응용통계연구 (2012)
25(1), 205–213
DOI: http://dx.doi.org/10.5351/KJAS.2012.25.1.205

회귀모형에서 이상치 검색을 이용한 로버스트 변수변환방법

서한손[1] · 이가연[2] · 윤민[3]

[1]건국대학교 응용통계학과, [2]오케이캐시백 서비스 전략기획팀, [3]부경대학교 통계학과

(2011년 10월 접수, 2011년 11월 수정, 2011년 11월 채택)

요 약

선형회귀모형에서 자료를 모형에 적합시킬 때 일반적으로 반응변수 변환을 시도하지만 적절한 변환함수의 결정은 몇 개의 이상치들에 민감하게 반응한다는 것이 잘 알려져 있다. 이에 따라 이상치에 영향을 받지 않는 변수변환 방법들이 연구, 개발되고 있으나 최근에 Cheng (2005)에 의해 최소절사제곱추정치에 기반을 둔 절사 우도추정치 방법처럼 이상치의 숫자를 미리 정해야한다거나 많은 계산량이 필요하다는 단점들을 갖고 있다. 본 논문에서는 그와 같은 문제점을 해결하고 추정치의 강건성을 개선하는 새로운 방법을 제안하며 제안된 방법에서는 반응변수 변환에 따른 이상치 탐색법에 있어서 Hadi와 Simonoff (1993)가 제시한 단계적 절차를 응용, 적용한다.

주요용어: 박스-콕스 변환, 변수변환, 이상치, 최소절사제곱추정량, 회귀모형.

1. 서론

통계분석의 대표적 도구로 사용되는 선형회귀모형은 다음과 같이 표현된다.

$$Y = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.1}$$

여기서 Y는 반응변수이며, $\boldsymbol{X}$는 $n \times p$행렬, $\boldsymbol{\beta}$는 $p \times 1$벡터, 그리고 $\boldsymbol{\varepsilon} \sim N(0, \sigma^2)$이다. 식 (1.1)과 같은 회귀식에서 모수를 추정하는 방법에는 여러 가지가 있으나 그 중 최소제곱추정법(least squared estimator; LSE)이 일반적으로 사용된다. 최소제곱추정법은 많은 장점과 아울러 이상치(outlier) 또는 영향값에 의하여 잘못된 모형을 추정할 수 있다는 취약점이 있다. 이에 따라 선형모형을 이용한 회귀분석(linear regression model)에서 다중이상치 식별을 위해 여러가지 방법들이 제시되었다. 예를 들면 Gentleman과 Wilk (1975)의 정상 데이터군 전수 조사방법, Marasinghe (1985)의 다단계 절차법(multistage method), Kianifard와 Swallow (1989)의 반복잔차(recursive residauls) 활용법, Paul과 Fung (1991)의 일반화된 극단치 스튜던트화 잔차(generalized extreme studentized residual; GESR) 활용법 등이 있으며 Hadi와 Simonoff (1993)는 순차적 탐색법을 제시하여 앞서 예를 든 방법들 보다 효율성이 더 높다는 것을 보였다. 식 (1.1)에서 정의된 선형회귀식에서 반응변수 변환을 허용하면 좀더 응용력이 확장된 식 (1.2)와 같은 회귀식을 정의할 수 있다.

$$Y(\lambda) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{1.2}$$

식 (1.2)의 반응변수의 변환에 관련된 대표적 방법은 다음과 같은 변환함수로 정의되는 Box-Cox 변환

[3]교신저자: (608-737) 부산시 남구 대연3동 599-1, 부경대학교 통계학과, 조교수. E-mail: myoon@pknu.ac.kr

(Box와 Cox, 1964)이다.

$$Y(\lambda) = \begin{cases} \dfrac{(Y^{\lambda}-1)}{\lambda}, & \lambda \neq 0, \\ \log Y, & \lambda = 0, \end{cases} \tag{1.3}$$

변환계수 λ의 추정 통계량은 최대우도추정량(Maximum Likelihood Estimator; MLE)을 사용하며 정의상 최대우도추정량은 이상치에 민감하다. 따라서 모형 (1.2)에서 λ를 추정할 때 이상치에 강건한 추정방법을 개발하는 것이 필요하다. 모형 (1.2)에서 λ의 추정에 관련하여 이상치를 탐지하는 기본 원리는 모든 관찰치를 사용하여 추정된 λ와 특정 관찰치 또는 관찰치군이 빠졌을 때 추정되는 λ의 차이를 보고 이상치의 여부를 판단하는 것이다. 그러나 이러한 원리를 적용할 때 어려운 점은 특정 관찰치 또는 관찰치군을 제외하고 λ의 최대우도추정량를 추정할 경우 각 추정치마다 쟈코비안이 달라져서 막대한 계산량이 요구된다는 점이다. 이에 대한 대안으로 Cook과 Wang (1983)은 관측값 제거 모델(case-deletion model)을 활용한 추정법을 제안하였고 Hinkley와 Wang (1988)은 일반화 선형모델(GLM)을 고려하여 Cook과 Wang의 추정량을 수정한 추정량을 제안하였으며 Tsai와 Wu (1990)는 관측값 제거 모델이 쟈코비안에 끼치는 영향력에 바탕을 둔 방법을 제안하였다.

본 논문에서는 식 (1.2)에서 λ를 추정할 때 이상치를 탐지하고 이것을 제외한 추정량을 계산하는 방법을 제안한다. 이러한 의도에서 기존에 제안된 방법 중 대표적인 것은 Cheng (2005)이 제안한 최소절사제곱추정량(Least Trimmed Squares estimator; LTS) 또는 최대절사우도추정량(Maximum Trimmed Likelihood Estimator; MTLE)을 이용한 방법이다. 본 논문에서는 순차적 검색법을 적용한 로버스트 추정법을 제시한다. 본 논문의 2장에서는 최소절사제곱추정량과 최대절사우도추정량의 관계를 설명하고 최소절사제곱추정량를 사용한 Cheng (2005)의 추정방법을 소개한다. 제 3장에서는 본 연구에서 제안하는 순차적 검색법을 통한 이상치 탐지와 변환계수 추정법을 설명한다. 제 4장에서는 시뮬레이션과 예제를 통해 Cheng의 방법과 새로 제안한 방법을 비교하고 5장에서는 연구의 결과를 요약, 정리한다.

2. 최소절사제곱추정량을 사용한 접근법

최대우도추정법에서 이상치는 낮은 우도에 해당하는 관찰치로 이해할 수 있으므로 이상치의 영향력을 배제하기 위해 최대절사우도법 (Hadi와 Luceno, 1997)은 낮은 우도의 관찰치를 제외한 우도함수를 설정하여 추정량을 계산한다. 모형 (1.2)에서 식 (1.3)과 같이 변환된 반응변수의 변환계수 λ에 대한 최대절사우도추정량의 목표함수는 식 (2.1)로 표시된다.

$$L_q\left(\hat{\beta}_q\right) = \sum_{i \in M} l\left(\hat{\beta}_q; y_{(i)}(\lambda)\right), \tag{2.1}$$

여기서 q는 우도함수에 포함되는 관찰치의 크기이며 M은 변환한 $y(\lambda)$를 이용하여 계산된 우도에서 크기가 가장 큰 q개의 관측치군이고 $\hat{\beta}_q$는 변환계수가 λ일 때 β의 최대절사우도추정량이며 $l(\hat{\beta}_q; y_{(i)}(\lambda))$는 n개 중 i번째로 큰 우도이다. 식 (2.1)은 다음과 같이 표현될 수 있다.

$$L_q\left(\hat{\beta}_q\right) = -\frac{q}{2}\log\hat{\sigma}_q^2(\lambda) = -\frac{q}{2}\log\sum_{i \in M}\frac{e_i^2(\lambda)}{(q-p)}, \tag{2.2}$$

이때 $e_i(\lambda) = y_i(\lambda) - x_i^T\hat{\beta}_q$, $i = 1, \ldots, n$이다. 따라서 변환계수 λ의 최대절사우도통계량은 변환계수가 λ일 때 σ^2의 최대절사우도추정량인 식 (2.3)을 최소화하는 추정치와 일치한다.

$$\hat{\sigma}_q^2(\lambda) = \sum_{i \in M}\frac{e_i^2(\lambda)}{(q-p)}. \tag{2.3}$$

이 때 M은 잔차가 가장 작은 q개의 관측치군이며 이것은 우도가 가장 큰 q개의 관측치군과 일치한다. λ가 고정된 상태에서 절사된 관찰치에 대한 변환변수의 쟈코비안은 β와 무관하므로 식 (2.1)를 최대화하는 것은 식 (2.3)을 최소화하는 것과 일치하여 최대절사우도량은 최소절사제곱추정량과 동일하게 된다.

이와 같은 관점에서 Cheng은 Box-Cox 변환의 변환모수 λ를 추정할 때 최소절사제곱추정법을 이용할 것을 제안하였다. Cheng이 제안한 방법은 최소절사제곱추정량을 구하기 위해 다수의 부표집(subsampling) 과정이 필요하며 k번째 부표집 과정은 다음과 같이 설명된다.

정상 데이터군의 크기는 q라고 사전에 정했다고 하자. $k-1$번째 부표집에서 추정된 변환계수 추정값을 $\hat{\lambda}_{k-1}$라고 하고 이때 계산된 최소절사제곱추정량을 $S^2(\hat{\lambda}_{k-1})$라고 할 때 $\hat{\lambda}_{k-1}$로 반응변수를 변환한 데이터에서 k번째 부표집으로 $s=p+1$개의 관찰치를 랜덤하게 뽑는다. 이때 뽑혀진 관찰치군을 $\mathfrak{T}_k$라고 표시하면 $\mathfrak{T}_k$만을 이용하여 최소제곱법에 의해 회귀계수 $\hat{\beta}_{(\mathfrak{T})}$를 추정하고, 이 회귀식에 의해 n개의 전체 데이터에 대한 잔차 $e_{i,\mathfrak{T}_k}$, $i=1,\ldots,n$를 계산한다. 잔차 $e_{i,\mathfrak{T}_k}$의 크기 순서에 따라 나열된 데이터를 가지고 Rousseeuw와 Driessen (2006)이 제안한 C-step을 적용하여 최소절사제곱추정량의 목적함수가 가장 작은 q개의 관찰치로 정상 데이터군 M_k를 만든다. 이렇게 생성된 정상 데이터군 M_k에 의해 반응변수 변환계수의 잠정적인 추정값 $\hat{\lambda}_k^{TP}$를 구한다. $\hat{\lambda}_k^{TP}$를 이용하여 반응변수를 변환시킨 후 정상 데이터 M_k만을 가지고 β의 최소제곱추정량인 $\hat{\beta}_{M_k}$와 잔차 $e_i(\hat{\lambda}_k^{TP})=z_i(\hat{\lambda}_k^{TP})-x_i^T\hat{\beta}_{M_k}$를 계산하고 $\hat{\lambda}_k^{TP}$에 따른 절사분산추정량 $S^2(\hat{\lambda}_k^{TP})=\sum_{i\in M}e_i^2(\hat{\lambda}_k^{TP})/(q-p)$을 계산한다. 만약 $S^2(\hat{\lambda}_k^{TP})<S^2(\hat{\lambda}_{k-1})$이면 $\hat{\lambda}_k=\hat{\lambda}_k^{TP}$으로 $S^2(\hat{\lambda}_k^{TP})\geq S^2(\hat{\lambda}_{k-1})$이면 $\hat{\lambda}_k=\hat{\lambda}_{k-1}$으로 결정하고 새로운 $k+1$번째 표본을 추출하고 위의 과정을 반복한다.

이와 같은 부표집 과정을 반복 수행한 후 가장 작은 절사분산추정량($S^2(\hat{\lambda})$)을 갖는 추정량이 λ의 최소절사제곱추정량의 근사적인 해가 된다. 전체 표본의 숫자가 작을 경우 가능한 모든 부표집의 경우를 전부 시도할 수 있으나 큰 데이터의 경우 적절한 횟수의 부표집을 수행한다. 부표집 과정에서 최초로 지정하는 λ의 추정치, 즉, $\hat{\lambda}_0$는 모든 관찰치에 의한 최대우도추정치 $\hat{\lambda}_{mle}$를 사용한다. Cheng의 방법에서 정상 데이터군의 크기 q는 주로 전체 데이터의 개수에 비례하여 지정된다.

3. 이상치 검정을 통한 제안한 로버스트 변수변환 방법

2장에서 설명한 Cheng의 방법은 최소절사제곱추정량을 구하기 위해 고려되어야 할 모든 가능한 정상 데이터군 M의 개수는 비정상 데이터군의 크기 $n-q$의 크기에 따라 ${}_nC_{n-q}$개가 되어 이를 확인하기 위하여 계산량은 상당히 늘어나게 되며 정상 데이터의 크기도 임의로 결정해야 하는 문제도 발생하게 된다. 이러한 점을 개선하고 보다 로버스트한 변환변수 추정을 위해 다단계 방법에 의한 이상치 검색법을 사용한 새로운 추정법을 제안한다. 새롭게 제안되는 방법은 비정상 데이터군의 크기를 미리 정할 필요가 없으며 적은 계산량을 통하여 정상 데이터를 찾아낼 수 있다.

모형 (1.2)에서 이상치를 고려하여 변환변수 모수 λ를 추정하는 문제에서 일단 λ값이 고정된다면 이상치 문제는 모형 (1.1)에서의 상황과 동일하게 되므로 모형 (1.1)에서 개발된 이상치 탐지방법을 활용할 수 있다. 새롭게 제안하는 변환변수 추정법에서는 λ에 대하여 잠정적으로 고정값을 가정하여 전진적으로 이상치를 탐색하고 이에 따라 λ를 추정해 나가는 과정을 채택한다. 이때 고정된 λ값 아래 이상치 탐색을 위해 Hadi와 Simonoff (1993)가 제안한 순차적 이상치 탐색법을 사용한다.

Hadi와 Simonoff (1993)는 기초적인 정상 데이터군을 선별한 후 이를 기반으로 나머지 관찰치에 대하여 이상치 여부를 검정해 나가는 순차적 검색법을 제안하였다. 그 과정을 요약하면 다음과 같다. 크기 $s=\text{int}[(n+p-1)/2]$인 정상 데이터의 기초군 M을 생성한 후 부분 데이터 M을 모형에 적용시켜 다

음과 같은 잔차(internally studentized residual) d_i를 계산한다.

$$d_i = \begin{cases} \dfrac{\left(y_i - x_i^T \hat{\beta}_M\right)}{\hat{\sigma}_M \sqrt{1 - x_i^T \left(X_M^T X_M\right)^{-1} x_i}}, & \text{if } i \in M, \\ \dfrac{\left(y_i - x_i^T \hat{\beta}_M\right)}{\hat{\sigma}_M \sqrt{1 + x_i^T \left(X_M^T X_M\right)^{-1} x_i}}, & \text{if } i \notin M, \end{cases}$$

여기서 $|d|_{(j)}$를 $|d_i|$의 크기순 j번째 순서통계량이라고 할 때, $|d|_{(s+1)} \geq t_{(\alpha/2, s-k)}$이면 마지막 $(n-s)$순서통계량에 해당하는 관찰치를 모두 이상치라고 판단하고 검정을 마친다. 만약 $|d|_{(s+1)} < t_{(\alpha/2, s-k)}$이면 $|d|_{(s+1)}$에 해당하는 관찰치를 정상 데이터군에 포함시킨 후 과정을 반복하고 $n = s + 1$인 경우에는 이상치에 해당하는 관찰치가 없다고 판정한다. 정상 데이터의 기초군을 생성하는 방법으로 Hadi와 Simonoff는 두가지 방법을 제안하였으며 이중에서 보다 더 효율적인 것은 LS 추정치를 이용한 점진적 선택법이다. 이 방법은 우선 모든 관찰치를 회귀모형에 적용시킨 후 잔차가 가장 작은 p개의 관찰치를 선택한다. 이와 같이 선택된 크기 p의 데이터에 모형을 적용시킨 후 $|d_i|$를 계산하여 가장 작은 $(p+1)$개의 관찰치를 새로운 기초군으로 형성한다. 정상 데이터의 기초군의 개수가 int$[(n+p-1)/2]$가 될 때 까지 이와 같은 과정을 반복하여 최종적인 정상 데이터 기초군을 선정한다.

본 연구에서 제안하는 변수변환모수에 대한 로버스트한 추정법은 Cheng이 제안한 방법에서 최소절사제곱추정법 대신 Hadi와 Simonoff의 이상치 탐색방법을 적용하여 이상치를 발견, 제외하고 이때 계산되는 절사제곱추정량을 최적추정치의 기준으로 간주하여 λ에 대한 최소절사제곱추정량을 찾는다.

새로운 로버스트 변수변환 방법의 절차를 요약하면 다음과 같다.

Step 0 현재 단계에서의 변환변수 추정값을 $\hat{\lambda}_{pr}$이라고 하고 이에 대한 절사제곱추정량을 $S^2(\hat{\lambda}_{pr})$이라고 하자.

Step 1 $\hat{\lambda}_{pr}$로 반응변수를 변환한다.

Step 2 앞에서 설명한 과정에 따라 이상치 검정을 위한 기초군을 구성하고 이상치 검정을 수행하여 정상 데이터군 M을 만든다. M의 크기를 r이라고 하자.

Step 3 정상 데이터군 M만을 가지고 반응변수 변환계수의 잠정적인 추정값 $\hat{\lambda}^{TP}$를 구한다.

Step 4 $\hat{\lambda}^{TP}$에 따라 반응변수를 변환시킨 후 정상 데이터 M만을 가지고 β의 최소제곱추정량(LSE)인 $\hat{\beta}_M$을 구하고 이에 따른 절사제곱추정량(TSE) $S^2(\hat{\lambda}^{TP})$를 다음과 같이 계산한다

$$S^2\left(\hat{\lambda}^{TP}\right) = \sum_{i \in M} \frac{e_i^2\left(\hat{\lambda}^{TP}\right)}{(r-p)}, \tag{3.1}$$

여기서 $e_i(\hat{\lambda}^{TP}) = z_i(\hat{\lambda}^{TP}) - x_i^T \hat{\beta}_M$이다. 만약 $S^2(\hat{\lambda}^{TP}) \geq S^2(\hat{\lambda}_{pr})$이면 $\hat{\lambda}_{pr}$을 λ의 최종 추정값으로 결정하고 절차를 끝낸다. $S^2(\hat{\lambda}^{TP}) < S^2(\hat{\lambda}_{pr})$이면 $\hat{\lambda}^{TP}$를 새로운 $\hat{\lambda}_{pr}$값으로 대체한 후 Step 1에서부터 위 과정을 반복한다.

최초로 지정하는 λ의 추정치 $\hat{\lambda}_0$는 Cheng의 알고리즘에서와 마찬가지로 모든 관찰치를 전부 사용하여 추정된 MLE 추정치를 사용하며 이에 해당하는 절사제곱추정량은 $S^2(\hat{\lambda}_0) = \sum_{i \in M} e_i^2(\hat{\lambda}_0)/(r-p)$이 된다. Cheng의 알고리즘은 최소절사제곱추정량에 의해서 정상 데이터군을 구하기 때문에 원칙적으로 샘플로 뽑혀질 가능한 모든 케이스를 모두 고려해야 하지만 제안한 알고리즘은 $S^2(\hat{\lambda}^{TP}) \geq S^2(\hat{\lambda}_{pr})$할 때까지만 과정을 반복하면 된다.

표 **4.1.** $p = 3$일 때 시뮬레이션 결과

표본 크기	이상치 비율(%)	λ							
		-1		-0.5		0		0.5	
		Cheng	new	Cheng	new	Cheng	new	Cheng	new
50	5	−0.9036	−0.9765	−0.4069	−0.4558	−0.0045	0.00004	0.4670	0.4996
		(0.0562)	(0.0737)	(0.0434)	(0.0750)	(0.0169)	(0.0008)	(0.0397)	(0.0029)
	10	−0.7624	−0.9805	−0.3425	−0.4369	−0.0239	−0.0011	0.3715	0.4992
		(0.0838)	(0.0756)	(0.0463)	(0.0930)	(0.0244)	(0.0096)	(0.0717)	(0.0072)
	15	−0.6405	−0.8450	−0.2913	−0.3763	−0.0449	−0.0214	0.2639	0.3934
		(0.1170)	(0.2095)	(0.0520)	(0.1189)	(0.0268)	(0.0423)	(0.0932)	(0.1713)
	20	−0.5784	−0.7041	−0.2559	−0.3991	−0.0773	−0.0126	0.1875	0.4495
		(0.1295)	(0.2275)	(0.0498)	(0.1127)	(0.0348)	(0.0309)	(0.1232)	(0.1425)
100	5	−0.8834	−0.9715	0.4517	0.4998	−0.0055	−8e-06	0.4551	0.4999
		(0.0441)	(0.0669)	(0.0309)	(0.0013)	(0.0133)	(0.0002)	(0.0325)	(0.0012)
	10	−0.7809	−0.9732	−0.3508	−0.4225	−0.0044	−0.0008	0.3787	0.4976
		(0.0546)	(0.0669)	(0.0309)	(0.0726)	(0.0088)	(0.0068)	(0.0447)	(0.0234)
	15	−0.6955	−0.8926	−0.3131	−0.3700	−0.0442	−0.0131	0.3019	0.4905
		(0.0744)	(0.1584)	(0.0325)	(0.0933)	(0.0221)	(0.0263)	(0.0555)	(0.0447)
	20	−0.5972	−0.7692	−0.2743	−0.3125	−0.0692	−0.0197	0.1949	0.4612
		(0.0824)	(0.1865)	(0.0355)	(0.0865)	(0.0196)	(0.0306)	(0.0806)	(0.1067)

4. 모의실험 및 예제

4.1. 모의실험

제안한 로버스트 변수변환 방법의 효율성을 Cheng의 방법과 비교하기 위하여 모의실험을 실시한다. 가상의 데이터는 Rousseeuw (1984)와 Cheng (2005)에서 사용한 방식과 유사하게 생성된다. 전체 데이터에서 비정상 데이터에 해당하는 관찰치 $(Y_i, X_i^T)^T$는 다변량 정규분포

$$\text{MVN}\left(\begin{pmatrix} 4 \\ (14+p)\mathbf{J}_{p-1} \end{pmatrix}, \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5\mathbf{I}_{p-1} \end{pmatrix}\right),$$

$\mathbf{J}_{p-1}$는 $(p-1)$차원인 단위벡터(unit vector)에 따라 발생시킨다. 정상 데이터에 해당하는 관찰치는 각각의 독립변수 X_i^T를 균등분포 $U(0,6)$에 따라 발생시키고 오차항 ε_i를 정규분포 $N(0, 0.04^2)$에 따라 생성한후 반응변수 Y는 주어진 변환변수 λ에 따라 Box-Cox 변환, $Y^{(\lambda)} = XI_p + \varepsilon$의 역함수를 적용하여 생성한다.

변수변환 방법의 효율성을 측정하기 위하여 200개의 데이터 셋을 생성하고 각 데이터에서 추정된 변환모수 추정값 $\hat{\lambda}$의 평균과 표준편차를 계산한다. Cheng의 방법에서 시행되는 부표집의 반복 횟수는 10,000번으로 한다. 모의실험 데이터의 중요한 모수값은 다음과 같이 다양하게 지정한다. 변환모수 λ는 $-1, -0.5, 0, 0.5$로, 데이터 차수 p는 $3, 5$로, 샘플크기 n은 $50, 100$으로, 이상치의 비율은 5%, 10%, 15%, 20% 등으로 지정한다. Cheng의 방법에서 정상 데이터군의 예상 개수 q는 이상치의 비율에 따라 다르게 할당하며 이상치의 비율이 5%와 10%, 15%일 때 $q = [0.75n]$으로, 20%일 때 $q = [0.70n]$으로 할당한다.

200번의 모의실험결과 추정된 변환모수 $\hat{\lambda}$의 평균과 표준편차(괄호부분)가 표 4.1과 표 4.2에 정리하였다. 표 4.2에서 알 수 있듯이 이상치의 비율(%), 샘플 크기(n), 데이터 차원(p)에 따라 추정방법의 효율성이 달라지지만 Cheng의 방법과 제안한 방법을 비교해 볼 때, 이러한 요인에 상관없이 제안한 방법이

표 **4.2.** $p = 5$일 때 시뮬레이션 결과

표본 크기	이상치 비율(%)	λ							
		-1		-0.5		0		0.5	
		Cheng	new	Cheng	new	Cheng	new	Cheng	new
50	5	−0.9250	−1.0009	−0.4695	−0.4998	−0.0002	−3.6e-0.5	0.4996	0.5001
		(0.0886)	(0.0147)	(0.0342)	(0.0015)	(0.0300)	(0.0004)	(0.0727)	(0.0026)
	10	−0.6157	−0.8533	−0.3068	−0.4251	−0.0258	−0.0084	0.2738	0.4225
		(0.0273)	(0.6272)	(0.0611)	(0.1230)	(0.0147)	(0.0195)	(0.0470)	(0.1556)
	15	−0.3214	−0.4136	−0.1974	−0.3006	−0.0685	−0.0510	0.0494	0.1564
		(0.0388)	(0.3958)	(0.0757)	(0.1984)	(0.0234)	(0.0393)	(0.0626)	(0.2707)
	20	−0.1779	−0.3031	−0.1535	−0.3355	−0.0894	−0.0709	−0.0807	0.0109
		(0.1694)	(0.4294)	(0.0770)	(0.1913)	(0.0234)	(0.0447)	(0.1171)	(0.2984)
100	5	−0.8910	−0.9997	−0.4981	−0.4999	0.0078	0.0012	0.4547	0.50003
		(0.0676)	(0.0018)	(0.0050)	(0.0008)	(0.0091)	(0.0206)	(0.0430)	(0.0011)
	10	−0.6032	−0.9105	−0.3228	−0.4553	−0.0298	−0.0055	0.2582	0.4445
		(0.0955)	(0.1930)	(0.0401)	(0.0881)	(0.0159)	(0.0160)	(0.0670)	(0.1251)
	15	−0.3641	−0.5197	−0.2298	−0.4251	−0.0597	−0.0326	0.0806	0.3823
		(0.1236)	(0.3473)	(0.0493)	(0.1424)	(0.0169)	(0.0314)	(0.0862)	(0.2241)
	20	−0.1859	−0.2904	−0.1582	−0.3504	−0.0632	−0.0411	−0.0781	0.3898
		(0.1394)	(0.3789)	(0.0578)	(0.1797)	(0.0364)	(0.0128)	(0.0922)	(0.2015)

표 **4.3.** 트리 데이터의 추정 결과

방법			$\hat{\lambda}_q$			이상치
new			0.3066			없음
Cheng	q	$0.85n$	평균	표준편차	최빈값	5, 6, 7, 14, 18
			0.3231	0.0735	0.3296	
		$0.95n$	평균	표준편차	최빈값	5, 14
			0.2915	0.0430	0.2968	

Cheng의 방법보다 변환모수 λ의 실제값에 보다 근접하게 추정하여 효율성이 높다는 것을 확인할 수 있다.

4.2. 예제 1: 미니탭 트리 데이터(Minitab tree data)

회귀변환(Regression transformation)에 관한 연구에서 많이 다루어진 미니탭 트리 데이터는 체리나무의 둘레와 높이에 따른 부피예측을 위한 31개의 데이터로 구성되어 있다. Atkinson (1985)은 변수변환에 대한 스코어검정을 통해 반응변수에 대한 변수변환이 필요하며 이상치로 판단되는 관찰치는 없다고 분석하였다. 이에 따라 31개의 관측값 전부를 이용한 반응변수에 대한 변환계수 λ의 추정치(MLE)는 0.3066으로 계산된다. Cheng의 방법을 이용하여 1000번의 부표집 과정을 통해 λ를 추정한 결과와 제안한 방법에 의한 λ의 추정결과는 표 4.3과 같다. 제안한 방법은 Atkinson이 언급한 것과 동일하게 이상치가 없으며 $\hat{\lambda} = 0.3066$으로 추정하였다. Cheng의 방법을 적용하면 q값에 따라 λ가 다르게 추정되었으며 다수의 관찰치가 이상치로 판정되었다.

4.3. 예제 2: 별 데이터(Hertzsprung-Russell star data)

별 데이터 (Rousseeuw와 Leroy, 1987, p.27)는 47개 항성에서 뿜어내는 빛의 세기와 표면 온도를 측정

표 **4.4.** 별 데이터의 추정 결과

방법			$\hat{\lambda}_q$	이상치
new			1.1168	7, 11, 20, 30, 34
Cheng	q	$0.75n$	1.7573	5, 7, 9, 11, 14, 18, 20, 23, 30, 34, 40, 47
		$0.85n$	1.3472	7, 11, 14, 20, 30, 34, 36, 39
		$0.90n$	1.1168	7, 11, 20, 30, 34

표 **4.5.** 등반경주 데이터의 추정 결과

방법			$\hat{\lambda}_q$			이상치
new			0.9144			7, 18
Cheng	q	$0.75n$	평균	표준편차	최빈값	5, 7, 11, 16, 17, 18, 31, 35
			0.7746	0.1719	0.9072	
		$0.85n$	평균	표준편차	최빈값	6, 7, 10, 11, 14, 18
			0.8076	0.1435	0.86	
		$0.90n$	평균	표준편차	최빈값	7, 11, 18, 35
			0.7682	0.1662	0.88	
		$0.95n$	평균	표준편차	최빈값	7, 11
			0.7073	0.1487	0.5968	

한 것이다. 대개의 관측치와는 달리 11, 20, 30, 34번째 관측치들은 표면온도가 낮음에도 불구하고 빛의 세기가 강한 것으로 조사되어 이상치로 간주되며 7번째 관측치는 경계선에 있다고 볼 수 있다. 제안한 방법에 의해 이상치를 제거하고 변환모수값을 추정했을 때 변환모수는 1.1168로 추정되었고 이때 이상치는 7, 11, 20, 30 그리고 34번째 관측치로 확인되었다. 변환모수 추정값이 1에 가까운 값이어서 이상치가 제거된 뒤 추정되는 모형은 선형식이 유력하다는 것을 암시하고 있다. 별 데이터는 변수의 수와 관찰치의 개수가 크지 않으므로 가능한 모든 조합의 부표집을 수행하여 Cheng의 방법을 적용하였다. 그 결과 q가 $0.90n$일 때 변환모수 추정값은 1.1168이고, 7, 11, 20, 30, 34번째 관측치를 이상치로 판정하여 제안한 방법과 동일한 결과를 보였으나 q가 이보다 작을 때는 정상적인 관찰치를 이상치로 판단하였고 추정된 λ도 1과 차이가 났다.

4.4. 예제 3: 언덕경주 데이터(Hill racing data)

언덕경주 데이터에는 2개의 독립변수인 경주거리(distances in miles)와 고도(climbs in feet)및 반응변수인 경주기록(record time)이 35개 포함되어있다. Atkinson (1986, 1988)의 분석에 의하면 데이터를 변환시킬 필요가 없고 7, 18, 33번째 관찰치들은 이상치로 간주된다. 35개 관찰치 전부를 이용한 최대우도변환모수 추정값은 $\hat{\lambda} = 0.512$이며 이를 초기치로 사용하여 Cheng의 방법과 제안한 방법을 언덕경주 데이터에 적용한 결과는 표 4.5와 같다. 새로운 방법에 의해 추정된 변환모수 값은 $\hat{\lambda} = 0.9144$로 1에 근사하였으며 이때의 이상치는 7번째와 18번째 케이스로 판단된다. Cheng의 방법에 따라 1,000번의 부표집 과정을 반복하여 변환모수를 추정한 결과 q가 $0.75n$이었을 때 추정값에 대한 최빈값이 0.9072로 다른 경우보다 1에 근사한 값으로 추정되었다.

5. 결론

본 논문에서는 이상치를 포함한 데이터에 대하여 이상치를 검출하여 로버스트한 변수변환을 이룰 수 있는 이론적인 배경을 설명하고, 기존의 방법보다 개선된 방법을 제안하고 있다. 제안한 방법으로 로버스

트 변수변환을 시도할 때 기존의 방법과 비교하여 다음과 같은 장점들을 가진다. 첫째 계산이 간편하다. Cheng의 방법에 따라 변수변환을 시도하기 위해서는 최소절사제곱추정량을 구하기 위하여 과도한 부표집 과정이 필요한데 반하여 제안한 방법은 검정을 통해 상대적으로 작은 계산량으로 작업을 수행할 수 있다. 둘째 정상 데이터군을 구하는데 있어 Cheng의 방법은 그 크기를 사전에 예상하여 지정해 주어야 하지만 제안한 방법은 단계적 방법에 의해 이상치에 대한 검정을 실시하게 되므로 미리 정상 데이터군의 크기를 지정할 필요가 없다. 마지막으로 시뮬레이션을 통한 두 방법의 비교 결과에서도 보여줬듯이, 제안한 방법을 통한 반응변수 변환모수 λ에 대한 추정이 Cheng의 방법에 의한 것보다 뛰어나다.

본 연구에서는 λ를 추정할 때 최우추정량을 사용하였지만 Cook과 Wang (1983) 또는 Tsia와 Wu (1990) 등이 제시한 추정량을 사용할 수도 있다. 제안한 로버스트 변수변환 방법이 전진형 탐색방법이므로 정상 데이터의 크기가 변할 때 마다 스코어 통계량을 표시한 팬 플랏(fan plot)을 그려봄으로써 이상치 제거 후 반응변수 변환모수의 추정범위를 시각적으로 예측하는 것도 시도해 볼 수 있다.

참고문헌

Atkinson, A. C. (1985). *Plots, Transformations and Regression: An Introduction to Graphical Method of Diagnostic Regression Analysis*, Oxford University Press, Oxford.

Atkinson, A. C. (1986). Aspects of diagnostic regression analysis (discussion of influential observations, high leverage points, and outliers in linear regression), *Statistical Science*, **1**, 397–402.

Atkinson, A. C. (1988). Transformations unmasked, *Technometrics*, **30**, 311–318.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations (with discussion), *Journal of the Royal Statistical Society. Series B (Methodological)*, **26**, 211–246.

Cheng, T.-C. (2005). Robust regression diagnostics with data transformations, *Computational Statistics & Data Analysis*, **49**, 875–891.

Cook, R. D. and Wang, P. C. (1983). Transformations and influential cases in regression, *Technometrics*, **25**, 337–343.

Gentleman, J. F. and Wilk, M. B. (1975). Detecting outliers. Ⅱ. Supplementing The direct analysis of residuals, *Biometrics*, **31**, 387–410.

Hadi, A. S. and Luceno, A. (1997). Maximum trimmed likelihood estimators: A unified approach, examples, and algorithms, *Computational Statistics & Data Analysis*, **25**, 251–272.

Hadi, A. S. and Simonoff, J. S. (1993). Procedures for the identification of multiple outliers in linear models, *Journal of the American Statistical Association*, **88**, 1264–1272.

Hinkley, D. V. and Wang, S. (1988). More about transformations and influential cases in regression, *Technometrics*, **30**, 435–440.

Kianifard, F. and Swallow, W. H. (1989). Using recursive residuals, calculated on adaptively-ordered observations, to identify outliers in linear regression, *Biometrics*, **45**, 571–585.

Marasinghe, M. G. (1985). A multistage procedure for detecting several outliers in linear regression, *Technometrics*, **27**, 395–399.

Paul, S. R. and Fung, K. Y. (1991). A generalized extreme studentized residual multiple-outlier-detection procedure in linear regression, *Technometrics*, **33**, 339–348.

Rousseeuw, P. J. (1984). Least median of squares regression, *Journal of the American Statistical Association*, **79**, 871–880.

Rousseeuw, P. J. and Driessen, K. V. (2006). Computing LTS regression for large data sets, *Data Mining and Knowledge Discovery*, **12**, 29–45.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*, John Wiley, New York.

Tsai, C. L. and Wu, X. (1990). Diagnostics in transformation and weighted regression, *Technometrics*, **32**, 315–322.

Robust Response Transformation Using Outlier Detection in Regression Model

Han Son Seo[1] · Ga Yoen Lee[2] · Min Yoon[3]

[1]Department of Applied Statistics, Konkuk University
[2]Strategy & Planning Team, Okcashbag Service
[3]Department of Statistics, Pukyong National University

(Received October 2011; Revise November 2011; Accepted November 2011)

Abstract

Transforming response variable is a general tool to adapt data to a linear regression model. However, it is well known that response transformations in linear regression are very sensitive to one or a few outliers. Many methods have been suggested to develop transformations that will not be influenced by potential outliers. Recently Cheng (2005) suggested to using a trimmed likelihood estimator based on the idea of the least trimmed squares estimator(LTS). However, the method requires presetting the number of outliers and needs many computations. A new method is proposed, that can solve the problems addressed and improve the robustness of the estimates. The method uses a stepwise procedure, suggested by Hadi and Simonoff (1993), to detect outliers that determine response transformations.

Keywords: Box-Cox transformation, variable transformation, outlier, least trimmed squares estimator, regression model.

[3]Corresponding author: Assistant Professor, Department of Statistics, Pukyong National University, 599-1 Daeyeon 3-Dong, Nam-Gu, Busan 608-737, Korea. E-mail: myoon@pknu.ac.kr