OFFICE AIRPORT	미국 미주리	37.23	-93.39	0	0
7	TURAIF	사우디아라비아	31.69	38.73	20	4.17

각 기상관측소들에 대하여 일 최고기온의 월평균 자료의 결측이 일어난 월은 그 관측소의 다른 월의 자료를 통해 메꾸어졌다. 결측값을 찾아주기 위하여 월평균 일 최고기온은 autocorrelation 없이 선형 추세와 1년을 주기로 하는 계절적인 추세가 동시에 존재한다고 가정했다. 즉, i번째 기상관측소 $(i=1,2,\cdots,7)$에서 t번째 월 $(t=1,2,\cdots,480)$의 월평균 일 최고기온 $y_{i,t}$가 다음 모형을 만족한다고 가정했다. 이 때, 오차 항 $\epsilon_{i,t}$는 서로 다른 i와 t에 대하여 독립이며 $N(0,\sigma_i^2)$를 따른다.

$$y_{i,t}=\beta_{i,0}+\beta_{i,1}t+\sum_{j=0}^{11}\gamma_{i,j}\cos(\frac{2\pi}{12}(t-j))+\epsilon_{i,t}$$

결측값이 존재하는 기상관측소 i에 대하여, 결측이 일어나지 않은 $y_{i,t}$들을 토대로 이 모형에 최소제곱법을 적용하여 $\widehat{\beta_{i,0}},\widehat{\beta_{i,1}},\widehat{\gamma_{i,j}}$을 얻었다. 기상관측소 i마다 서로 자료가 독립임을 가정하기 때문에 추정값 $\widehat{\beta_{i,0}},\widehat{\beta_{i,1}},\widehat{\gamma_{i,j}}$은 그 i번째 기상관측소의 자료인 $y_{i,t}$만을 이용하여 도출되었다. 만약 t^*번째 월에서 결측되었다면 $\widehat{y_{i,t^*}}$는 위 모형에 각 모수의 추정치를 대입한 값인 다음 값으로 두었다. 이 방법을 통해 우리는 자료를 balanced한 종단 자료로 변환시킬 수 있다.

$$\widehat{y_{i,t^*}}=\widehat{\beta_{i,0}}+\widehat{\beta_{i,1}}t^*+\sum_{j=0}^{11}\widehat{\gamma_{i,j}}\cos(\frac{2\pi}{12}(t^*-j))$$

예를 들어, 7번째 기상관측소(Turaif)는 본 연구에서 다루는 480개월 중 460개월만이 기록되어있고 20개월은 결측값이다. 그림 1은 이 기상관측소의 결측값을 추정치(빨간 색)로 메꾼 그래프이다.

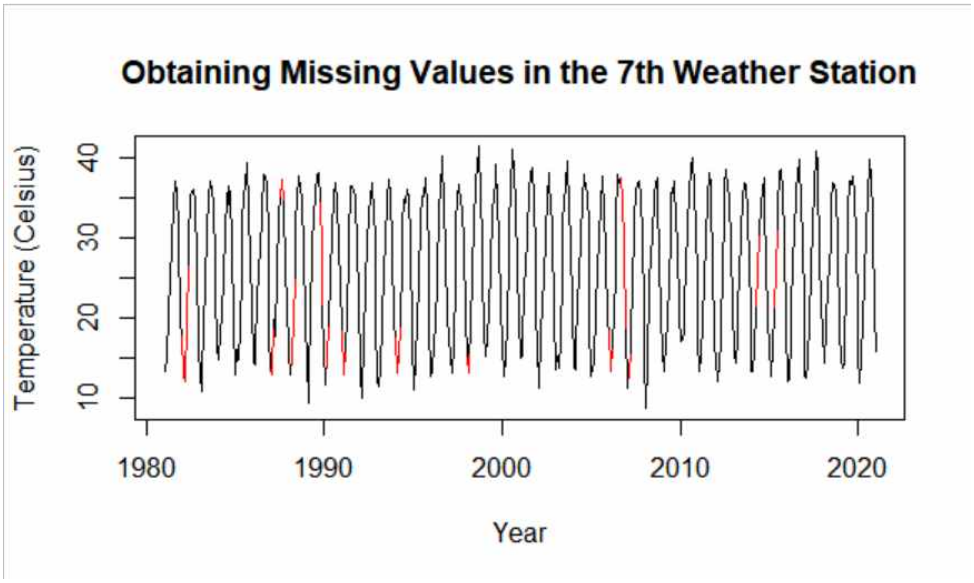

그림 1 기상관측소 7(Turaif)의 월평균 일 최고기온의 결측값의 추정치(빨간 색)

3.2. Autocorrelation의 확인

종단자료인 월평균 일 최고기온 자료에 autocorrelation이 존재하는지 여부를 확인하기 위하여 다음과 같은 모형을 고려한다. $T_{i,t}$는 선형 deterministic trend이고 $S_{i,t}$는 계절형 deterministic trend이다. 이는 앞서 결측값을 찾기 위한 모형과 유사한 형태이지만 이 모형은 오차 항 $u_{i,t}$의 autocorrelation을 가정한다.

$$y_{i,t} = T_{i,t} + S_{i,t} + u_{i,t}$$

Autocorrelation을 확인하기 위하여 본 연구는 최소제곱법을 통해 $T_{i,t}$와 $S_{i,t}$를 구한 후, 오차 항인 $u_{i,t}$에 각각 독립적인 Ljung-Box test를 적용한다. 기상관측소 i에 대한 p-value들은 표 2와 같다. 유의수준 $\alpha = 0.01$으로 Ljung-Box test 검정했을 때, $u_{i,t}$의 autocorrelation이 없다는 귀무가설을 $i = 2,3,4$에 대해서만 기각할 수 있다. 그런데 $y_{i,t}$는 종단자료로서 서로 다른 i에 걸쳐 동일한 특성을 가진다고 가정할 때, 모든 i에 대하여 $u_{i,t}$에 autocorrelation이 모두 있거나 모두 없어야 한다. 즉, 우리는 Bonferroni test를 활용하여 귀무가설 H_0: "$u_{i,t}$ 모두 autocorrelation 없음"이라는 새로운 귀무가설을 설정한다. 이에 대하여 유의수준 $\alpha = 0.01$으로 검정을 실시하면, 7개의 p-value 중 가장 작은 3번째 기상관측소의 값은 $\alpha/N = \frac{1}{700}$보다 작아지므로 귀무가설을 기각할 수 있다. 즉, 모든 i에 대하여 $u_{i,t}$에 autocorrelation이 존재한다는 가정이 가능하다는 것을 뜻한다.

표 2 Ljung-Box Test 결과

기상관측소(i)	1	2	3	4	5	6	7
p-value	0.986	5.64e-4	1.67e-5	2.18e-3	0.158	0.691	0.210

적합된 $u_{i,t}$의 값은 1번째와 4번째 기상관측소에 대하여 다음과 같은 형태를 보인다. Ljung-Box test의 p-value가 왼쪽은 0.986이고, 오른쪽은 0.002인데, 육안으로 차이를 구분하기 어렵다.

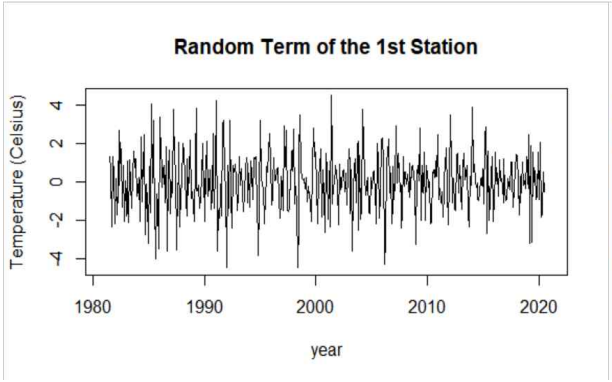

그림 2 1번째 기상관측소의 $u_{i,t}$

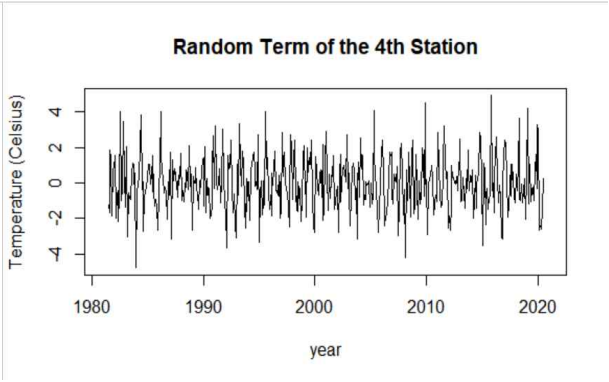

그림 3 4번째 기상관측소의 $u_{i,t}$

3.3. Panel data로서 Unit root test의 적용

앞서 인용한 논문들에 등장한 unit root test들은 계절적인 deterministic trend가 없는 선형 추세까지를 가정한 모형에 대해서만 적용된다. 따라서 본 연구는 $y_{i,t} = T_{i,t} + S_{i,t} + u_{i,t}$에 unit root test를 적용하기 위하여 $y_{i,t}$에서 계절 성분 $S_{i,t}$를 배제한 새로운 변수 $v_{i,t} = y_{i,t} - S_{i,t}$를 도입하여 종단 자료 시계열 분석을 진행하였다. $v_{i,t}$에는 선형 deterministic trend와 drift의 존재성을 가정할 수 있다. 우리는 $v_{i,t}$가 i에 따라 독립인 종단 자료라고 가정하였다.

만약 $v_{i,t}$를 종단 자료로 관찰하여 unit root test를 적용하였을 때 unit root가 존재한다면, 이는 월평균 일 최고기온 데이터가 계절성에 추가적으로 random walk 형태의 통계적인 추세 (stochastic trend)를 따른다는 것을 의미한다. 구체적으로는 시간의 흐름에 따라 기온 자료의 분산이 증가함을 암시하며 기상학적인 관점에서 지구온난화의 추세가 추측하기 어려운 방향으로 진행됨을 뜻한다.

$$v_{i,t} = \alpha_i + \beta_i t + \rho_i v_{i,t-1} + \theta_{i,1}\Delta v_{i,t-1} + \theta_{i,2}\Delta v_{i,t-2} + \cdots + \theta_{i,p_i}\Delta v_{i,t-p_i} + \epsilon_{i,t}$$

기상관측소 $i = 1, 2, \cdots 7$에 대하여 ρ_i가 모두 동일하다고 가정하는 LL test와 ρ_i가 모두 다를 수 있다는 IPS test를 적용하였다. R의 'plm' 패키지를 참고하여 "drift가 존재"하는 모형(ADF test의 (2))과 "drift와 선형 deterministic trend가 동시에 존재"하는 모형(ADF test의 (3))에 대한 검정을 실시하였다. p_i는 AIC를 비교하는 방법으로 결정되었다. 통계량과 p-value는 표 3에 제시되어 있다.

표 3 LL test와 IPS test 적용 결과

	LL test (통계량: t_ρ)		IPS test (통계량: W_{tbar})	
	drift만 존재	drift & trend	drift만 존재	drift & trend
통계량	-39.116	-65.988	-37.992	-47.772
p-value	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16

검정 결과로, 모든 검정에 대하여 p-value가 2.2×10^{-16} 미만의 매우 작은 값이 되어 유의수준 $\alpha=0.001$ 하에서 unit root가 존재한다는 귀무가설을 기각할 수 있음이 확인된다. 따라서, 99.9% 신뢰수준으로 $v_{i,t}$에 종단 자료로서의 unit root가 존재하지 않음이 통계적으로 입증된다. 즉, $v_{i,t}$가 stationary하다는 결론이 도출된다.

4. 결론 및 제언

본 연구는 1981년 1월부터 2020년 12월까지 40년간 월평균 일 최고기온 자료를 종단 자료로 고려하여 분석하였다. 이 자료들은 계절성과 선형성 추세를 제거하였을 때 autocorrelation이 존재함이 확인되었다. 또한, 계절성 추세를 제거한 자료에서 LL test와 IPS test를 비롯한 종단 자료 unit root test를 적용한 결과 unit root가 존재하지 않으며 stationary함이 강력하게 입증되었다. 즉, 월평균 기온 자료는 계절성 추세에 시간에 따른 분산의 변화가 없는 stationary 항이 추가된 값을 가진다. 지구과학적인 관점에서는 현대에 대두되는 지구온난화를 비롯한 기후 변화가 기온의 편차에 중대한 변화를 일으키지 않았다는 결과가 시사된다.

LL test의 경우 $p_i \geq 1$인 $ADF(p_i)$ 모형이 적합된 자료는 그 검정력이 현저히 감소하여, 대립가설이 옳음에도 대립가설을 채택하지 않는 제2종 오류의 확률이 높아짐이 알려져 있다. 실제 적합 결과, LL test의 7개 단위 중 일부의 단위는 $ADF(1)$ 모형으로 적합되었음이 확인되었고, p-value가 매우 낮은 값이라는 사실로부터 대립가설이 높은 신뢰도로 채택된다. 즉, 본 연구는 기온 자료에 unit root가 존재하지 않는다는 것이 매우 유의함을 통계적으로 입증한다.

또한, 실제로 LL test와 IPS test의 검정통계량은 N과 T가 무한대로 발산할 때 극한분포로 수렴한다. 본 연구에서 적용된 종단자료의 시간의 길이 T는 480으로 크지만, 단위의 개수 N은 7로 작다. 만약 연구에 쓰인 40년의 기간을 줄이거나 다른 기간으로 옮기는 과정을 통해 기온 자료를 결측 없이 보존하는 다른 기상관측소를 추가로 발견한다면, N이 증가하여 두 unit root test의 결과에 대한 신뢰도가 증가할 것이다. 두 검정의 p-value가 지나치게 낮게 등장한 이유도 N이 T에 비하여 지나치게 작기 때문으로 추정된다.

본 연구에 이용된 unit root test는 deterministic trend로 최대 선형성만을 가정한다. 따라서 우리는 어쩔 수 없이 기온 자료에 계절성 추세를 제거한 자료로 검정을 실시하였다. 계절성 추세를 포함한 종단 자료의 unit root test는 Otero의 2005년과 2007년 연구에서 등장한다. 추후에 이 기법을 본 연구에 쓰인 기온 자료에 적용하여 효율적으로 unit root test를 적용할 수 있을 것이다.

5. 참고문헌

[1] Im, K.S., Pesaran, M.H., Shin, Y., 2003. Testing for unit roots in heterogeneous panels, Journal of Econometrics, Volume 115, Issue 1, 53-74,

[2] Dickey, D.A., Fuller, W.A., 1979. Distribution of the estimators for autoregressive time series with a unit root. Journal of the American Statistical Association 74, 427-431.

[3] Hamilton, J.D., 1994. Time Series Analysis. Princeton University Press, Princeton.

[4] Levin, A., Lin, C.F., 1993. Unit root tests in panel data: asymptotic and finite-sample properties. Unpublished manuscript, University of California, San Diego.
[5] National Centers for Environmental Information, 과거 세계 기상 자료 제공, U.S. Government - National Oceanic and Atmospheric Administration, https://www.ncdc.noaa.gov/data-access
[6] Croissant, Y., 2021, Linear Models for Panel Data (R Package 'plm'), https://cran.r-project.org/web/packages/plm/plm.pdf
[7] 이상열, 2013, 시계열분석 이론 및 SAS 실습, 자유아카데미
[8] 조신섭, 손영숙, 2019, SAS/ETS를 이용한 시계열분석, 율곡출판사
[9] Kirchgässner, G., Wolters, J., Hassler, U., 2007, Introduction to Modern Time Series Analysis. Springer