응용통계연구 제 **19**권 2호, *2006*년, *pp. 283–290*

이동-멱변환에 관한 연구*

조기종[1] 정석오[2] 신기일[3]

요 약

일반적으로 Box-Cox 변환과 같은 류의 멱변환은 분산 안정화 혹은 분포의 대칭성 향상 등을 목적으로 사용된다. 그러나 원 자료의 평균의 크기가 크면서 분산이 상대적으로 작은 경우, 즉 변동계수가 작은 경우에는 제대로 작동하지 않는 것이 알려져 있다. 본 논문에서는 이러한 문제점을 해결하기 위한 이동-멱변환을 제안하고 모의실험과 실제 자료 분석을 통하여 그 효과를 확인하였다.

주요용어: 멱변환, 이동-멱변환, Box-Cox 변환, 이동모수

1. 서론

자료에 적절한 변환을 취하여 변환된 자료의 분포가 가능한 한 정규분포에 가깝도록 만들고자 할 때 가장 널리 사용되는 Box-Cox 변환(Box and Cox, 1964)은 다음과 같다:

$$\Psi_{\mathrm{BC}}(x;\lambda) = \begin{cases} (x^{\lambda}-1)/\lambda, & (x \geq 0, \lambda \neq 0) \\ \log x, & (x > 0, \lambda = 0) \end{cases} \tag{1.1}$$

멱변환모수 λ의 값은 보통, 관측된 원래 자료값을 $\mathbf{x} = \{x_1, x_2, \cdots, x_n\}$로 나타낼 때

$$\ell_{\mathrm{BC}}(\lambda|\mathbf{x}) = -\frac{n}{2}\log s_y^2 + (\lambda-1)\sum_{i=1}^{n}\log x_i \tag{1.2}$$

를 최대로 하는 λ의 값으로 선택한다. 단, $i = 1, 2, \cdots, n$에 대하여

$$y_i = \Psi_{\mathrm{BC}}(x_i;\lambda), \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2$$

* 이 논문은 한국과학재단 목적기초연구지원(조기종)과 2006년도 한국외국어대학교 교내학술연구비의 지원(정석오, 신기일)에 의하여 이루어진 것임.

1) (136-701) 서울 성북구 안암동 고려대학교 환경생태학부, 교수
E-mail: kjcho@korea.ac.kr
2) (449-791) 경기도 용인시 처인구 모현면 왕산리 산 89, 한국외국어대학교 정보통계학과, 조교수
E-mail: seokohj@hufs.ac.kr
3) (교신저자) (449-791) 경기도 용인시 처인구 모현면 왕산리 산 89, 한국외국어대학교 정보통계학과, 교수
E-mail: keyshin@hufs.ac.kr

이다.

그러나 Bickel and Doksum(1981)에서 지적된 바와 같이, 원 자료의 분산이 평균에 비해 매우 작은 경우, 거꾸로 말해 평균이 분산에 비해 현격하게 큰 경우에 Box-Cox 변환이 전혀 작동하지 않는 경우를 종종 만나게 된다. μ_x를 x_i들의 기대값이라 하면 μ_x가 충분히 큰 경우

$$\begin{cases} x_i^{\lambda} \approx \mu_x^{\lambda} + \lambda\mu_x^{\lambda-1}(x_i-\mu_x), & \lambda \neq 0 \text{ 일 때} \\ \log x_i \approx \log\mu_x + \mu_x^{-1}(x_i-\mu_x), & \lambda = 0 \text{ 일 때} \end{cases} \quad (i=1,2,\cdots,n)$$

의 근사식이 성립한다. 이를 이용하면

$$\frac{y_i - \mathrm{E}(y_i)}{\sqrt{\mathrm{Var}(y_i)}} \approx \frac{x_i - \mathrm{E}(x_i)}{\sqrt{\mathrm{Var}(x_i)}},\ i=1,2,\cdots,n \tag{1.3}$$

가 됨을 간단히 보일 수 있는데, 이는 위에서 언급한대로 자료의 평균이 분산에 비해 매우 큰 경우 Box-Cox 변환이 거의 무의미해짐을 알려준다. 간단한 모의실험을 통해 이러한 예를 쉽게 만들어 보일 수 있다. 그림 1.1의 (a)는 평균이 1인 지수분포에서 100개의 난수를 생성한 후 Box-Cox 변환을 실시한 결과로서 Box-Cox 변환이 성공적으로 작동했음을 알 수 있다. (b), (c), (d)는 원래 생성된 자료에 각각 3, 5, 10을 더한 값에 Box-Cox 변환을 실시한 결과인데, 평균 대 분산의 비가 커지면서 Box-Cox 변환을 실시하더라도 그 분포의 형태가 오히려 원래 자료의 분포의 형태와 점점 비슷해져서 정규분포와 가까워지지 않는 것을 확인할 수 있다. 참고로 각 경우 λ의 값은 위 식 (1.2)를 최대로 하는 값으로 선택하였다. 따라서 분산에 비해 평균이 큰 자료에 대해 Box-Cox류의 멱변환을 적용할 때에는 각별한 주의와 함께 특별한 조처가 필요하다.

Atkinson, Pericchi and Smith(1991)는 이를 위해 이동-멱변환, 즉 자료에 적절한 이동모수를 뺀 후 멱변환을 취하는 방법에 관하여 심도있게 연구하였다. 이들은 식 (1.2)의 우도함수에 이동모수를 포함시켜 이를 최적화하는 방법으로 이동모수와 멱변환모수를 동시에 추정하는 방법을 제안하였는데, 이러한 방식으로 정의한 보통의 우도함수는 그 최적해가 존재하지 않는 등의 심각한 문제가 있어 그룹화된 우도함수를 이용하는 접근법을 사용하였다. 그러나 이를 위해서는 이동모수, 멱변환모수 외에 그룹화를 위한 또다른 모수(Δ)가 추가적으로 요구되며, 매우 주의하여 Δ를 선택해야 하는 점이 지적되었다(Atkinson, Pericchi and Smith, 1991, p. 479).

본 논문에서는 위에서 지적된 어려움을 피하기 위해 2-단계 과정을 거치는 이동-멱변환 방법을 제안한다. 이를 간단히 요약하면 다음과 같다. 첫번째 단계에서 적절한 이동모수의 값을 결정하여 원 자료에서 빼어 준다. 두번째 단계에서 이동된 자료를 이용해 식 (1.2)를 최대화하는 멱변환모수 λ를 결정한다. 제안된 방법의 장점은 다음과 같다. 첫째, 간단하고 이해하기 쉽다. 둘째, 계산량이 작고 수치해석적 문제가 발생하지 않는다. 세째, 이동모수의 값을 구할 때 필요한 평활모수의 선택이 변환 결과에 미치는 영향이 미미하다.

관련 연구로는, 변환모수의 대표본 성질을 연구한 Hernandez and Johnson (1981), 실수 전체에서 값을 취하는 자료의 멱변환에 관한 연구인 Yeo and Johnson (2000), 분산에 비해 평균이 큰 시계열 자료를 분석할 경우 멱변환 효과에 관한 연구인 Shin and Kang(2001), Park and Shin(2006) 등을 들 수 있다.

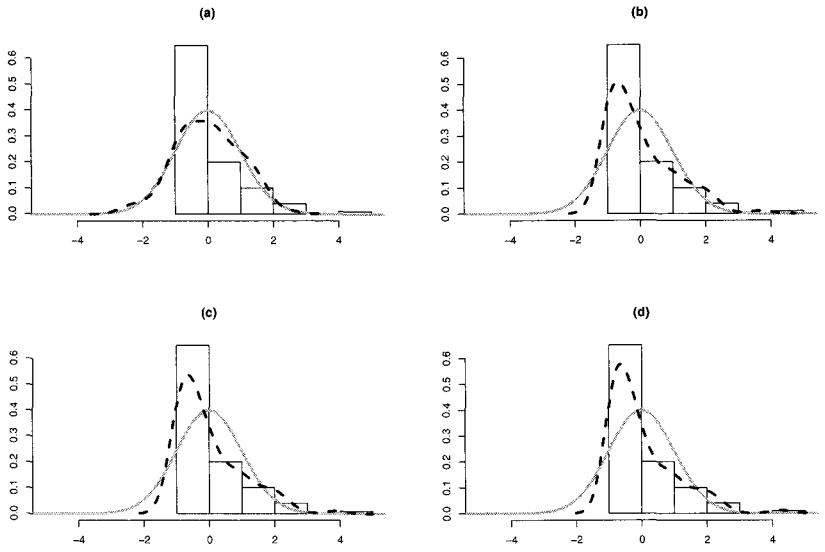

그림 1.1: 평균 변화에 따른 멱변환의 효과 : 회색실선은 표준정규분포, 히스토그램은 원래 자료를 표준화한 $(x_i - \bar{x})/s_x$의 분포, 굵은 점선은 Box-Cox 변환한 자료를 표준화한 $(y_i - \bar{y})/s_y$의 커널밀도함수추정치를 나타낸 것이다. 단 $\bar{x}$와 s_x는 원래 자료의 표본평균과 표본표준편차이고, $\bar{y}$와 s_y는 Box-Cox 변환을 취한 후의 표본평균과 표본표준편차이다.

이 논문은 다음과 같이 구성되었다. 다음 절에서는 본 논문이 제안하는 이동-멱변환에 대해 구체적으로 소개하고, 그 이동시키는 양 즉 이동모수의 값을 결정하는 방법에 대해 논의한다. 3절에서 제안된 방법의 효과를 모의실험과 실제 자료 분석을 통해 보인 후, 4절에서 결론을 맺는다.

2. 이동모수

그림 1.1의 모의실험 결과는 이동모수를 도입하는 아이디어에 대한 간단한 힌트를 제공한다. 원래 멱변환의 효과가 좋았던 자료에 단순히 상수를 더하였다고 해서 변환 결과가 좋지 않다는 것은, 거꾸로 말해 멱변환 결과가 좋지 않은 자료의 경우 적절한 크기의 상수를 빼주면 그 효과가 개선될 가능성이 높을 것이라는 기대를 갖게 한다. 따라서 이 논문에서는 주어진 자료 $\mathbf{x} = \{x_1, \cdots, x_n\}$에 대해 적당한 양의 상수 γ를 뺀 후 Box-Cox 변환을 적용하는 다음과 같은 형태의 변환을 제안한다.

$$\Psi(x; \gamma, \lambda) = \begin{cases} \{(x-\gamma)^\lambda - 1\}/\lambda, & (x - \gamma \geq 0, \lambda \neq 0) \\ \log(x - \gamma), & (x - \gamma > 0, \lambda = 0) \end{cases} \tag{2.1}$$

이 때 양수 γ의 값은 모든 $i = 1, \cdots, n$에 대해 $x_i - \gamma$의 값이 음이 되지 않도록 즉 자료의 최소값을 넘지 않도록 정하기만 하면 위의 변환은 잘 정의된다. 다만 (1.3)와 같은 상황을 피

하려면 이동 후의 평균값이 가능한 한 작아야 하므로 γ는 x_i들의 분포 받침(support)의 왼쪽 끝값, 즉 $f(x)$를 x_i들의 밀도함수라 할 때

$$\gamma_0 = \inf\{x > 0 \mid f(x) > 0\} \tag{2.2}$$

로 정하는 것이 가장 좋을 것으로 기대된다. 물론 이 값이 잘 알려져 있는 경우에는 문제가 없지만 그렇지 않은 경우에는 이 값을 추정해 사용해야 한다. 이 때 가장 쉬운 방법은 자료의 최소값

$$\hat{\gamma}_0 = x_{(1)}$$

을 사용하는 것이다. 단 $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$은 자료의 순서통계량을 뜻한다.

그러나 이 방법은 이동 후 자료의 값에 하나 이상의 0을 항상 포함하게 되어 식 (2.1)가 정의되지 않는 경우를 발생시킬 뿐 아니라, 태생적으로 $\hat{\gamma}_0$의 값이 언제나 식 (2.2)의 γ_0의 값보다 큰 값을 가지게 되는 편의(bias)를 피할 수 없다. 이 때 발생하는 편의는

$$0 < \mathrm{Bias}(\hat{\gamma}_0, \gamma_0) \equiv \mathrm{E}(\hat{\gamma}_0 - \gamma_0) \approx \{nf(\gamma_0)\}^{-1}$$

의 근사식을 만족함을 이용하면, 다음과 같은 잘 알려진 방법으로 간단히 추정할 수 있다(Bloch and Gastwirth, 1968):

$$\widehat{\mathrm{Bias}}(\hat{\gamma}_0, \gamma_0) = m^{-1}\{x_{(m+1)} - x_{(1)}\}. \tag{2.3}$$

단, m은 n이 증가할 때 $m \to \infty$, $m/n \to 0$을 만족하는 적당한 양의 수열이다. 이상을 종합하여 편의를 수정한

$$\tilde{\gamma}_0 = x_{(1)} - m^{-1}\{x_{(m+1)} - x_{(1)}\} \tag{2.4}$$

을 식 (2.1)의 이동모수 자리에 대입하면 된다.

3. 모의실험과 예제

이 절에서는 모의실험과 실제 자료 분석을 통해 제안된 방법의 효과를 조사한다. 모의실험을 위한 모형은 다음과 같다.

[모형 1] $X_0 \sim \mathrm{Exp}(1)$, $X = X_0 + M$, $M = 3, 5, 10$.

[모형 2] $X_0 \sim \mathrm{Gamma}(2, 1/2)$, $X = X_0 + M$, $M = 5, 10, 20$.

두 모형 모두 X_0가 평균이 1인 분포를 따르지만, 모형 1은 왼쪽 받침이 끝이 점프가 있는 분포의 형태이고 모형 2는 그렇지 않다는 점에 유의하기 바란다. M은 인위적으로 만든 이동량이다. 각 모형에 대해 500회의 몬테칼로 모의실험을 실시하였고, 각 실험은 크기가 $n = 20$인 난수 X를 발생시켜 Box-Cox 변환 및 제안하는 변환을 적용하여 정규성 검정을 위한 Shapiro-Wilk 검정통계량 및 유의확률을 계산하는 과정으로 이루어졌다. 500회 반복실험에서 얻은 유의확률의 평균을 표 3.1에 정리하였다. 두 모형에서 모두 일관되게 Box-Cox

변환보다 이동모수를 고려한 변환을 시행했을 때 정규분포에 훨씬 더 가까워질 뿐 아니라 m의 선택에 거의 영향을 받지 않았다. 그림 3.1는 500회 반복 실험 중 한 실험의 결과를 도시한 것이다. 이 역시 제안된 방법의 Box-Cox 변환에 대한 우수성과 함께, m의 값에 따른 변환 후 자료의 분포의 변화가 거의 없음을 잘 보여주고 있다.

표 3.1: Shapiro-Wilk 검정의 유의확률

			m						
	M	Box-Cox	1	2	3	4	5	6	7
[모형 1]	+3	0.0855	0.6776	0.6785	0.6742	0.6705	0.6659	0.6610	0.6556
	+5	0.0721	0.6689	0.6696	0.6680	0.6632	0.6591	0.6535	0.6494
	+10	0.0511	0.6636	0.6706	0.6673	0.6619	0.6575	0.6535	0.6465
[모형 2]	+5	0.2075	0.7033	0.7087	0.7090	0.7064	0.7043	0.7029	0.7008
	+10	0.1554	0.7016	0.7062	0.7060	0.7041	0.7024	0.7003	0.6990
	+20	0.1563	0.6726	0.6783	0.6782	0.6757	0.6741	0.6719	0.6705

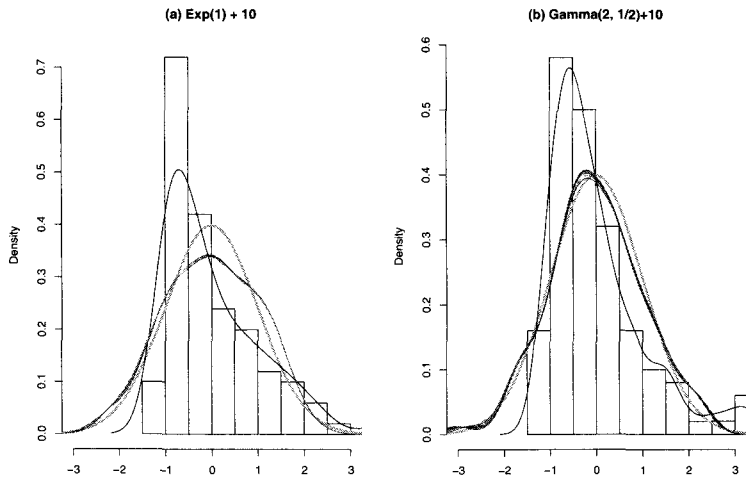

그림 3.1: 이동-멱변환의 효과 : 회색 굵은 실선은 표준정규분포곡선, 히스토그램은 원래 자료를 표준화한 것의 분포, 검은 실선은 Box-Cox 변환한 자료를 표준화한 것의 커널밀도함수추정치, 다른 색깔의 실선들은 다양한 m값($m = 1, 2, \cdots, 10$)을 사용해 얻은 제안된 방법에 의한 변환을 적용한 것을 표준화한 것의 커널밀도함수추정치이다. (a) [모형 1] $M = 10$. (b) [모형 2] $M = 10$.

다음은 1930년 미국 49개 대도시의 인구 자료(Cochran (1977)의 p.152, Table 6.1)에 적용한 결과이다. 그림 3.2와 표 3.2를 살펴보면 이동모수를 고려하지 않은 Box-Cox 변환은 전혀 작동하지 않고 있음을 확인할 수 있다. 반면에 제안된 변환 방법은 m의 값에 관계없이 정규분포에 잘 적합된 결과를 보이고 있다.

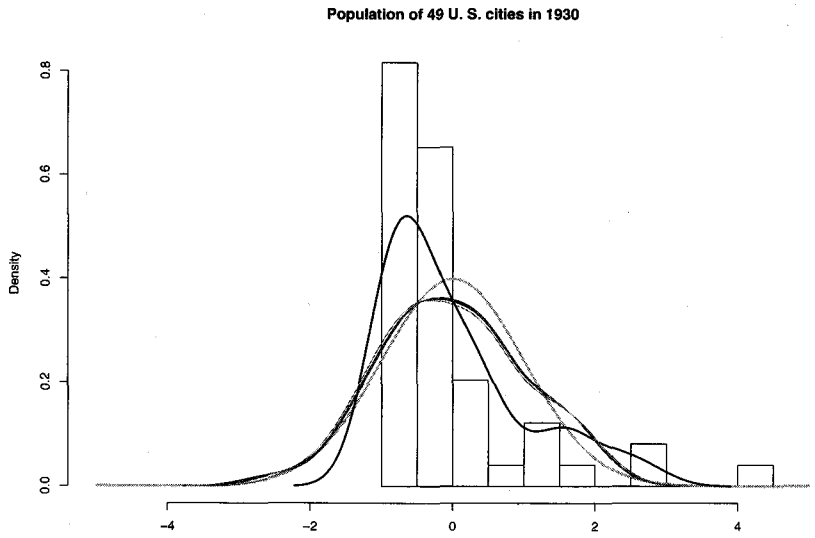

그림 3.2: 1930년 미국 49개 도시 인구 자료: 회색 굵은 실선은 표준정규분포곡선, 히스토그램은 원래 자료를 표준화한 것의 분포, 검은 실선은 Box-Cox 변환한 자료를 표준화한 것의 커널밀도함수추정치, 다른 색깔의 실선들은 다양한 m값($m = 1, 2, \cdots, 10$)을 사용해 얻은 제안된 방법에 의한 변환을 적용한 것을 표준화한 것의 커널밀도함수추정치이다.

표 3.2: 1930년 미국 49개 도시 인구 자료: Shapiro-Wilk 정규성 검정 결과

		m								
	Box-Cox	1	2	3	4	5	6	7	8	9
검정 통계량	0.8629	0.9855	0.9855	0.9883	0.9887	0.9886	0.9886	0.9887	0.9887	0.9885
(유의확률)	4.1404e-5	0.8025	0.8025	0.9032	0.9174	0.9127	0.9144	0.9174	0.9174	0.9115

4. 결론

분포의 대칭성 향상 혹은 분산 안정화 등을 위해 널리 사용되고 있는 Box-Cox 변환은 자료의 분포의 평균이 분산에 비해 큰 경우 잘 작동하지 않는 단점이 있다. 이에 대한 대안으로 이동모수를 고려하여 이를 수정한 후 Box-Cox 변환을 적용하는 것을 제안하였으며, 이동모수를 자료에 의존하여 결정하는 방법을 제시하였다. 또한 제안된 방법의 효과를 모의실험과 실제 자료 분석을 통해 성공적으로 실증하였다.

참고문헌

Atkinson, A. C., Pericchi, L. R. and Smith, R. L. (1991). Grouped likelihood for the shifted power transformation, *Journal of the Royal Statistical Society B* **53**, 473–482.

Bickel, P. J. and Doksum, K. A. (1981). An analysis of transformations revisited, *Journal of the American Statistical Association* **76**, 296–311.

Bloch, D. A. and Gastwirth, J. L. (1968). On a simple estimate of the reciprocal of the density function, *Annals of Mathematical Statistics* **39**, 1083–1085.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations (with Discussion), *Journal of the Royal Statistical Society B* **26**, 211–252.

Cochran, W. G. (1977), *Sampling Techniques.* 3rd edition, Wiley.

Hernandez, F. and Johnson, R. A. (1980). The large-sample behavior of transformations to normality, *Journal of the American Statistical Association* **75**, 855–861.

Park, H. and Shin, K.-I. (2006). A shrinked forecast in stationary processes favouring percentage error, *Journal of Time Series Analysis* **27**, 129–139.

Shin, K.-I. and Kang, H. (2001). A study of the effect of power transformation in the ARMA(p,q) model, *Journal of Applied Statistics* **28**, 1019–1028.

Yeo, I.-K. and Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry, *Biometrika* **87**, 954–959.

[2005년 12월 접수, 2006년 3월 채택]

Shift-Power Transformation*

Kijong Cho[1] Seok-Oh Jeong[2] Key-Il Shin[3]

ABSTRACT

Generally speaking, power transformations such as Box-Cox transformation(1964) is applied for variance stabilization and symmetry. But, when the distribution of the original data has a large mean with a small variance or the coefficient of variation is very small, they don't work at all. This paper propose a simple method to introduce a shift parameter before applying power transformations and showed the numerical evidence by Monte Carlo simulation and a real data analysis.

Keywords: Power transformation; shifted power transformation; Box-Cox transformation; shift parameter.

* This work was partially supported by grant No. R01-2003-000-1024-0 from the Basic Research program of the Korea Science and Engineering Foundation to K. Cho and the research fund of Hankuk University of Foreign Studies, 2006 to S.-O. Jeong and K.-I. Shin.

1) Professor, Division of Environmental Science and Ecological Engineering, Korea University. E-mail: kjcho@korea.ac.kr

2) Assistant professor, Department of Statistics, Hankuk University of Foreign Studies. E-mail: seokohj@hufs.ac.kr

3) (Corresponding author) Professor, Department of Statistics, Hankuk University of Foreign Studies. E-mail: keyshin@hufs.ac.kr