The Korean Journal of Applied Statistics (2016) DOI: http://dx.doi.org/10.5351/KJAS.2016.29.1.157
29(1), 157–169

A comparison study on regression with stationary nonparametric autoregressive errors

Kyusang Yu[a,1]

[a]Department of Applied Statistics, Konkuk University

(Received December 15, 2015; Revised December 31, 2015; Accepted December 31, 2015)

Abstract

We compare four methods to estimate a regression coefficient under linear regression models with serially correlated errors. We assume that regression errors are generated with nonlinear autoregressive models. The four methods are: ordinary least square estimator, general least square estimator, parametric regression error correction method, and nonparametric regression error correction method. We also discuss some properties of nonlinear autoregressive models by presenting numerical studies with typical examples. Our numerical study suggests that no method dominates; however, the nonparametric regression error correction method works quite well.

Keywords: nonparametric autoregressive model, regression, efficiency

1. 서론

고전적인 회귀모형에서 오차항이 서로 독립이고 오차항의 분산이 동일한 경우 통상적 최소제곱(OLS) 추정법이 최소분산불편 추정량임이 잘 알려져 있다. 또한 오차항에 이분산성이 있거나 상관관계가 있는 경우 오차항의 공분산행렬을 이용한 일반화 최소제곱(GLS) 추정법이 최소제곱추정량의 분산을 줄여준다는 사실도 잘 알려져 있다. 경험적 연구에서 사용되는 자료들은 시간에 따라 관측되는 경우가 많다. 특히, 경제학 등에서 사용되는 변수들은 경제현상을 시간의 진행에 따라 측정하는 경우가 많다. 이러한 자료를 이용한 연구 중에서 외생변수(exogenous variable)의 효과를 추정하기 위한 회귀모형을 고려하는 경우가 있다. 이러한 연구에서는 설명변수와 오차항은 서로 무상관을 가정하지만 오차항은 시간에 따른 자기 상관을 가정하는 경우가 많다. 또한 시계열 분석에서 외부적 요인에 대한 추세를 제거하는 선행 작업이 요구되는 경우가 많다. 따라서, 이러한 자료들을 분석하는 경우에는 자료의 자기상관 구조가 중요한 고려요소가 된다. 시간 $t = 1, 2, \ldots$에 대하여 아래의 회귀 모형을 고려해보자.

$$Y_t = \beta_0 + \beta^T X_t + Z_t, \tag{1.1}$$

$$Z_t = m(Z_{t-1}, \ldots, Z_{t-p}) + \epsilon_t, \tag{1.2}$$

This paper was written as part of Konkuk University's research support program for its faculty on sabbatical leave in 2015.

[1]Department of Applied Statistics, Konkuk University, 120 Neungdong-ro, Gwangjin-gu, Seoul 05029, Korea. E-mail: kyusangu@konkuk.ac.kr

여기서 Z_t는 회귀모형의 오차항으로 식별 조건 $E(Z_t|X_t) = 0$을 만족한다고 가정한다. 또한 ϵ_t는 서로 독립이고 같은 분포를 갖는 평균이 0이고 유한한 분산($\sigma_e^2 < \infty$)을 갖는 확률변수들로 가정한다. 만약 식 (1.2)에서 함수 m을 0으로 가정하면 위의 모형은 임의 표본(random sample)에서의 회귀모형을 나타낸다. 또한, 함수 m을 선형함수 $\sum_{j=1}^{p} \phi_j Z_{t-j}$로 가정하면 $\{Z_t\}$는 p차 자기회귀과정 AR(p)을 따르는 확률과정이 된다. 이러한 자기상관이 있는 오차항을 갖는 모형에서는 통상적 최소제곱추정량이 일반화 최소제곱추정량에 비해 큰 분산을 갖는 것이 알려져있다. 따라서 회귀계수 β를 추정할 때 오차항 확률과정 $\{Z_t\}$의 공분산 구조를 반영하는 일반화 최소제곱추정량을 사용하는 것이 더 효율적이다. 위의 모형에서 함수 함수 m이 선형함수 $\sum_{j=1}^{p} \phi_j Z_{t-j}$일 때, 회귀계수 β의 효율적인 추정에 대한 연구는 Sheather (2009) 등 다양한 문헌에서 찾아 볼 수 있다.

한편, 비대칭적 경기 순환(asymmetric business cycles), 생물 계체수의 주기적 변화, 주식시장에서의 변동성 등에서 경험적으로 관측이 된 극한주기궤도(limit cycles) 현상이나 다중고정점(multiple fixed points) 현상, 혹은 상태 변환이 있는 시계열 자료들은 함수 m이 선형함수인 정상 시계열 모형으로 설명하기 어렵다는 것이 알려져 있다. 이러한 경우 비선형 시계열 모형을 사용할 수 있다. 이러한 현상을 모형화하는 모수적 비선형 회귀모형에는 Tong과 Lim (1980)에 의해 제안된 SETAR(self-exciting threshold autoregressive) 모형, Haggan과 Ozaki (1981)에 의해 제안된 ExpAR(amplitude-dependent exponential autoregressive) 모형 등이 있다. 이러한 모수적 비선형 회귀모형에 대한 자세한 논의는 Tong (1990)에서 찾아 볼 수 있다. 이와 같은 모수적 비선형 모형은 선형 시계열 모형에서 설명할 수 없는 여러 현상들에 대해 부분적인 해법을 주지만 일반적인 비선형 시계열 자료에 적용하는데는 한계가 있다. 이러한 경우 비모수 모형을 사용하는 것이 대안이 될 수 있으며 모형 (1.2)에서 함수 m의 특정한 형태를 가정하지 않는 비모수자기회귀(NPAR) 모형을 고려할 수 있다. Truong과 Stone (1992)은 정상 비모수자기회귀모형(stationary NPAR)을 포함한 정상 비모수 시계열 모형에서 기저함수급수를 이용한 추정법의 점근적 성질을 규명하였으며 국소선형추정량(local linear estimator)의 성질은 Fan과 Gijbels (1996)의 저서에 자세히 설명되어 있다.

이 논문에서는 오차항이 정상(stationary) 비선형 자기회귀 모형을 따르는 경우, 비선형 자기회귀모형의 성질을 몇 가지 비선형 함수 m을 설정하여 수치적 특성과 이에 따라 발생된 시계열자료의 특징을 모의실험을 통해 설명하고 이러한 구조를 활용한 회귀계수 추정의 효율성 개선 방법들의 통계적 성질에 대하여 기술하는 것이 목적이다. 이 논문의 구성은 다음과 같다. 2장에서는 비모수 자기회귀모형에 대한 개괄적 설명을 하고 3장에서 오차항이 자기상관이 있을 때 회귀계수의 추정량의 효율을 개선하는 방법들을 설명하며 4장에서는 여러가지 모의실험의 방법과 결과를 소개한다. 마지막으로 모의실험 결과를 종합하며 제시점을 요약한다.

2. 비모수 자기회귀 모형

서론의 회귀모형 중 오차항 모형을 따로 고려하자.

$$Z_t = m(Z_{t-1}, \ldots, Z_{t-p}) + \epsilon_t, \tag{2.1}$$

여기서 함수 m을 선형함수가 아닌 일반적인 연속함수로 가정하면 p차 비모수 자기회귀모형 NPAR(p)가 된다. 위의 모형에서 백색잡음(white noise) 시계열인 $\{\epsilon_t\}$가 서로 독립이고 같은 분포를 갖으며 다음의 조건들을 만족한다고 가정하자.

E1. $E(\epsilon_t|Z_{t-1}, Z_{t-2}, \ldots) = 0$.

E2. $E(\epsilon_t^2|Z_{t-1}, Z_{t-2}, \ldots) = \sigma_e^2$.

위의 가정에서 E2는 일반적인 경우 $Z_{t-1}, Z_{t-2}, \ldots$에 의존하는 이분산 모형으로 약화시킬 수 있으나 이 논문에서는 단순한 설명을 위해 등분산 구조를 가정하였다.

모형 (2.1)에서 함수 m이 선형함수 $\sum_{j=1}^{p} \phi_j Z_{t-j}$인 경우에는 시계열 $\{Z_t\}$의 정상성 등의 성질이 특성방적식의 해의 크기를 통해서 결정된다. 또한 Yule-Walker 방정식을 이용하거나 회귀모형을 이용하여 계수 ϕ_j를 추정할 수 있다. 그러나 함수 m에 선형 가정을 하지 않는 비모수 자기회귀모형인 경우 정상성 등의 성질이 쉽게 이해되지 않고 추정법 또한 자명하지 않다. 비모수 자기회귀모형을 포함한 비선형 자기회귀모형의 정상성에 대해서는 Bhattachary와 Lee (1995) 등의 많은 연구자들이 충분조건들을 제사하고 있다. 여기서는 Biscay 등 (2005)이 사용한 충분 조건을 소개한다. 벡터 $X_t = (Z_t, \ldots, Z_{t-p+1})^T$라 하고 벡터 $F(X_{t-1}) = (m(Z_{t-1}, \ldots, Z_{t-p}), Z_{t-1}, \ldots, Z_{t-p+1})^T$라 하자. 이러한 표현법은 모형 (2.1)을 상태공간(state-space) 형식 $X_t = F(X_{t-1}) + U_t$으로 표현하는데 사용된다. 여기서 $U_t = (\epsilon_t, 0, \ldots, 0)^T \in \mathbb{R}^p$이다. 이러한 표현법을 사용하여 비모수 자기회귀모형을 따르는 시계열 $\{Z_t\}$가 정상분포를 갖는 충분조건은 다음과 같다.

S1. 오차항 ϵ_t가 모든 영역에서 양의 확률밀도함수값을 갖는다.

S2. $\mathbb{R}^p$의 임의의 0이 아닌 벡터열 $\{X_n\}$에 대하여 다음의 조건을 만족한다.

$$\limsup_n \frac{\|F(X_n)\|_2}{\|X_n\|_2} < 1.$$

서론에서 설명하였듯이 비선형 자기회귀모형은 일반적 자기회귀모형과 다른 성질을 갖는다. 다음의 예를 통하여 함수 m이 비선형 부드러운 함수인 비모수자기회귀모형을 살펴보자. 아래의 예는 모두 정상성을 만족하는 1차 비모수자기회귀 모형들 이다.

예제 2.1:

$$Z_t = Z_{t-1} \exp\left(-Z_{t-1}^2\right) + \epsilon_t, \tag{2.2}$$

여기서 ϵ_t는 서로 독립이고 평균이 0, 분산이 σ_e^2인 정규분포를 따른다.

예제 2.1은 $p = 1$이고 함수 $m(x) = x\exp(-x^2)$로 주어진 경우이다. 이 함수 m은 한개의 고정점을 갖는다. 즉, $x = m(x)$의 방정식을 $x = 0$가 푼다. Figure 2.1의 첫 번째 열 좌측 그림은 함수 m과 직선 $y = x$를 겹쳐 그린 그림이다. 이 그림에서 볼 수 있는 구조는 함수 m이 고정점을 0에서만 갖고 x값이 커지면 함수값이 0 근방으로 다시 돌아오는 성질이다. 첫 열의 우측 그림은 $\sigma_e^2 = 0.5$로 설정하여 모형 (2.2)에서 2,000개의 자료를 생성한 후 확률밀도함수를 추정한 그림이다. 0근방에서의 확률 밀도함수 함수가 뾰족하게 커짐을 볼 수 있다. 그림의 하단은 생성된 자료 전반부 500개를 그림 시계열도이다. 굵은 쇄선은 초기치를 각각 $-1, 0, 1$로 주고 $\sigma_e^2 = 0$, 즉, 잡음이 없는 시계열의 그림으로 고정점인 0 근방으로 수렴하고 있는 모습을 보인다. 점선은 $\sigma_e^2 = 0.5$로 설정한 모형에서 생성한 자료에 대한 시계열도이다. 이 자료에 대하여 선형 자기회귀 모형을 적합한 결과 모형의 차수는 4차가 선택되었고 자기상관계수들은 $\phi_1 = 0.3309,\ \phi_2 = -0.0231,\ \phi_3 = -0.0459,\ \phi_4 = 0.0337$로 추정되었다.

예제 2.2:

$$Z_t = \log\left(1 + Z_{t-1}^2\right) + \epsilon_t, \tag{2.3}$$

여기서 ϵ_t는 서로 독립이고 평균이 0, 분산이 σ_e^2인 정규분포를 따른다.

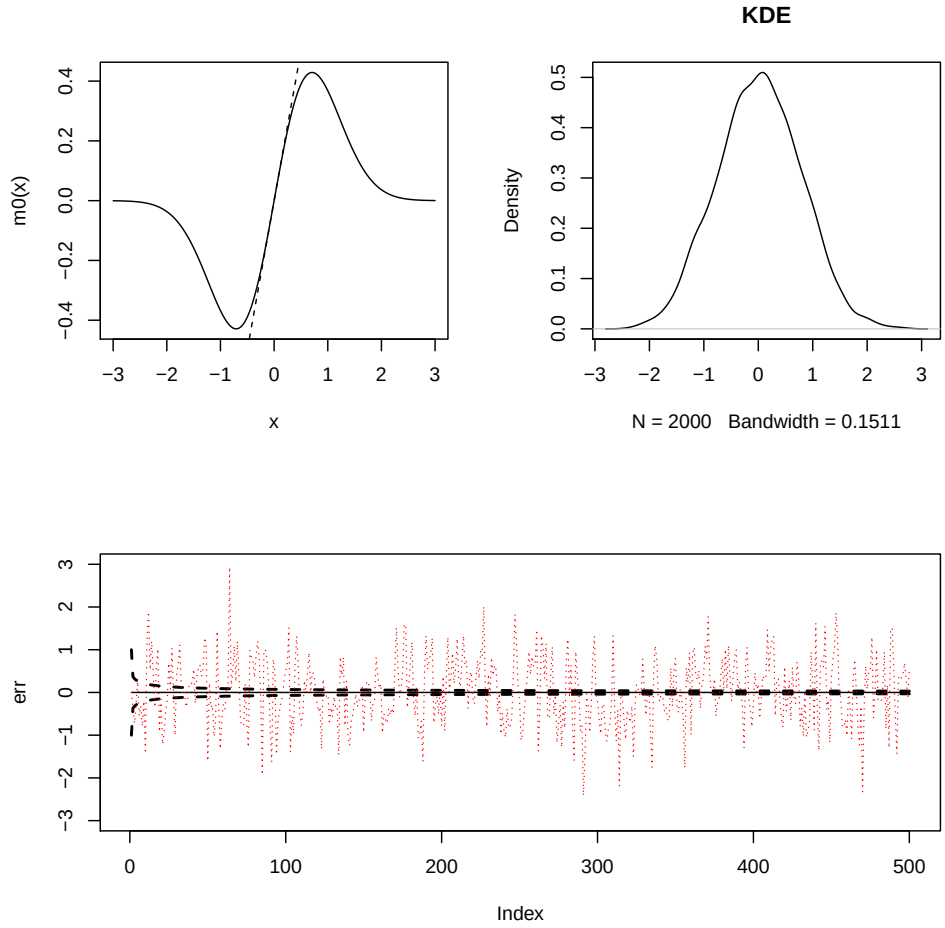

Figure 2.1. NAR(1) with $m(x) = x\exp(-x^2)$. The upper left panel shows the shape of m with the reference line $y = x$. The upper right panel shows the density estimator of the stationary density of the process. The lower panel shows the time series plot of the generated time series.

예제 2.2는 $p = 1$이고 함수 $m(x) = \log(1 + x^2)$으로 주어진 경우이다. 이 함수 m은 한개의 고정점을 갖는다. 즉, $x = m(x)$의 방정식을 $x = 0$가 푼다. Figure 2.2의 첫 번째 열 좌측 그림은 함수 m과 직선 $y = x$를 겹쳐 그린 그림이다. 이 그림에서 볼 수 있는 구조는 함수 m이 고정점을 0에서만 갖고 x값이 커지면 함수값이 $y = x$ 보다 천천히 증가하는 특징이 있다. 첫 열의 우측 그림은 $\sigma_e^2 = 0.5$로 설정하여 모형 (2.3)에서 2,000개의 자료를 생성한 후 확률밀도함수를 추정한 그림이다. 모형 (2.2)와 달리 0 근방에서 뾰족한 모양이 관측되지 않는다. 그림의 하단은 생성된 자료 전반부 500개를 그림 시계열도이다. 굵은 쇄선은 초기치를 각각 $-1, 0, 1$로 주고 $\sigma_e^2 = 0$, 즉, 잡음이 없는 시계열의 그림으로 고정점인 0 근방으로 빠르게 수렴하고 있는 모습을 보인다. 점선은 $\sigma_e^2 = 0.5$로 설정한 모형에서 생성한 500개 자료에 대한 시계열도이다. 이 자료에 대하여 선형 자기회귀 모형을 적합한 결과 모형의 차수는 5차가 선택되었고 자기상관계수들은 $\phi_1 = 0.3836,\ \phi_2 = 0.0549,\ \phi_3 = 0.0361,\ \phi_4 = -0.0333,\ \phi_5 = 0.0464$로 추정되었다.

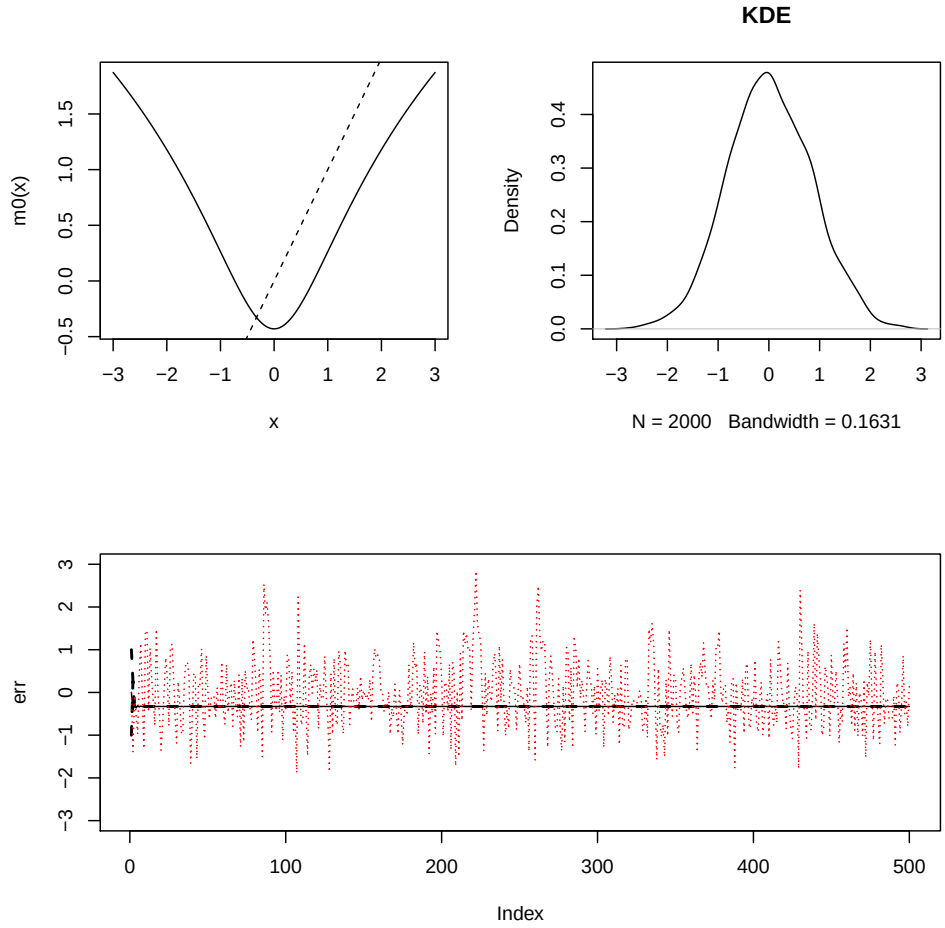

Figure 2.2. NAR(1) with $m(x) = \log(1 + x^2)$. The upper left panel shows the shape of m with the reference line $y = x$. The upper right panel shows the density estimator of the stationary density of the process. The lower panel shows the time series plot of the generated time series.

예제 2.3:

$$Z_t = \sin\left(\frac{\pi}{2} Z_{t-1}\right) + \frac{1}{2} Z_{t-1} + \epsilon_t, \tag{2.4}$$

여기서 ϵ_t는 서로 독립이고 평균이 0, 분산이 σ_e^2인 정규분포를 따른다.

예제 2.3은 $p = 1$이고 함수 $m(x) = \sin((\pi/2)x) + (1/2)x$로 주어진 경우이다. 이 함수 m은 세 개의 고정점을 갖는다. 즉, $x = m(x)$의 방정식을 $x = 0, \pm 1.472969$가 푼다. Figure 2.3의 첫 번째 열 좌측 그림은 함수 m과 직선 $y = x$를 겹쳐 그린 그림이다. 이 그림에서 볼 수 있는 구조는 함수 m이 고정점을 $0, \pm 1.472969$에서 갖고 x값이 커지면 함수값이 $y = x$ 보다 천천히 증가하는 특징이 있다. 첫 열의 우측 그림은 $\sigma_e^2 = 0.5$로 설정하여 모형 (2.3)에서 2,000개의 자료를 생성한 후 확률밀도함수를 추정한 그림이다. 모형 (2.2), (2.3)과 달리 두 고정점 ± 1.472969 근방에서 봉우리가 관측되는 양봉 분포형태를 갖

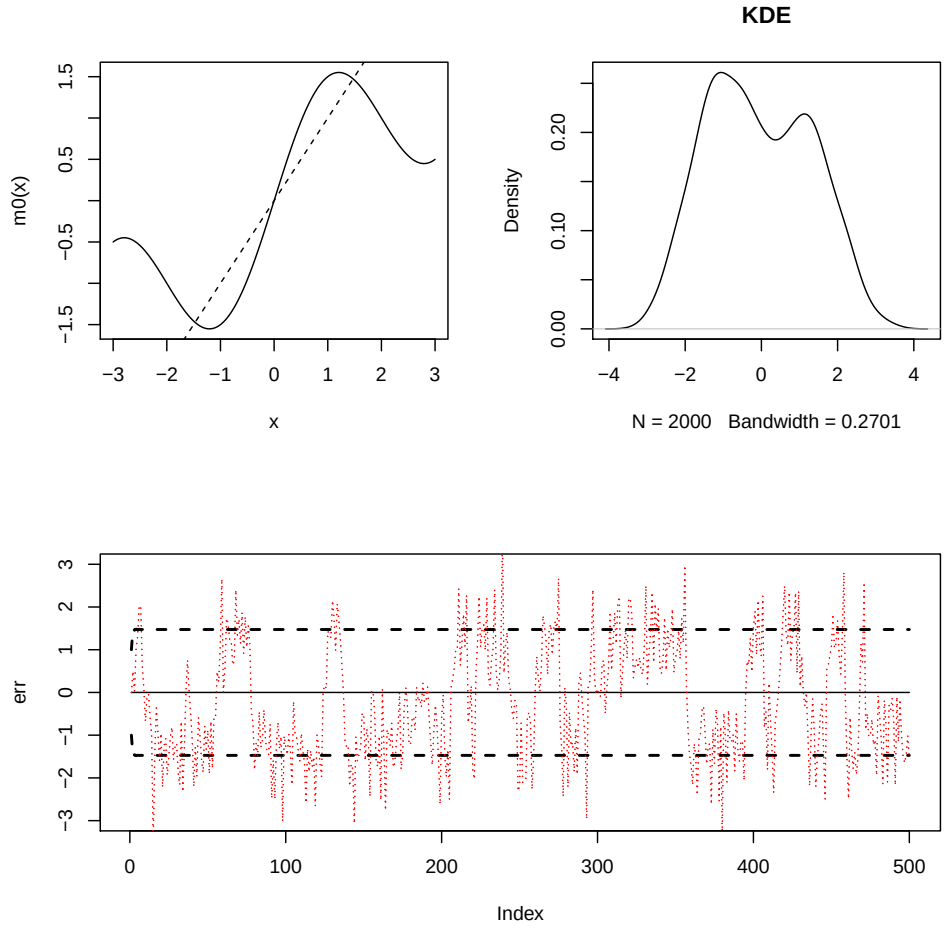

Figure 2.3. NAR(1) with $m(x) = \sin((\pi/2)x) + (1/2)x$. The upper left panel shows the shape of m with the reference line $y = x$. The upper right panel shows the density estimator of the stationary density of the process. The lower panel shows the time series plot of the generated time series.

는다. 이는 함수 m의 0에서의 미분값이 1.570796이고 ± 1.472969에서의 미분값이 -0.5625724이어서 0 근방에서는 국소적으로 폭발하는 성질이 있어 0 근방의 값을 취하면 임의 변동이 누적되어 다른 점으로 이동하려는 경향이 강하고 ± 1.472969 근방에서는 국소적으로 정상 자기회귀모형의 성질을 갖기 때문에 그 근방에서 변동하는 성질을 갖기 때문이다. 그림의 하단은 생성된 자료 전반부 500개를 그림 시계열도이다. 굵은 쇄선은 초기치를 각각 $-1, 0, 1$로 주고 $\sigma_e^2 = 0$, 즉, 잡음이 없는 시계열의 그림으로 고정점인 $0, \pm 1.472969$ 근방으로 빠르게 수렴하고 있는 모습을 보인다. 점선은 $\sigma_e^2 = 0.5$로 설정한 모형에서 생성한 500개 자료에 대한 시계열도이다. 위의 정상 확률밀도함수 추정에서 관측했듯이 시계열도에서도 자료들이 고정점 ± 1.472969 근방에서 변동하고 있는 현상을 볼 수 있다. 이 자료에 대하여 선형 자기회귀 모형을 적합한 결과 모형의 차수는 13차가 선택되었고 자기상관계수들은 $\phi_1 = 0.4772$, $\phi_2 = 0.2048$, $\phi_3 = 0.0708$, $\phi_4 = 0.0353, \phi_5 = 0.1042, \phi_6 = -0.0068$, $\phi_7 = -0.0582$, $\phi_8 = -0.0744$, $\phi_9 = 0.0410$, $\phi_{10} = 0.0374$, $\phi_{11} = -0.0787$, $\phi_{12} = -0.0942$, $\phi_{13} = 0.1208$로 추정되었다.

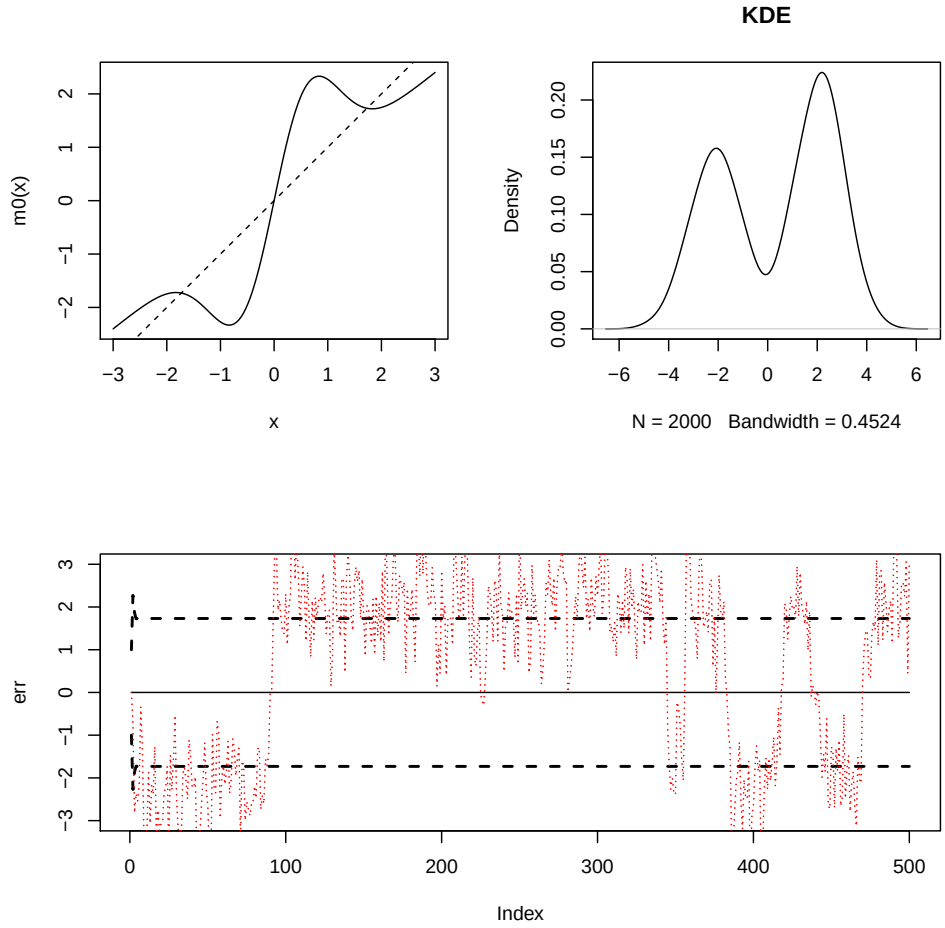

Figure 2.4. NAR(1) with $m(x) = (0.8 + 4 * \exp(-x^2))x$. The upper left panel shows the shape of m with the reference line $y = x$. The upper right panel shows the density estimator of the stationary density of the process. The lower panel shows the time series plot of the generated time series.

예제 2.4:

$$Z_t = \left(0.8 + 4 * \exp\left(-Z_{t-1}^2\right)\right) Z_{t-1} + \epsilon_t, \tag{2.5}$$

여기서 ϵ_t는 서로 독립이고 평균이 0, 분산이 σ_e^2인 정규분포를 따른다.

예제 2.4는 $p = 1$이고 함수 $m(x) = (0.8 + 4 * \exp(-x^2))x$로 주어진 경우이다. 이 함수 m은 세 개의 고정점을 갖는다. 즉, $x = m(x)$의 방정식을 $x = 0, \pm 1.730818$가 푼다. Figure 2.4의 첫 번째 열 좌측 그림은 함수 m과 직선 $y = x$를 겹쳐 그린 그림이다. 이 그림에서 볼 수 있는 구조는 함수 m이 고정점을 $x = 0, \pm 1.730818$에서 갖고 x값이 커지면 함수값이 $y = x$ 보다 천천히 증가하는 특징이 있다. 첫 열의 우측 그림은 $\sigma_e^2 = 0.75$로 설정하여 모형 (2.5)에서 2,000개의 자료를 생성한 후 확률밀도함수를 추정한 그림이다. 모형 (2.4)보다 더 뚜렷하게 두 고정점 ± 1.730818 근방에서 봉우리가 관측

되는 양봉 분포형태를 갖는다. 이는 함수 m의 0에서의 미분값이 4.8이고 ± 1.730818에서의 미분값이 -0.19829이어서 0 근방에서는 국소적으로 강하게 폭발하는 성질이 있어 0 근방의 값을 취하면 임의 변동이 누적되어 다른 점으로 이동하려는 경향이 강하고 ± 1.730818 근방에서는 국소적으로 정상 자기회귀모형의 성질을 갖기 때문에 그 근방에서 변동하는 성질을 갖기 때문이다. 그림의 하단은 생성된 자료 전반부 500개를 그림 시계열도이다. 굵은 쇄선은 초기치를 각각 $-1, 0, 1$로 주고 $\sigma_e^2 = 0$, 즉, 잡음이 없는 시계열의 그림으로 고정점인 $0, \pm 1.730818$ 근방으로 빠르게 수렴하고 있는 모습을 보인다. 점선은 $\sigma_e^2 = 0.75$로 설정한 모형에서 생성한 500개 자료에 대한 시계열도이다. 위의 정상 확률밀도함수 추정에서 관측했듯이 시계열도에서도 자료들이 고정점 ± 1.730818 근방에서 변동하고 있는 현상을 볼 수 있다. 이 자료에 대하여 선형 자기회귀 모형을 적합한 결과 모형의 차수는 6차가 선택되었고 자기상관계수들은 $\phi_1 = 0.5504,\ \phi_2 = 0.2041,\ \phi_3 = 0.0866,\ \phi_4 = -0.0188,\ \phi_5 = 0.0275,\ \phi_6 = 0.0849$로 추정되었다.

3. 자기상관구조를 이용한 회귀계수 추정에서의 효율성 개선

서론의 회귀모형 (1.2)에서 설명변수 X_t의 차원이 1이고 $p = 1$일 때, 함수 m이 선형함수이면 모형 (1.2)는 다음과 같이 정리할 수 있다.

$$Y_t = \beta_0 + \beta_1 X_t + Z_t,$$
$$Z_t = \phi Z_{t-1} + \epsilon_t.$$

여기서 자기회귀계수 ϕ가 $|\phi| < 1$을 만족하면 확률과정 $\{Z_t\}$는 정상과정(stationary process)가 된다. 이러한 경우, 시점 t와 $t-h$에서의 오차항 Z_t의 상관계수는 시점 t와 무관하게 ϕ^h으로 주어지고 분산 또한 시점 t와 무관하게 $\sigma_e^2(1-\phi^2)^{-1}$로 주어진다. 따라서 시점 $t = 1, \ldots, T$에서 관측된 자료의 오차항의 공분산 행렬 Σ_e는 다음과 같은 형태로 주어진다.

$$\Sigma_e = \frac{\sigma_e^2}{1-\phi^2} \begin{pmatrix} 1 & \phi & \cdots & \cdots & \phi^{T-1} \\ \phi & 1 & \phi & \cdots & \phi^{T-2} \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \phi \\ \phi^{T-1} & \cdots & \cdots & \phi & 1 \end{pmatrix}.$$

이러한 공분산 구조를 이용한 일반화 최소제곱추정법이 회귀계수 β_1의 추정효율을 높일 수 있다. 여기서 $\mathcal{X}$를 절편을 포함한 계획행렬(design matrix), $\mathcal{Y}$를 반응터(response vector)라 하면, 오차항 Z_t의 자기공분산 행렬 Σ_e가 알려진 경우, 일반화 최소제곱추정량 $\hat{\beta}_1^{o,GLS}$는 $(0,1)(\mathcal{X}^T\Sigma_e^{-1}\mathcal{X})^{-1}\mathcal{X}^T\Sigma_e^{-1}\mathcal{Y})$로 정의된다. 하지만 실제 자료분석에서는 일반적으로 오차항 확률과정 $\{Z_t\}$의 모형 구조나 자기상관이 알려져 있지 않다. 따라서, 통상적 최소제곱법 추정을 통해 잔차를 구한 후 잔차에 대한 시계열 분석을 통해 오차항에 대한 모형을 선택하고 자기상관을 추정하는 과정을 거치게 된다. 이때, 시계열 모형 선택과 모수추정에서 우연 변이(random variation)가 생기므로 오차항의 공분산 구조를 이용하는 GLS의 효율성 개선이 크지 않은 경우가 발생할 수 있다. 또한, 선택된 자기회귀 모형의 차수 p가 큰 경우에는 GLS의 해가 수치적으로 불안정한 경우가 많다. 특히, 오차항의 자기회귀 구조가 비선형인 경우 GLS의 적용에는 어려운 점이 많다. 이러한 경우 오차항의 자기회귀 구조를 이용하여 회귀모형 (1.2)를 다음의 모형으로 표현할 수 있다.

$$Y_t = \beta_0 + \beta_1 X_t + m(Z_{t-1}) + \epsilon_t. \tag{3.1}$$

만약, 회귀모형 오차항 Z_t가 관측이 가능하다면 새로운 회귀모형은 X_t와 Z_{t-1}을 설명변수로 갖는 부분선형 모형이 된다. 여기서, 만약 설명변수 X_t가 회귀모형의 오차항 Z_t의 과거와 무상관이면, 즉, $E(X_t|Z_{t-1},\ldots) = E(X_t)$이면, 완전한 외생변수로 간주할 수 있고 $E(X_t|Z_{t-1},\ldots) = h(Z_{t-1},\ldots)$이면 내생성을 갖는 모형으로 해석할 수 있다. 이 논문에서는 X_t가 외생 변수인 경우에 중점을 두어서 설명한다.

모형 (3.1)과 모형 (1.1)을 비교하면 모형 (1.1)에서는 회귀모형의 오차항이 Z_t이고 모형 (3.1)에서는 ϵ_t이다. X_t가 외생 변수이고 Z_t가 정상 비모수 1차 자기회귀 모형을 따르는 경우, 모형 (1.1)의 회귀계수 β_1의 통상적 최소제곱 추정량 $\hat{\beta}_1^{OLS}$의 분산은 $\mathrm{var}(m(Z_t)) + \mathrm{var}(\epsilon_t)$에 비례한다. 또한, 같은 가정의 모형 (3.1)에서 Z_t가 관측 가능하다면 회귀계수 β_1의 추정 문제에서 효율추정의 분산 하한은 $\mathrm{var}(\epsilon_t)$에 비례한다. 따라서 오차항의 자기상관 구조를 활용하면 모형 (1.1)에서 회귀계수 β_1의 통상적 최소제곱 추정량의 효율성에 대한 개선의 여지가 있다. 여기서는 Su와 Ullah (2006)의 방법에 기반한 추정량을 소개한다.

Su와 Ullah (2006)의 방법을 이해하기 위해 모형 (1.2)에서 회귀오차 Z_t가 관측이 가능한 경우를 생각해보자. 설명의 편의를 위해 $p = 1$인 경우를 가정한다. 이러한 경우 모형 (1.2)의 방정식은 비모수 자기회귀모형이 된다. 따라서 함수 m이 두번 미분가능하고 조건 E1, E2, S1, S2를 만족하며 적당한 mixing-계수 조건과 적률 조건을 만족하는 경우 함수 m에 대한 다음의 조건을 만족하는 비모수 커널 추정량을 구할 수 있다.

NE1. $\bar{m}(x) = C_1 h^2 + R_n(x)$이고 랜덤함수 $\sup|R_n(x)| = O_p((\log n/nh)^{1/2})$를 만족한다. 또한, 랜덤함수 $R_n(\cdot)$은 거의 모든 점에서 확률 1로 Lipschitz 연속이다.

여기서 h는 자료의 개수 n이 증가할수록 0으로 수렴하는 수열이다. 이러한 조건을 만족하는 비모수 추정량 $\bar{m}$을 이용하여 모형 (3.1)에서 반응 변수 Y_t를 $\check{Y}_t = Y_t - \bar{m}(Z_{t-1})$로 수정하고 이를 설명변수 X_t를 이용하여 최소제곱법을 사용하여 적합한다. 이렇게 얻어진 회귀계수 β_1의 추정량은 띠폭(bandwidth) h가 $h = o(n^{-1/4})$이고 $1/(nh) = o(1)$인 조건을 만족하면 점근적 분산(asymptotic variance)이 가상의 모형 $Y_t = \beta_0 + \beta_1 X_t + \epsilon_t$에서의 통상적 최소제곱추정량의 점근적 분산과 같아진다. 하지만 이러한 추정량은 회귀모형의 오차항 Z_t가 관측이 불가능하기 때문에 실제로 구현될 수는 없다. 따라서 이러한 발상을 현실에 적용하기 위해서는 관측이 불가능한 회귀오차 Z_t을 대체할 수 있는 예측치들이 필요하다. 이 논문에서는 이러한 예측치를 모형 (1.1)에서 통상적 최소제곱법으로 추정한 회귀식에서 구한 잔차들을 이용한다. 아래는 이러한 과정을 정리한 것이다.

Step 1. 모형 $Y_t = \beta_0 + \beta_1 X_t + Z_t$에서 통상적 최소제곱법을 이용하여 잔차 $\tilde{Z}_t$를 구한다.

Step 2. 모형 $Z_t = m(Z_{t-1}) + \epsilon_t$에서 $\{Z_t\}$를 $\{\tilde{Z}_t\}$로 대체하여 함수 m에 대한 비모수추정량$\hat{m}$을 구한다. 여기서 국소선형회귀(local linear regression) 또는 Nadaraya Watson 추정량을 사용할 수 있다.

Step 3. 새로운 반응변수 $\tilde{Y}_t = Y_t - \hat{m}(\tilde{Z}_{t-1})$을 사용하여 통상적 최소제곱추정량을 구한다.

Su와 Ullah (2006)은 회귀모형 (1.1)에서 회귀함수가 선형이 아닌 비모수 모형을 고려하여 이러한 절차로 구한 추정량이 회귀오차 Z_t를 관측한 경우와 같은 점근 분포(asymptotic distribution)를 갖는 것을 보였다. 모형 (1.1)에서도 같은 성질을 증명할 수 있을 것으로 기대되고 그 증명은 Su와 Ullah (2006)의 증명을 확장함으로 가능할 것이라 예상된다. 이 논문에서는 방법론의 소개와 수치적 결과에 주목하는 것이 목적이므로 이에 대한 증명은 시도하지 않는다. 다음 장에서는 위에 소개한 방법과 선형 자기회귀모형을 활용한 방법들을 모의실험으로 비교하여 유한 표본에서의 성질에 대한 논의를 진행한다.

4. 모의실험 연구

이 장에서는 비선형 자기회귀모형을 따르는 회귀모형의 오차항을 2장에서 기술한 예 2.1–2.4에 의하여 생성하여 3장에서 기술한 방법론과 선형자기회귀모형을 이용한 몇 가지 방법론을 모의실험을 통하여 비교한다. 모의실험에서 비교하는 추정량은 다음과 같다.

방법 1. 통상적 최소제곱 추정량 $\hat{\beta}_1^{ols}$.

방법 2. 통상적 최소제곱 추정량을 이용한 잔차를 AR(p) 모형을 적합하여 구한 오차항의 자기공분산행렬을 사용한 일반화 최소제곱 추정량 $\hat{\beta}_1^{gls}$.

방법 3. 통상적 최소제곱 추정량을 이용한 잔차를 AR(p) 모형을 적합하여 3장에서 기술한 수정한 반응변수를 이용한 추정량 $\hat{\beta}_1^{AR.C}$.

방법 4. 통상적 최소제곱 추정량을 이용한 잔차를 NAR(1) 모형을 적합하여 3장에서 기술한 수정한 반응 변수를 이용한 추정량 $\hat{\beta}_1^{NAR.C}$.

방법 2와 방법 3은 정상 시계열이 AR(∞)로 표현될 수 있음에 착안하여 비선형 정상 자기회귀 모형을 AR(p)모형으로 근사하는 방법을 이용한 것이다. 방법 2와 방법 3에서는 자기회귀모형의 차수 p의 선택과 계수의 추정은 R의 ar( )함수를 사용하여 구했으며 방법 4에서 띠폭은 R의 npregbw( )함수를 이용하여 교차검증(cross validation)방법으로 구했다. 또한 비모수 추정을 위해서는 국소선형회귀방법을 사용하였다. 또한 회귀계수는 $\beta_1 = 2$로 설정하였으며 설명변수 X_t는 표준 정규분포에서 생성하였다. 각 모의실험은 200회씩 반복하였으며 표본의 크기 $n = 100, 200, 300$으로 설정하였다. 자료수가 $n = 100, 200$인 경우는 자기회귀모형의 차수 p를 10으로 제한하였으며 $n = 300$인 경우는 20으로 제한하였다. 방법 4의 경우 같은 추정된 회귀 모형을 이용하여 다시 잔차를 구하고 같은 방법을 반복적으로 적용하는 방법도 고려할 수 있다. 이러한 경우 추정량의 점근적 성질은 차이가 없을 것으로 예상되지만 유한 표본에서의 추정량의 분산이 개선될 가능성은 있다. Table 4.1에 보고하지는 않았지만 방법 4를 반복적으로 적용한 경우 표본의 크기 $n = 100, 200$인 예제 2.3과 예제 2.4의 모형에서 약 3–5%의 분산이 감소되었다. 표본의 크기 $n = 300$인 경우는 반복적용으로 인한 개선이 없었다. 방법 4의 반복 적용은 표본수가 비교적 작은 경우에 다중 고정점을 갖는 경우 약간의 개선이 있으나 전반적으로는 일회 적용한 방법과 큰 차이는 없었다.

Table 4.1은 모의실험에서 구해진 추정량들의 분산을 보여준다. 모든 경우에서 편의의 제곱은 분산에 비해 매우 작은 값으로 무시할 수 있는 정도이다. 전반적으로 방법 2의 $\hat{\beta}_1^{gls}$와 방법 4의 $\hat{\beta}_1^{NAR.C}$ 분산이 다른 추정량들에 비해 작음을 알 수 있다. 각 모든 추정량에서 표본수가 증가하면 분산이 작아지는 것을 관측할 수 있지만 통상적 최소제곱추정량 $\hat{\beta}_1^{ols}$에 비해 $\hat{\beta}_1^{gls}$과 $\hat{\beta}_1^{NAR.C}$에서 그러한 경향이 더 두드러진다. 이는 잔차를 이용하여 공분산을 추정하거나 회귀 오차항의 비선형 자기상관에 대한 비모수추정이 더 안정적이기 때문에 발생하는 현상으로 해석할 수 있다. 특이한 점은 $\hat{\beta}_1^{gls}$의 분산이 $\hat{\beta}_1^{AR.C}$의 분산보다 일반적으로 작은 것이 관측된 결과이다. 이는 특히 예제 2.3과 예제 2.4에서 두드러지는데 두 모형의 경우 두 개의 고정점 근방에서 비교적 안정적이고 비슷한 자기회귀 모형, 즉, 평균 수준은 다르면서 비슷한 자기상관 구조를 갖는 모형으로 근사되는 특징이 있다. 따라서 잔차의 평균에 대한 선형 수정을 이용한 방법보다 잔차의 자기상관 구조를 이용하는 일반화 최소제곱법의 성능이 더 좋은 것으로 나타날 수 있다. 예제 2.1 모형에서는 네 방법이 큰 차이를 보이지 않는다. 이는 예제 2.1의 비선형 자기회귀모형을 정의하는 함수 $m(x) = x\exp(-x^2)$의 그래프에서 볼 수 있듯이 회귀오차항의 값이 커지면 0근처로 다시 돌아오는 성질이 있어 $\{Z_t\}$의 동역학계(dynamic system)가 분산의 크기가 커지는 것을 제어하고 있기 때문으로 생각된다. 예제 2.2와 예제 2.3의 모형은 3장에서 제안된 방법이 다른 방법에 비해 매우

Table 4.1. Variances of estimators: The variances are multiplied by 10^3 and the numbers in the parenthesis represent the relative efficiency compared to $\hat{\beta}_1^{NAR.C}$

		$\hat{\beta}_1^{ols}$	$\hat{\beta}_1^{gls}$	$\hat{\beta}_1^{AR.C}$	$\hat{\beta}_1^{NAR.C}$
$n = 100$	예제 2.1	8.483	9.199	9.550	9.005
		(0.942)	(1.021)	(1.060)	(1.000)
	예제 2.2	6.493	6.449	6.685	5.511
		(1.178)	(1.170)	(1.213)	(1.000)
	예제 2.3	15.249	8.290	12.161	5.933
		(2.570)	(1.397)	(2.050)	(1.000)
	예제 2.4	30.732	9.532	14.722	9.142
		(3.362)	(1.043)	(1.610)	(1.000)
$n = 200$	예제 2.1	5.117	5.057	5.127	4.666
		(1.097)	(1.084)	(1.099)	(1.000)
	예제 2.2	3.874	3.908	3.913	2.962
		(1.308)	(1.320)	(1.321)	(1.000)
	예제 2.3	9.562	4.250	6.089	3.099
		(3.086)	(1.372)	(1.965)	(1.000)
	예제 2.4	20.753	4.192	7.202	4.214
		(4.924)	(0.995)	(1.709)	(1.000)
$n = 300$	예제 2.1	3.514	3.239	3.583	3.310
		(1.062)	(0.979)	(1.083)	(1.000)
	예제 2.2	2.904	3,014	2.971	1.956
		(1.484)	(1.541)	(1.518)	(1.000)
	예제 2.3	5.554	2.206	4.504	1.941
		(2.856)	(1.136)	(2.321)	(1.000)
	예제 2.4	12.609	2.691	5.510	2.933
		(4.299)	(0.917)	(1.879)	(1.000)

Table 4.2. Distribution of selected order for AR(p) approximation

		1사분위수	중앙값	3사분위수	평균
$n = 100$	예제 2.1	1.00	1.00	2.00	1.33
	예제 2.2	0.00	0.00	1.00	0.78
	예제 2.3	2.00	2.00	3.00	2.55
	예제 2.4	1.00	2.00	3.00	2.00
$n = 200$	예제 2.1	1.00	1.00	1.00	1.39
	예제 2.2	0.00	0.00	1.00	0.89
	예제 2.3	2.00	2.00	3.00	2.82
	예제 2.4	2.00	3.00	4.00	3.29
$n = 300$	예제 2.1	1.00	1.00	2.00	2.00
	예제 2.2	0.00	0.00	1.00	0.88
	예제 2.3	2.00	3.00	4.00	3.19
	예제 2.4	3.00	4.00	5.00	4.23

우수한 경우이다. 두 경우는 $\{Z_t\}$를 생성하는 동역학계가 비선형성을 강하게 갖으며 큰 값에 대하여 증가하는 경향이 있는 경우이다. 예제 2.4의 모형은 0이 아닌 두 고정점 주위로 시계열이 밀집하는 형태이다. 이때는 방법 2의 $\hat{\beta}_1^{gls}$와 방법 4의 $\hat{\beta}_1^{NAR.C}$가 다른 방법들에 비해 잘 작동하는 것으로 나타난다.

Table 4.2는 표본수 별 각 모형에서 선택되어진 자기회귀모형 AR(p)의 선택된 차수의 분포를 정리한

것이다. 예제 2.1과 예제 2.2의 비선형 모형의 경우 선택되는 차수가 표본수에 따라 크게 변하지 않는다. 또한 예제 2.2의 모형에서 정상 자기회귀모형에 기반한 방법들이 통상적 최소제곱법과 크게 차이가 나지 않는 이유를 보여준다. 예제 2.4의 모형에서는 자료의 수가 늘어날수록 선택되는 차수가 커지는 경향을 볼 수 있다. 이를 통하여 자료의 수가 증가함에 따라 방법 2의 $\hat{\beta}_1^{gls}$의 성능이 좋아지는 이유를 어느정도 설명할 수 있다. 표에는 보고되지 않았지만 예제 2.3의 모형에서는 자기회귀 차수 p가 12보다 크게 선택되는 경우가 모형에 비해서 많았다. 이는 2장에서 2,000개의 자료를 생성해서 적합한 결과에서 13차 자기회귀모형이 선택된 결과에서 유추가 가능하다.

5. 결론

본 연구에서는 몇 가지 유형의 비선형 자기회귀 모형의 성질을 살펴보고 이러한 비선형 자기회귀 모형을 따르는 오차항을 갖는 회귀모형에서 최소제곱추정량의 효율을 개선하는 방법을 살펴보았다. 모의실험 연구에서 고려한 세가지 개선방법은 비선형 구조를 직접 활용하는 비모수자기회귀적합을 이용하는 방법과 정상 시계열의 자기회귀모형으로 근사를 이용하는 방법을 비교해보았다. 모의실험결과 구조적 비선형성이 강하고 또한 비선형함수가 계속 증가하거나 감소하는 경우에 방법 4의 $\hat{\beta}_1^{NAR.C}$가 다른 방법에 비해 우수한 성능을 보였고, 비선형함수가 양 극단에서 0으로 수렴하는 경우는 네 방법이 큰 차이를 보이지 않았다. 또한, 예제 2.4의 모형과 같이 다수의 고정점에 강하게 밀집하는 경우는 방법 2의 $\hat{\beta}_1^{gls}$의 성능이 우수함을 알수있었다. 이는 각 고정점 사이의 실현값이 매우 적게 나타나서 비모수 함수 추정에 어려움이 있기 때문이라고 추측한다. 본 연구에서는 회귀모형의 설명변수 X_t가 외생변수인 경우만 고려하였으나 X_t가 회귀오차 Z_t의 과거에 상관이 있는 내생변수인 경우에는 더욱 흥미로운 결과를 도출할 수 있을 것으로 기대된다. 이러한 경우는 준모수 부분선형모형의 방법과 이론을 전개할 수 있을 것으로 기대된다.

References

Bhattacharya, R. and Lee, C. (1995). Ergodicity of nonlinear first order autoregressive models, *Journal of Theoretical Probability*, **8**, 210–219.

Biscay, R. J., Lavielle, M., and Ludeña, C. (2005). Estimation of nonparametric autoregressive time series models under dynamical constraints, *Journal of Time Series Analysis*, **26**, 371–397.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*, Chapman & Hall, London.

Haggan, V. and Ozaki, T. (1981). Modelling nonlinear random vibrations using an amplitude dependent autoregressive time series model, *Biometrika*, **68**, 189–196.

Sheather, S. (2009). *A Modern Approach to Regression with R*, Springer, New York.

Su, L. and Ullah, A. (2006). More efficient estimation in nonparametric regression with nonparametric autocorrelated errors, *Econometric Theory*, **22**, 98–126.

Tong, H. (1990). *Nonlinear Time Series: A Dynamical Approach*, Oxford University Press, Oxford.

Tong, H. and Lim, K. (1980). Threshold autoregression, limit cycles and cyclical data (with Discussion), *Journal of Royal Statistical Society B*, **42**, 245–292

Truong, Y. K. and Stone, C. (1992). Nonparametric function estimation involving time series, *The Annals of Statistics*, **20**, 77–97.

정상 비모수 자기상관 오차항을 갖는 회귀분석에 대한 비교 연구

유규상[a,1]

[a]건국대학교 응용통계학과

(2015년 12월 15일 접수, 2015년 12월 31일 수정, 2015년 12월 31일 채택)

요 약

이 논문에서는 비선형 자기회귀 과정을 따르는 오차항을 포함한 회귀모형에서 계수추정법의 비교를 다룬다. 비교를 위해 통상적 최소제곱추정량, 일반화 최소제곱추정량, 모수적 회귀오차 수정법, 비모수적 회귀오차 추정법을 비교하였다. 본 논문에서는 또한 비선형 자기회귀모형의 성질을 전형적인 몇가지 비선형자기회귀 모형을 예를 들어 설명한다. 비교연구의 결과 네 가지 추정량 중에 모든 상황에서 최선인 추정량은 존재하지 않았으나 비모수 회귀오차 수정방법이 일반적으로 우수한 성능을 보임을 알 수 있다.

주요용어: 비모수자기회귀모형, 회귀분석, 효율성

이 논문은 2015학년도 건국대학교의 연구년교원 지원에 의하여 연구되었음.

[1](05029) 서울 광진구 능동로 120, 건국대학교 응용통계학과. E-mail: kyusangu@konkuk.ac.kr