자기회귀계수에 대한 소표본 점근추론

나종화[1]　김정숙[2]　장영미[3]

요 약

본 논문에서는 1차 자기회귀모형에서 자기회귀계수에 대한 여러 가지 추정량들의 분포함수에 대한 근사적추론 방법에 대해 연구하였다. 이차형식에 대한 안장점근사의 결과를 이용한 이 근사법은 여러 형태의 추정량들에 대해 근사분포의 유도과정이 불필요하며, 소표본은 물론 통계적 추론의 주요 관심영역에서의 근사정도가 매우 뛰어난 장점을 가지고 있다. 모의실험을 통해 Edgeworth근사를 비롯한 기존의 여러 근사법보다 효율이 뛰어남을 확인하였다.

주요용어 : 자기회귀계수, Edgeworth 근사, 안장점근사, 이차형식

1. 서론

다음과 같이 정의되는 1차 자기회귀모형(first-order autoregressive model)을 생각하자.

$$X_t = \alpha X_{t-1} + e_t, \quad t = 1, \cdots, n\ . \tag{1.1}$$

여기서 모든 t 에 대해 $|\alpha| < 1$이고, e_t 는 평균 0, 분산 σ^2이며 서로독립인 정규분포를 따른다. 그리고 X_0는 평균이 0이고 분산이 $\sigma^2/(1-\alpha^2)$ 인 정규분포를 가정한다. 식 (1.1)의 자기회귀계수 α에 대한 최소제곱추정량은 다음과 같다.

$$\hat{\alpha} = \frac{\sum_{t=2}^{n} X_t X_{t-1}}{\sum_{t=1}^{n} X_t^2}. \tag{1.2}$$

자기회귀계수 α에 대한 추론은 $\hat{\alpha}$의 점근분포가, 충분히 큰 n에 대해, 근사적으로 평균이 α이고, 분산이 $(1-\alpha)^2/n$인 정규분포를 따른다는 사실에 기초하고 있다. 그러나 정규분포에 기초한 추론은 매우 큰 표본의 크기를 요구하고 있어, 실제 중· 소표본의 자료에 대해 큰 제한으로 작용하고 있다. 이를 개선하기 위한 많은 연구가 진행되어 왔으며, 특히 Sargan (1976)과 Phillips (1977, 1978)는 $\hat{\alpha}$의 분포함수에 대한 Edgeworth전개를 통한 근사법을 제안하여, 정규근사의 결과를 부분적으로 개선하였으나 자기회귀계수의 값이 커짐에 따라 근사의 정확도가 여전히 만족스럽지 못한 결과를 보여주고 있다. Daniels (1956)는 다음의 변형된 추정량

$$\hat{\alpha}^* = \frac{X_1X_2 + \cdots + X_{n-1}X_n}{\frac{1}{2}X_1^2 + X_2^2 + \cdots + X_{n-1}^2 + \frac{1}{2}X_n^2}. \tag{1.3}$$

의 밀도함수(density function)에 대한 효과적인 근사로 안장점근사(saddlepoint approximation)를 적용한 바 있다. 이 근사는 대단히 정확한 결과를 제공하며, 식 (1.2)의 최소제곱추정량에도

1) (361-763) 청주시 흥덕구 개신동 충북대학교 정보통계학과, 교수

2) (121-749) 서울시 마포구 염리동 168-9 건강보험심사평가원 평가정보부, 책임연구원

3) (361-763) 청주시 흥덕구 개신동 충북대학교 정보통계학과, 박사과정

적용이 가능하다. 그러나 자기회귀계수의 추론에 필요한 꼬리확률에 대한 결과를 얻기 위해서는 밀도함수에 대한 근사식을 수치적분 해야 하는 어려움이 존재하기 때문에 그다지 효율적인 방법이라 말할 수 없다.

안장점근사는 특정 통계량의 밀도함수나 분포함수의 근사에 매우 효과적인 방법으로 알려져 있다. 특히 안장점근사의 장점은 소표본의 경우에서도 그 정확도가 매우 뛰어난 것으로 알려져 있어, 정밀한 추론이 요구되는 많은 응용문제에 효과적인 대안으로 활용되고 있다. 본 논문에서는, 최근 Kuonen (1999)과 Na & Kim (2005)에 의한 이차형식에 대한 안장점근사의 결과를 이용하여, 자기회귀계수의 여러 추정량들의 분포함수에 대한 안장점근사를 수행하고 자 한다. 2절에서는 이차형식에 대한 안장점근사의 결과를 소개하고, 3절에서는 자기회귀계수 추정량들의 분포함수에 대한 안장점근사의 적용과정을 다루었으며, 4절에서는 모의실험을 통해 이 근사법의 정확도 및 기존의 방법과의 비교분석을 수행한다.

2. 이차형식에 대한 안장점근사

$X=(X_1,\cdots,X_n)'$를 평균이 $\mu=(\mu_1,\cdots,\mu_n)'$이고 공분산행렬이 Σ인 다변량 정규분포를 따르는 확률벡터라 하자. 이차형식 $Q(X)$는 다음과 같이 정의된다.

$$Q(X)=\sum_{1\leq i\leq j\leq n} a_{ij}X_iX_j=X'AX. \tag{2.1}$$

위 식에서 $A=(a_{ij})_{n\times n}$ 는 일반성을 잃지 않고 대칭(symmetric)으로 가정할 수 있다. 이차형식 $Q(X)$의 누율생성함수를 $K(t)$라 할 때, Lugannani & Rice (1980)의 분포함수에 대한 안장점근사식은 다음과 같다. (Daniels, 1987)

$$\Pr\{Q(X)\leq q\}\simeq\begin{cases}\Phi(w)+\phi(w)\left\{\dfrac{1}{w}-\dfrac{1}{\zeta}\right\}, & q\neq E(Q(X))\\[2ex] \dfrac{1}{2}-\dfrac{K^{(3)}(0)}{6\sqrt{2\pi(K''(0))^3}}, & q=E(Q(X))\end{cases} \tag{2.2}$$

위식에서 w와 ζ는 다음과 같이 정의된다.

$$w=sgn(t_0)\,[2n\{t_0q-K(t_0)\}]^{1/2},\quad \zeta=t_0\{nK''(t_0)\}^{1/2}.$$

여기서 안장점(saddlepoint)으로 불리는 t_0는 다음의 안장점방정식(saddlepoint equation)에 대한 수치해를 통해 구해진다.

$$K'(t_0)=q$$

또한 $\Phi(\cdot)$와 $\phi(\cdot)$는 표준정규분포의 분포함수와 밀도함수를 나타낸다.

다음으로 Jensen (1992, 1995)의 근사식은 다음과 같다.

$$\Pr\{Q(X)\leq q\}\simeq\Phi\left\{w+\frac{1}{w}log\left(\frac{\zeta}{w}\right)\right\}. \tag{2.3}$$

위 근사식에 사용되는 이차형식 $Q(X)$의 누율생성함수와 관련된 성질은 다음의 정리와 보조정리를 이용하여 구할 수 있다. 먼저 $\Sigma=\Gamma\Gamma'$, $B=\Gamma A\Gamma$이라 하고, 행렬 B의 고유치들을 $\lambda_i(i=1,\cdots,n)$라 하자.

정리 2.1 이차형식 $Q(X)=X'AX$의 누율생성함수는 다음과 같이 주어진다.

$$\begin{aligned}K_Q(t)&=-\frac{1}{2}\sum_{i=1}^{n}\log(1-2t\lambda_i)+\frac{1}{2}\sum_{i=1}^{n}\frac{\delta_i^2\lambda_i t}{(1-2t\lambda_i)}\\&=-\frac{1}{2}\sum_{i=1}^{n}\log(1-2t\lambda_i)+\frac{1}{2}\sum_{i=1}^{n}\frac{\delta_i^2}{(1-2t\lambda_i)}-\frac{1}{2}\sum_{i=1}^{n}\delta_i^2.\end{aligned}$$

여기서 $\delta = P \cdot \Gamma^{-1} \cdot \mu$으로 정의된다.
증명 Na와 Kim (2005)을 참고할 것.
따름정리 2.1 이차형식 $Q(X)$의 1, 2차 미분식은 다음과 같이 주어진다.

$$K_Q'(t) = \frac{d}{dt}K_Q(t) = \sum_{i=1}^{n}\frac{\delta_i^2\lambda_i}{(1-2t\lambda_i)^2} + \sum_{i=1}^{n}\frac{\lambda_i}{1-2t\lambda_i}.$$

$$K_Q''(t) = \frac{d^2}{dt^2}K_Q(t) = \sum_{i=1}^{n}\frac{4\delta_i^2\lambda_i^2}{(1-2t\lambda_i)^3} + \sum_{i=1}^{n}\frac{2\lambda_i^2}{(1-2t\lambda_i)^2}.$$

증명 정리2.1로부터 쉽게 구할 수 있다.

3. 자기회귀계수에의 응용

식 (1.1)로 부터 $X=(X_1,\cdots,X_n)'$는 평균이 $\mu=(0,\cdots,0)'$이고 공분산행렬이

$$\Sigma = \frac{\sigma^2}{1-\alpha^2}\begin{pmatrix} 1 & \alpha & \alpha^2 & \cdots & \alpha^{n-1} \\ \alpha & 1 & \alpha & \cdots & \alpha^{n-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \alpha^{n-1} & \cdots & \cdots & \cdots & 1 \end{pmatrix} \tag{3.1}$$

인 다변량 정규분포를 따른다. 또한 본 논문의 주된 관심사인 식 (1.2)의 통계량의 분포함수는

$$\Pr(\hat{\alpha} \leqq q) = \Pr(V_n \leqq 0), \tag{3.2}$$

$$V_n = \sum_{t=2}^{n} X_t X_{t-1} - q\sum_{t=1}^{n} X_t^2.$$

이며, 통계량 V_n은 다음과 같이 이차형식의 형태의 통계량으로 재 표현 될 수 있다.

$$V_n = (X_1,\cdots,X_n)\begin{pmatrix} -q & 1/2 & 0 & 0 & 0 & \cdots & 0 \\ 1/2 & -q & 1/2 & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{2} & -q & 1/2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1/2 & -q & 1/2 \\ 0 & 0 & 0 & \cdots & 0 & 1/2 & -q \end{pmatrix}\begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \tag{3.3}$$

$$= X'AX.$$

이제 식 (3.2)에 대해 앞 절에서 소개한 이차형식에 대한 안장점근사의 결과를 적용할 수 있다. 이 과정은 식 (1.3) 및 다음과 같이 표현되는 Durbin (1980)의 추정량

$$\hat{\alpha}' = \frac{\frac{1}{2}X_1^2 + X_1X_2 + \cdots + X_{n-1}X_n + \frac{1}{2}X_n^2}{X_1^2 + \cdots + X_n^2}. \tag{3.4}$$

에 대해서도 동일하게 적용될 수 있다. 예를 들어, 식 (1.3)의 변형된 추정량 $\hat{\alpha}^*$의 분포함수는

$$\Pr(\hat{\alpha}^* \leqq q) = \Pr(V_n^* \leqq 0).$$

$$V_n^* = \sum_{t=2}^{n} X_t X_{t-1} - y\left(\frac{1}{2}X_1^2 + X_2^2 + \cdots + X_{n-1}^2 + \frac{1}{2}X_n^2\right).$$

이며, V_n^*는 다시 다음의 이차형식

$$V_n^* = (X_1, \cdots, X_n)\begin{pmatrix} -q/2 & 1/2 & 0 & 0 & 0 & \cdots & 0 \\ 1/2 & -q & 1/2 & 0 & 0 & \cdots & 0 \\ 0 & 1/2 & -q & 1/2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1/2 & -q & 1/2 \\ 0 & 0 & 0 & \cdots & 0 & 1/2 & -q/2 \end{pmatrix}\begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$$

$$= X'BX.$$

으로 표현되어 안장점근사의 결과를 쉽게 적용할 수 있게 된다. 동일한 방법을 식 (3.4)의 통계량에 적용하면, 대응하는 이차형식은 다음과 같이 주어진다.

$$V_n' = (X_1, \cdots, X_n)\begin{pmatrix} 1/2-q & 1/2 & 0 & 0 & 0 & \cdots & 0 \\ 1/2 & -q & 1/2 & 0 & 0 & \cdots & 0 \\ 0 & 1/2 & -q & 1/2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1/2 & -q & 1/2 \\ 0 & 0 & 0 & \cdots & 0 & \frac{1}{2} & \frac{1}{2}-q \end{pmatrix}\begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$$

$$= X'CX.$$

이 과정에서 알 수 있듯이 자기회귀계수에 대한 안장점근사의 적용은 다양한 추정량의 형태에 따라 이차형식에 나타나는 행렬의 형태만 약간의 변형이 일어나며, 동일한 방법으로 근사될 수 있음을 알 수 있다.

4. 모의실험

앞 절에서 소개한 여러 가지 유형의 추정량 가운데, 가장 보편적으로 사용되는 식 (1.2)의 최소제곱추정량에 대한 안장점근사의 결과를 제시하고, 기존의 여러 근사법과의 비교를 수행한다. 모의실험은 식 (3.1)에서 $\mu=0,\ \sigma^2=1$이고, 자기회귀계수가 $\rho=0.2, 0.8$이며, 표본의 크기 $n=5, 30$ 인 경우에 대해 수행하였다. 모의실험의 결과에서 Exact는 정확한 값으로 10만 번의 모의실험을 통해 구해진 값이며, Normal은 정규근사로 아래 Edgeworth근사식(Phillips, 1978)의 첫째항을 이용한 것이며, Edge1과 Edge2는 아래의 근사식에서 $O(n^{-1/2})$와 $O(n^{-1})$항까지의 근사를 나타낸다.

$$\Pr\{\sqrt{n}(\hat{\alpha}-\alpha)/(1-\alpha^2)^{1/2} \le x\} = \Phi(x) + \phi(x)[n^{-1/2}\alpha(1-\alpha^2)^{-1/2}(x^2+1) + \frac{1}{4n}(1-\alpha^2)^{-1}\{(1-\alpha^2)x + (1+\alpha^2)x^3 - 2\alpha^2x^5\}] + O(n^{-3/2}).$$

또한 Saddle1과 Saddle2는 각각 식 (2.3)과 (2.2)의 안장점근사의 결과를 나타낸다.

아래의 <표4.1>에서 알 수 있듯이 안장점근사(Saddle1, Saddle2)의 정도는 매우 뛰어난 결과를 제공하는데 비해, 기존의 정규근사 및 Edgeworth근사는 통계적 추론의 관심 영역인 꼬리부분에서 그 정도가 매우 떨어지는 것을 확인할 수 있다. 안장점근사의 경우에는 α의 값이 큰 경우에도 매우 정확한 근사값을 제공할 뿐 이니라, 매우 작은 소표본($n=5$)의 경우에도 기존의 근사법에 비해 상당히 정확한 값을 제공하고 있음을 알 수 있다.

5. 결론

본 논문에서는 자기회귀계수의 여러 추정량들을 이차형식의 구조로 이해하고, 이차형식의 분포함수에 대한 안장점근사의 결과를 이들 추정량에 적용한 결과 매우 정확한 근사결과를 얻을 수

있었다. 특히, 자기회귀계수의 값이 비교적 큰 값을 가지는 경우, 기존의 여러 근사법들은 매우 불안정하며 대표본하에서도 매우 부정확한 결과를 제공하는데 비해, 안장점근사는 소표본의 경우에도 매우 뛰어난 근사를 제공함을 확인하였다. 따라서 본 논문에서 제시한 추정량들에 대한 효과적인 근사를 통해 자기회귀계수에 대한 매우 정확한 근사적 추론이 가능하다.

<표4.1> 자기회귀계수의 분포함수 근사($n=5, \alpha=0.2$)

		$\Pr\{\hat{\alpha} \le q\}$에 대한 근사				
q	Exact	Normal	Edge1	Edge2	Saddle1	Saddle2
-1.9	0.0133	0.0287	0.0563	0.0325	0.0147	0.0147
-1.5	0.0535	0.0668	0.1052	0.0759	0.0556	0.0556
-1.1	0.1365	0.1357	0.1796	0.1534	0.1379	0.1379
-0.7	0.2620	0.2420	0.2844	0.2679	0.2683	0.2683
-0.3	0.4358	0.3821	0.4200	0.4138	0.4384	0.4384
0.1	0.6176	0.5398	0.5764	0.5784	0.6260	0.6261
0.3	0.7110	0.6179	0.6559	0.6621	0.7160	0.7161
0.5	0.7956	0.6915	0.7316	0.7428	0.7974	0.7975
0.7	0.8736	0.7580	0.8005	0.8170	0.8667	0.8668
1.1	0.9660	0.8643	0.9083	0.9345	0.9633	0.9633
1.5	0.9999	0.9332	0.9716	1.0009	0.9999	0.9999

참고문헌

Daniels, H. E. (1956). The approximate distribution of serial correlation coefficients, Biometrika, 43, 169-185.

Daniels, H. E. (1987). Tail probability approximations, International Statistical Review, 55(1), 37-48.

Durbin, J. (1980). The approximate distribution of partial serial correlation coefficients calculated from residuals from regression on Fourier series, Biometrika, 67(2), 335-349.

Jensen, J. L. (1992). The modified signed loglikelihood statistic and saddlepoint approximations, Biometrika, 79, 693-703.

Jensen, J. L. (1995). Saddlepoint Approximations, Oxford.

Kuonen, D. (1999). Saddlepoint approximations for distributions of quadratic forms in normal variables, Biometrika, 86(4), 929-935.

Lugannani, R. and Rice, S. O. (1980). Saddlepoint approximation for the distribution of the sum of independent random variables, Advanced Applied Probability, 12, 475-490.

Na, J. H. and Kim, J. S. (2005). Saddlepoint approximations to the distribution functions of non-homogeneous quadratic forms, The Korean Journal of Applied Statistics, 18(1), 183-196.

Phillips, P. C. B. (1977). A general theorem in the theory of asymptotic expansions as approximations to the finite sample distributions of econometric estimators, Econometrica, 45, 1517-1534.

Phillips, P. C. B. (1978). Edgeworth and saddlepoint approximations in the first-order noncircular autoregression, Biometrika, 65(1), 91-98.

Sargan, J. D. (1976). Economic estimators and the Edgeworth approximation. Econometrica, 44, 421-428.