Journal of the Korean Data & Information Science Society
2013, **24**(1), 33–39

http://dx.doi.org/10.7465/jkdi.2013.24.1.33
한국데이터정보과학회지

부분선형모형에서 반응변수변환을 위한 회귀진단[†]

서한손[1] · 윤민[2]

[1]건국대학교 응용통계학과 · [2]부경대학교 통계학과

접수 2012년 11월 7일, 수정 2012년 12월 2일, 게재확정 2012년 12월 7일

요 약

반응변수의 변환을 고려하는 부분선형모형에서 이상치 문제는 선형모형에서와 마찬가지로 반응변수 변환모수의 추정에 왜곡된 결과를 초래할 수 있다. 이를 해결하기 위해서는 부분선형모형에서 반응변수 변환 모수 추정과 이상치 탐지 과정이 수행되어야 하지만 모형에 포함된 비모수 함수의 비정형성에 따른 어려움이 크다. 본 연구에서는 부분선형모형의 비모수함수에 대한 추정과 순차적 검정, 최대절사우도추정 등과 같은 이상치 제거방법의 적용을 통하여 부분선형모형에서 이상치에 강건한 반응변수 변환 과정을 제안한다. 제안된 방법들은 모의실험과 예제를 통해 효과를 비교 검증한다.

주요용어: 강건모형, 박스-칵스 변환, 씨-단계, 최대절사우도법.

1. 서론

회귀모형에 대한 추정은 이상값이나 영향값에 민감하게 반응하기 때문에 이들을 확인하고 제거해 주는 것이 중요하다. 만약 데이터 내에 이상값이 작은 비율로 포함되어 있다면 이를 식별하는데 크게 문제될 것이 없겠으나, 데이터 규모에 비해 이상값의 비율이 크다면 이상값 식별 문제는 해결하기 쉽지 않으며 가면화 효과 (masking effects)나 수렁효과 (swamping effects)와 같은 오류가 발생할 가능성이 높다. 이와 같은 논의는 회귀모형의 반응변수변환 문제에도 적용되며 변수변환의 필요성을 검정하거나 구체적인 변환 값을 추정할 때 이상치에 강건한 변수변환방법이 필요하다. 이상치를 고려한 변수변환 문제에 관련하여 본 연구에서 고려하는 모형은, 반응변수와 설명변수 간 선형적 관계뿐만 아니라 형태가 알려지지 않은 비모수 함수가 포함된 아래와 같은 부분선형모형이다.

$$Y^{(\lambda)} = X\beta + h(Z) + \epsilon \tag{1.1}$$

여기서 $Y^{(\lambda)}$는 Box와 Cox (1964)가 제안한 누승변환이며 β는 알려지지 않은 $p \times 1$벡터이고 X는 Z를 제외한 설명변수 벡터를 나타낸다. 오차항 ϵ은 X와 Z에 대하여 독립이고 $h(Z)$는 알려지지 않은 함수이다.

부분선형모형에 있어서 이상치 탐지에 대한 연구는 분산에 관련된 특정 모델을 가정하여 주로 수행되었다 (Fung 등, 2002). 본 논문은 부분선형모형에서 이상치를 고려하여 반응변수 변환 값을 추정하는 방법들을 제안한다. 여기에는 이상치 탐지, 부분선형모형 추정, 반응변수변환값 추정 등에 관련된 여러 방법들이 복합적으로 활용되어야 한다. 부분선형모형의 비모수 추정을 위해서는 편잔차 (partial

[†] 이 논문은 2012학년도 건국대학교의 지원에 의하여 연구되었음.
[1] (143-701) 서울시 광진구 화양동 1번지, 건국대학교 응용통계학과, 교수.
[2] 교신저자: (608-737) 부산광역시 남구 대연3동 599-1번지, 부경대학교 통계학과, 조교수.
E-mail: myoon@pknu.ac.kr

residual) (Cook, 1993; Larsen과 Mccleary, 1972; Weisberg, 2005), 덧편잔차 (augmented partial residual plot) (Mallows, 1986), CERES (Cook, 1993) 등을 이용할 수 있으나 본 연구에서는 과대적합의 위험 등을 고려하여 덧편잔차를 이용한다. 이상치 탐지에 관련된 여러 방법들은 Jajo (2005) 등에 정리 비교되어 있으며 부분선형모형에서 이상값을 고려한 변수변환의 연구는 선형모형에서 제안된 방법을 확장하여 적용할 수 있다. 이상치 문제를 해결하는 방법에는 M 추정량, 최소중위수제곱 추정법, 최소절사제곱추정법등 이상치에 민감하지 않는 추정량 또는 기준을 사용하는 것과 직접적으로 이상값을 탐지하여 제거하는 것이 있다. 이상치를 탐지하는 방법은 RSS (residual sum of square) 최소축소법 (Gentleman과 Wilk, 1975), 다단계 RSS 최소축소법 (Marashingshe, 1985), 일반적 극대 스튜던트 잔차 사용법 (Paul과 Fung, 1991)등과 같이 이상치의 크기를 사전에 정해두는 방법과 반복 잔차 사용법 (Kianifard와 Swallow, 1989, 1990), 순차적 검정법 (Hadi와 Simonoff, 1993)과 같이 순차적 검정을 통해 이상치의 크기를 결정하는 방법으로 구분된다. 이러한 방법들 중에서 본 연구는 간접적인 방법으로, Cheng (2005)이 선형모형에서 이상값에 강건한 반응변수 변환을 위해 적용한 최대절사우도추정 절차를 부분선형모형에 적합하도록 확장하며 직접적인 방법으로는 과정의 간결성, 계산량, 방법의 효율성을 고려하여 순차적 검정법을 부분선형모형에 적용하여 이상치를 탐지한다. 본 논문의 2절에서는 각각 순차적검정과 최소절사제곱추정법을 적용하여 부분선형모형에서 이상치를 고려한 반응변수변환 과정을 제시하고 3절에서는 모의실험과 예제를 통해 제안한 두 방법과 이상치를 고려하지 않은 방법의 효율성을 비교한다. 마지막으로 연구의 결론과 향후 연구방향에 대하여 정리한다.

2. 부분선형 모형에서 이상치 진단과 반응변수변환 지수 추정

선형모형의 경우 반응변수 변환모수 λ의 추정법이나 이상치 탐지법이 다양하게 제안되고 있으나 부분선형모형의 경우에는 극히 제한적이다. 본 연구에서는 부분선형모형의 비모수 함수를 적절히 추정한 후 이상치를 고려한 반응변수 변환모수 추정법을 적용하고자 한다. 부분선형모형 $Y = X\beta + f(Z) + \epsilon$에서 비모수 함수 f를 추정하기 위하여 편잔차, 덧편잔차 또는 CERES 그림을 이용할 수 있다. 예를 들어 편잔차를 이용하는 경우 자료를 선형모형 $Y = Xa + Zb + \epsilon$에 적합시키고 임의의 볼록 목적함수 $L_n(a,b) = (1/n)\sum_{i=1}^{n} L(y_i - x_i a - z_i b)$을 최소화하는 값으로 모형의 각 계수를 추정하면 $\widehat{a}$은 $\widehat{\beta}$에 접근 (almost surely converge)하고 편잔차 $e + Z\widehat{b}$는 상수 $+ f(z) + \epsilon$에 접근하게 된다 (Cook, 1993). 본 연구에서는 모형 (1.1)에서 덧편잔차를 이용하여 함수 $h(Z)$를 추정한다. 덧편잔차는 부분선형모형에서 Y를 모형 $Y = X\rho + \phi_1 Z + \phi_2 Z^2 + \epsilon$에 적합하여 계산되는 값 $u = (Y - X\widehat{\rho})$이며 덧편잔차 u와 변수 Z의 산점도인 덧편잔차 그림에서 비모수적 방법 또는 모수적인 방법으로 $h(Z)$를 추정한다.

부분선형모형의 함수 (1.1)에서 임의의 λ를 가정하여 $h(Z)$를 추정한 후 이를 기반으로 새로운 반응변수 변환모수를 최대우도추정법으로 추정하는 경우 이상치는 낮은 우도를 갖는 관찰치로 이해할 수 있다. 이와 같은 관점에서 이상치의 영향력을 배제한 추정법은 최대절사우도법이며 $(n-q)$개의 이상치를 고려한 최대절사우도법에서 목표함수는 식 (2.1)과 같은 절사분산추정량과 일치한다 (Hadi와 Luceno, 1997).

$$\widehat{S}^2(\lambda) = \sum_{i \in M} \frac{e_i^2(\lambda)}{(q-p)} \tag{2.1}$$

여기서 $e_i^2(\lambda)$는 잔차이고 M은 잔차가 가장 작은 q개의 관측치 집단을 표시한다. 최대절사우도 추정량을 계산하기 위한 q개의 정상치군 M은 Rousseeuw와 Driessen (2006)이 제안한 C-단계를 이용하여 구할 수 있으며 Cheng (2005)등은 선형모형에서 C-단계를 이용한 변환모수 추정과정을 제시하였다. C-단계를 이용하여 q개의 정상치군을 선택하는 과정은 아래와 같다.

1. 현재의 λ를 추정치 $\widehat{\lambda}$이라고 하자.
2. 전체 n개의 관찰치로부터 $(p+1)$개 관찰치를 임의로 추출한다.

3. (a) 추출된 $(p+1)$개의 관찰치에 의하여 회귀모형을 추정한 후 이 모형으로부터 전체 n개의 데이터에 대한 잔차를 계산한다.

   (b) 잔차의 절대값을 기준으로 가장 작은 잔차를 갖는 q개의 관찰치 (M_0)를 선발하며 이 관찰치에 의한 절사분산 추정량 $S_0^2(\widehat{\lambda})$를 계산한다.

   (c) 3.(b)에서 선발된 q개의 관찰치(M_0)만 으로 회귀모형을 추정한 후 3.(b)에서와 같이 가장 작은 잔차를 갖는 q개의 관찰치(M_1)를 선발하고 해당하는 절사분산추정량 $S_1^2(\widehat{\lambda})$을 계산한다. 만약 $S_0^2(\widehat{\lambda}) = S_1^2(\widehat{\lambda})$ 이면 최적 절사분산 추정량과 최적 정상치군을 각각 $S_0^2(\widehat{\lambda})$과 M_0으로 결정하고 단계3을 마치며 $S_0^2(\widehat{\lambda}) \neq S_1^2(\widehat{\lambda})$이면 M_1을 가지고 $S_{i-1}^2(\widehat{\lambda}) = S_i^2(\widehat{\lambda})$가 될 때 까지 3.(b)와 3.(c)의 과정을 반복한다.

4. 또 다른 $(p+1)$개 관찰치를 임의로 추출한 후 3.(a), 3.(b), 3.(c)의 과정을 통하여 해당 부표집에 따른 최적 절사분산 추정량과 최적 정상군을 계산한다.

5. k 개 부표집에 의해 계산된 최적 절사분산 추정량들 중 가장 작은 값에 해당하는 최적정상군을 크기 q개의 최종 최적정상군으로 결정한다.

이와 같은 C-단계를 적용하여 본 연구에서 제안하는 모형 (1.1)에서 최대절사우도추정법에 의하여 변수변환 모수를 추정하는 절차는 아래와 같다.

단계 0 현 단계에서의 변환변수 추정값을 $(\widehat{\lambda}_{pr})$, 절사제곱추정량을 $S^2(\widehat{\lambda}_{pr})$이라고 하자.

단계 1 반응변수를 $Y^{\widehat{\lambda}_{pr})}$으로 변환 한다.

단계 2 $Y^{(\widehat{\lambda}_{pr})}$를 모형 $Y^{(\widehat{\lambda}_{pr})} = X\rho + \phi_1 Z + \phi_2 Z^2 + \epsilon$에 적합시켜 계산된 덧편잔차와 변수 Z의 산점도로 부터 $\widehat{h}(Z)$를 모수적 또는 비모수적 방법에 따라 추정한다.

단계 3 $Y^{(\widehat{\lambda}_{pr})}$를 X와 $\widehat{h}(z)$에 의해 적합시키고 C-단계를 수행하여 새로운 정상치군 M을 선발한다.

단계 4 정상치군 M을 이용하여 새로운 잠정적인 추정값 $\widehat{\lambda}^{TP}$를 구한다. $\widehat{\lambda}^{TP}$는 Tsai와 Wu (1990)에서 제안한 비정상치군 제거 반응변수 변환모수에 대한 최대우도추정량으로 계산한다.

단계 5 $\widehat{\lambda}^{TP}$에 따라 반응변수를 변환시킨 후 정상데이터 M만을 가지고 단계 2와 단계 3과 같이 부분선형모형을 적합시킨 후 식 (2.1)의 $S^2(\widehat{\lambda}^{TP})$을 계산한다. 만약 $S^2(\hat{\lambda}^{TP}) \geq S^2(\widehat{\lambda}_{pr})$이면 $\widehat{\lambda}_{pr}$을 λ의 최종 추정값으로 결정하고 $S^2(\widehat{\lambda}^{TP}) < S^2(\widehat{\lambda}_{pr})$이면 $\widehat{\lambda}^{TP}$를 새로운 $\widehat{\lambda}_{pr}$값으로 대체한 후 단계 1 에서부터 위 과정을 반복한다.

위의 단계에서 변환변수는 Y와 $X, \widehat{h}(Z)$의 모형에서 Box-Cox 추정법을 적용하여 추정한다. 또한 변환변수의 초기값은 λ를 -2에서 2까지 0.5단위로 각각 적용하여 단계 2를 수행한 후 Y와 $X, \widehat{h}(Z)$에 의해 추정되는 $\widehat{\lambda}$중에서 우도값이 가장 큰 것을 사용한다. 최대절사우도법에서 목표함수인 식 (2.1)은 반응변수 변환 관점에서 최소절사제곱법의 목표함수와 일치한다 (Cheng, 2005). 따라서 식 (2.1)을 목표함수로 지정하고 C-단계 대신 특정 이상치 탐지법을 적용하면 반응변수의 변환모수 추정에 대해 일종의 최소절사제곱법이 된다. 본 연구에서는 이와 같은 과정에서 이상치 탐지를 위해 Hadi와 Simonoff (1993)가 제안한 순차적 검정법을 사용한다. Hadi와 Simonoff (1993)의 순차적 검정법은 기초정상치군을 구성한 후 이 데이터 군을 기반으로 정상치의 크기를 늘려가면서 이상치를 탐지해간다. 기초 정상치군은 모든 관찰치에 의하여 추정된 회귀모형으로부터 계산된 가장 작은 p개의 잔차에 해당하는 관찰치들을 선별 한 후 여기에 회귀모형을 적용시켜 n개의 모든 데이터에 대해 내부 스튜던트 잔차 (internally studentized residual)를 계산하고 그 값이 가장 작은 $p+1$개의 관찰치를 다시 선별하는 과정을 반복하여 크기 $int[(n+p-1)/2]$인 군을 선발한다. 기초정상데이터군이 생성되면 이상치 검정의 과정을 수행

한다. 이상치 검정 과정에서는 우선 정상 데이터군 만을 가지고 모형을 추정한 후 다음과 같은 내부 스튜던트 잔차 d_i를 계산하여 t 통계량과 비교한다.

$$d_i = \begin{cases} \frac{(y_i - x_i^T \hat{\beta_M})}{\hat{\sigma}_M \sqrt{1 - x_i^T (X_M^T X_M)^{-1} x_i}}, & \text{만약 } i \in M \\ \frac{(y_i - x_i^T \hat{\beta_M})}{\hat{\sigma}_M \sqrt{1 + x_i^T (X_M^T X_M)^{-1} x_i}}, & \text{만약 } i \notin M. \end{cases}$$

여기서 X_M은 정상데이터군 M만으로 구성된 독립변수 행렬이고, $\hat{\beta}_M$와 $\hat{\sigma}_M$은 정상데이터군으로 추정된 β와 σ의 추정치이다. 현재의 정상데이터군 M의 크기를 s라고 하고 $d_{(s+1)}$를 $|d_i|$의 $(s+1)$번째 순서 통계량이라고 할 때, $d_{(s+1)} \geq t_{(\alpha/2(s+1),s-3)}$이면, $|d_i| \geq t_{(\alpha/2(s+1),s-3)}$를 만족하는 모든 관측값들을 이상값으로 판정하며 만약 $d_{(s+1)} < t_{(\alpha/2(s+1),s-3)}$이면, $(s+1)$개의 가장 작은 $|d_i|$에 해당하는 관측치로 새로운 정상 데이터군 M을 구성하여 위의 과정을 반복한다.

3. 모의실험과 예제

C-단계 활용법과 순차적 검정 활용법을 적용한 두 가지 과정과 이상치 탐지를 적용하지 않고 반응변수 변환모수를 추정하는 방법의 효율성을 비교하기 위해 모의실험을 수행한다. 한 번의 실험에 사용된 가상데이터의 크기는 $n = 50$이며 이중 정상 관찰치는 47개, 비정상 관찰치는 3개이고 독립변수 X의 차원은 2이다. 정상 관찰치에서 독립변수 X는 적절한 범위의 균등분포에 의해 생성되며 변수 Z는 $(0.5, 2.5)$ 사이에서 균일한 간격의 값을 갖는다. 반응변수 Y는 변환변수 λ에 따라 $y^{(\lambda)} = XI + c \times (z - 1.5)^2$의 관계식에 의해 생성되며 상수 c는 임의로 지정된다. 비정상 관찰치는 다음과 같은 다변량 정규분포에 의해 생성된다.

$$\begin{pmatrix} y \\ X_1 \\ X_2 \\ Z \end{pmatrix} = MN \left(\begin{pmatrix} 3 \\ 15 \\ 15 \\ 15 \end{pmatrix}, \begin{pmatrix} 0.5 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0.5 \end{pmatrix} \right)$$

200번의 모의실험에서 각 방법에 의해 추정된 변환변수 $\hat{\lambda}$의 평균과 표준편차가 계산되었으며 이상치 탐지와 관련된 세 가지 비율을 함께 계산되었다. p_1은 정확하게 이상치를 탐지한 비율이며, p_2는 적어도 한 개 이상 정확한 이상치를 탐지한 비율이고 p_3는 정상치를 이상치로 탐지한 비율이다.

Table 3.1 Simulation results

λ	-1			-0.5			0	
methods	sequential	C-step	all data used	sequential	C-step	all data used	sequential	C-step
mean	-1.01	-1.22	-1.36	-0.51	-0.56	-0.67	-0.01	-0.22
st. dev.	0.50	0.32	0.22	0.14	0.11	0.11	0.32	0.36
p_1	0.82	0.73	-	0.72	0.68	-	0.78	0.76
p_2	0.98	0.92	-	0.88	0.92	-	0.86	0.90
p_3	0.08	0.27	-	0.08	0.32	-	0.05	0.24

λ	0	0.5			1		
method	all data used	sequential	C-step	all data used	sequential	C-step	all data used
mean	-0.33	0.46	0.41	0.23	0.94	0.88	0.70
st. dev.	0.26	0.11	0.16	0.08	0.24	0.30	0.28
p_1	-	0.85	0.77	-	0.81	0.76	-
p_2	-	0.98	0.88	-	0.95	0.93	-
p_3	-	0.05	0.23	-	0.10	0.24	-

따라서 $1-p_2$는 모의실험을 통하여 가면화 효과가 발생한 비율을 표시하고 p_3는 수렁화 효과가 발생한 비율이 된다. C-단계 방법 적용시 p_1을 계산하기 위하여 이상치의 크기는 3개로 지정하였으며 부표집의 횟수는 가능한 부표집 전부를 수행하였다.

모의실험 결과 Table 3.1에서 알수 있듯이 순차적 검정을 적용한 방법이 모든 λ값에 있어서 C-단계 적용 방법이나 이상치를 고려하지 않은 방법 보다 변환변수를 보다 더 정확하게 추정하는 것을 알 수 있다. C-단계를 사용한 경우 이상치 개수를 실제 이상치 갯수와 동일한 3으로 고정하였으므로 $p_3 = 1-p_1$의 관계가 되어 순차적 검정방법과 p_3값을 비교하는 것은 무의미하지만 p_1, p_2 관점에서도 순차적 검정을 적용한 방법이 C-단계를 적용한 방법보다 정확하게 이상치를 탐지하고 있으며 수렁화나 가면화의 비율도 낮은 것을 확인할 수 있다.

예제 3.1 호수 질소농도 자료 (Nitrogen in lakes)

크기 $n = 29$의 호수 질소농도 자료 (Atkinson과 Riani, 2000)는 유입질소 평균농도 (x_1), 물 보존시간 (x_2), 연 평균 질소농도 (y)등의 변수로 구성되어 있다. 이 자료에 대하여 Stromberg (1993)와 Atkinson과 Riani (2000)는 다음과 같은 모형을 이용하여 분석하였으며 10, 23번째 관찰치를 이상치로 진단하였다.

$$y_i = \frac{x_{1i}}{1+\beta_1 x_{2i}^{\beta_2}} + \epsilon_i,\ i = 1, \ldots, 29 \tag{3.1}$$

식 (3.1)이 적절한 모형이라는 가정아래 로그 변환을 하면 $\log E(y_i) = \log x_{1i} + h(x_{2i})$가 되어 Y를 $\log x_1$과 x_2를 이용하여 부분선형모형에 적합시켰을 때 10, 23번째 관찰치를 이상치로 판정하고 으로 추정하는 것을 기대할 수 있다. Table 3.2는 호수 질소농도 자료에 대해 본 순차적 검정방법과 이상치 크기 각각 1, 2, 3의 C-단계 적용방법, 그리고 이상치를 고려하지 않은 방법의 결과를 보여준다.

Table 3.2 Outlier identification and estimation results for nitrogen in lake data

methods	sequential	C-step 1	C-step 2	C-step 3	all data used
λ	0.03	-0.08	0.03	0.07	0.12
outliers	10, 23	23	10, 23	2, 10, 23	-

순차적 검정방법을 적용한 경우 10과 23번째 관찰치를 이상치로 분류하고 $\hat{\lambda} = 0.03$으로 추정하였으며 C-단계를 적용한 방법도 이상치의 크기를 2로 고정하였을 때 동일한 결과를 보여주며 이상치 개수를 한 개 또는 세 개를 고정하였을 때도 선행 연구에 부합하는 결과를 보여준다. 이상치를 고려하지 않고 λ를 추정하였을 경우 $\hat{\lambda} = 0.12$가 되어 순차적 검정법을 적용한 경우나 C-단계를 적용한 경우보다 log 변환에서 멀어진 추정 값을 제시하고 있다.

4. 결론

부분선형모형에서 이상값을 고려하여 반응변수 변환을 수행하는 과정에 대하여 본 연구에서 제안하는 방법은 최대우도추정법에서 낮은 우도를 이상치로 간주하는 최대절사우도추정법과 순차적 검정방법이다. 반응변수 변환모수 추정에 있어서 최대절사우도추정법은 최소절사제곱 (least trimmed square) 추정법과 일치하므로 두 가지 방법을 적용하는 과정에서 동일한 목표함수를 사용할 수 있다. 이상치 탐지 과정과 최대절사우도법을 적용하여 부분선형모형에서 반응변수변환값을 추정하기 위해서는, 먼저 반응변수변환 지수 λ를 임의로 정하고 지정된 λ값에 의해 변환된 반응변수로 부분선형모형에서 이상치를 탐지한다. 탐지된 이상치를 고려하여 반응변수변환 지수 λ를 새롭게 추정하며 새롭게 추정된 λ값으로 변환된 반응변수로 모형을 추정한다. 추정된 모형의 목표함수를 계산하여 이전에 추정된 λ의 절사제곱추정량과 비교하여 λ의 갱신 여부를 결정한다.

본 연구에서 제안한 두 가지 방법중 순차적 검정법을 적용한 과정이 보다 더 효율적임이 실험과 예제를 통해 검증하였다. 이러한 결과와 더불어 C-단계의 과도한 계산량이나 이상치군의 크기를 사전에 정해야 하는 점 등을 고려할 때 일정크기 이상의 자료에서는 순차적 검정법을 적용하는 과정이 효과적이라고 할 수 있다.

본 연구에서는 C-단계나 순차적 검정을 수행할 때 $Y^{\hat{\lambda}_{pr}}$과 X, $\hat{h}(z)$을 이용하였지만 Seo와 Yoon (2010)에서와 같이 덧편잔차와 Z를 이용할 수도 있으며 Seo와 Yoon (2009)에서 제시한 과정과 유사하게 동적그림을 작성하여 이상치와 반응변수 변환모수를 탐색적으로 추정하는 것도 시도해 볼 수 있다. 본 연구에서는 부분선형모형에 포함된 비모수 함수의 추정 과정에 이상치 탐지 방법을 적용하지 않았지만 이 부분에서도 이상치를 고려하는 통합적인 방법을 개발하는 것이 추후 연구 과제로 고려할 수 있다.

참고문헌

Atkinson, A. C. and Riani, M. (2000). *Robust diagnostic regression analysis*, Springer, New York.

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

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

Cook, R. D. (1993). Exploring partial residual plots. *Technometrics*, **35**, 351-362.

Fung, W., Zhu, Z., Wei, B. and He, X. (2002). Influence diagnostics and outlier tests for semiparametric mixed models. *Journal of the Royal Statistical Society B*, **64**, 565?579.

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.

Jajo, N. K. (2005). A Review of robust regression an diagnostic procedures in linear regression. *Acta Mathematicae Applicatae Sinica*, **21**, 209-224.

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.

Kianifard, F. and Swallow, W. H. (1996). A review of the development and application of recursive residuals in linear models. *Journal of the American Statistical Association*, **91**, 391-400.

Larsen, W. A. and McCleary, S. J. (1972). The use of partial residual plots in regression analysis. *Technometrics*, **14**, 781-790.

Mallows, C. L. (1986). Augmented partial residual plots. *Technometrics*, **28**, 313-320.

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

Seo, H. S. and Yoon, M. (2009). A dynamic graphical method for transformations and curvature specifications in regression. *The Korean Journal of Applied Statistics*, **22**, 189-195.

Seo, H. S. and Yoon, M. (2010). Outlier detection methods using augmented partial residual plots in a partially linear model. *Journal of the Korean Data Analysis Society*, **12**, 1125-1133.

Stromberg, A. J. (1993). Computation of high breakdown nonlinear regression parameters. *Journal of the American Statistical Association* **88**, 237-244.

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

Weisberg, S. (2005). *Applied linear regression*, 3rd Ed., John Wiley, New York.

Journal of the Korean Data & Information Science Society
2013, **24**(1), 33–39

http://dx.doi.org/10.7465/jkdi.2013.24.1.33
한국데이터정보과학회지

Regression diagnostics for response transformations in a partial linear model[†]

Han Son Seo[1] · Min Yoon[2]

[1]Department of Applied Statistics, Konkuk University
[2]Department of Statistics, Pukyong National University

Received 7 November 2012, revised 2 December 2012, accepted 7 December 2012

Abstract

In the transformation of response variable in partial linear models outliers can cause a bad effect on estimating the transformation parameter, just as in the linear models. To solve this problem the processes of estimating transformation parameter and detecting outliers are needed, but have difficulties to be performed due to the arbitrariness of the nonparametric function included in the partial linear model. In this study, through the estimation of nonparametric function and outlier detection methods such as a sequential test and a maximum trimmed likelihood estimation, processes for transforming response variable robust to outliers in partial linear models are suggested. The proposed methods are verified and compared their effectiveness by simulation study and examples.

Keywords: Box-Cox transformation, C-step, maximum trimmed likelihood estimation, robustness.

[†] This work was supported by the Konkuk University 2012.
[1] Professor, Department of Applied Statistics, Seoul 143-701, Korea.
[2] Corresponding author: Assistant professor, Department of Statistics, Busan 608-737, Korea. E-mail: myoon@pknu.ac.kr