The Korean Journal of Applied Statistics (2013) DOI: http://dx.doi.org/10.5351/KJAS.2013.26.4.687
26(4), 687–696

An Outlier Detection Method in Penalized Spline Regression Models

Han Son Seo[a] · Ji Eun Song[a] · Min Yoon[b,1]

[a]Department of Applied Statistics, Konkuk University
[b]Department of Statistics, Pukyong National University

(Received July 15, 2013; Revised August 13, 2013; Accepted August 21, 2013)

Abstract

The detection and the examination of outliers are important parts of data analysis because some outliers in the data may have a detrimental effect on statistical analysis. Outlier detection methods have been discussed by many authors. In this article, we propose to apply Hadi and Simonoff's (1993) method to penalized spline a regression model to detect multiple outliers. Simulated data sets and real data sets are used to illustrate and compare the proposed procedure to a penalized spline regression and a robust penalized spline regression.

Keywords: Outlier detection methods, penalized spline regression, robust method.

1. 서론

선형회귀분석에서 회귀계수 추정을 위해 주로 사용되는 최소제곱추정은 잔차제곱합의 크기에 의존하므로 상대적으로 큰 잔차값은 모수추정에 큰 영향을 줄 수 있으며 이것이 이상치인 경우 모형 적합에 왜곡된 결과를 초래할 수 있어 이상치 탐지가 매우 중요하다.

이상치 탐지는 이상치로 의심되는 관찰값들을 찾아내고, 그 값들이 실제로 이상치인지 아닌지 판별하는데 목적이 있다. 이상치 탐지방법은 크게 '직접적 이상치 탐지방법'과 강건추정의 잔차를 이용한 '간접적 이상치 탐지방법'으로 나뉜다. 간접적 이상치 탐지방법은 이상치에 영향을 덜 받는 강건 추정량을 사용하거나 이상치에 영향을 받지 않는 기준을 적용하여 이상치의 영향을 배제하는 방법이다. 간접적 이상치 탐지방법으로는 Huber (1973)에 의해 소개된 M-추정량, Rousseeuw (1984)가 제안한 최소메디안 제곱추정량, 최소 절사 제곱추정량등이 있다. 직접적 이상치 탐지방법은 검정등을 통하여 이상치를 탐지한 후 제거하는 방법이다. 예를 들면 Kianifard와 Swallow (1989)는 이상치를 탐색하는 방법으로 스튜던트화 잔차나 Cook's Distance와 같은 회귀진단 기준에 따라 관측값을 순서대로 정렬한 뒤 잔차를 계산하는 방법을 제안하였으며, 그 외에도 RSS 최소축소법 (Gentleman과 Wilk, 1975), 다단계 RSS 최소축소법 (Marasinghe, 1985), 순차적 검정법 (Hadi와 Simonoff, 1993) 등이 있다. 이중 Hadi와 Simonoff (1993)는 일정 크기의 양호 관찰치군으로부터 모형을 추정한 후, 양호치군과 이상치군에서 계산된 내적 스튜던트화 잔차(internally studentized residual)의 순서통계량에 대한 t 검정 결과에 따라 최종 이상치군을 결정하거나 양호치군의 크기를 한 개 늘려서 반복된 절차를 수행한다.

[1]Corresponding author: Associate Professor, Department of Statistics, Pukyong National University, 45, Yongso-ro, Nam-Gu, Busan 608-737, Korea. E-mail: myoon@pknu.ac.kr

본 논문에서는 벌점 스플라인 회귀모형(penalized spline regression model)에서 이상치를 탐지하기 위하여 Hadi와 Simonoff (1993)의 이상치군 식별법을 적용하며 제안된 방법은 강건 벌점 스플라인 회귀방법, 이상치 탐지를 하지 않은 벌점 스플라인 회귀방법과 비교한다. 2장에서는 벌점 스플라인 회귀와 강건 벌점 스플라인 회귀방법을 설명하고 본 연구에서 제안하는 Hadi와 Simonoff (1993)의 정상데이터군 식별법을 통하여 이상치를 검출한 후 벌점 스플라인 회귀를 수행하는 알고리즘을 설명한다. 제 3장에서는 모의실험과 실제데이터를 이용하여 각 방법의 적합결과와 비교하고 4장에서 연구결과를 요약한다.

2. 벌점 스플라인 회귀모형의 강건 추정법과 이상치 탐지법

일반적으로 회귀모형은 다음과 같은 관계식을 가정한다.

$$Y_i = m(X_i) + \epsilon_i, \qquad i = 1, \ldots, n,$$
$$m(x_i) = E(Y|X = x_i),$$

여기서 $m(x_i)$는 알려지지 않은 회귀함수이고 ϵ_i는 평균이 0, 분산은 σ^2이며, ϵ_i들은 독립적이다. 함수 $m(x_i)$를 추정하기 위한 H개의 knot이 있는 일반적인 p차 스플라인 모형은 다음과 같다.

$$m(x;\beta) = \beta_0 + \beta_1 x + \cdots + \beta_p x^p + \sum_{h=1}^{H} \beta_{p+h}(x - K_h)_+^p,$$

여기서 $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_p, \beta_{p+1}, \ldots, \beta_{p+H}]^T$, β_{p+h}는 h번째 knot에 대한 회귀계수이고, K_h는 h번째 knot이며 μ_+는 μ가 양수이면 μ, 그 외에는 0의 값을 갖는다.

벌점 스플라인 회귀는 $\sum \beta_{p+h}^2 < C$와 같은 제약조건을 고려한 $\boldsymbol{\beta}$의 추정문제와 동일하여, $\lambda \geq 0$일 때 $||\boldsymbol{y} - \boldsymbol{X\beta}||^2 + \lambda \boldsymbol{\beta}^T D \boldsymbol{\beta}$를 최소화하는 $\boldsymbol{\beta}$를 찾는 것이다. 여기서 λ는 평활의 정도를 결정하는 평활모수이며 $\boldsymbol{y}$는 $(y_1, \ldots, y_n)^T$이고, 행렬 D는 $\text{diag}\{\mathbf{0}_{(p+1)\times 1}, \mathbf{1}_{H\times 1}\}$이며 행렬 X는 다음과 같다:

$$X = \begin{bmatrix} 1 & x_1 & \cdots & x_1^p & (x_1 - K_1)_+^p & \cdots & (x_1 - K_H)_+^p \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \cdots & x_n^p & (x_n - K_n)_+^p & \cdots & (x_n - K_H)_+^p \end{bmatrix}.$$

벌점 스플라인 회귀에 의해 적합한 값은 일종의 능형회귀이고 이때의 해는 $\hat{\beta}_\lambda = (X^T X + \lambda D)^{-1} X^T \boldsymbol{y}$, $\hat{\boldsymbol{y}} = X(X^T X + \lambda D)^{-1} X^T \boldsymbol{y}$이다.

벌점 스플라인 회귀에서 이상치가 존재하면 회귀함수를 추정할 때 왜곡된 결과를 얻을 수 있다. 벌점 스플라인 모형에서 강건성을 갖기 위하여 M-추정법을 적용하여 다음과 같은 기준으로 모형을 추정한다.

$$\sum_{i=1}^{n} \rho_c(y_i - m(x_i;\boldsymbol{\beta})) + \lambda \sum_{k=1}^{K} \beta_{p+k}^2,$$

여기서 $\rho_c(x)$는 다음과 같은 Huber의 손실함수(loss function)이다($c > 0$):

$$\rho_c(x) = \begin{cases} x^2, & |x| \leq c, \\ 2c|x| - c^2, & |x| > c. \end{cases}$$

그러나 위와 같이 정의된 M-추정법을 사용하여 모형을 추정할 때 λ의 적절한 값을 결정하는 문제와 최적화 기준을 최소화하는 현실적인 알고리즘을 개발하기가 쉽지 않다는 문제가 발생한다. 이를 해결하기

위해 Lee와 Oh (2007)는 가상 데이터(pseudo data)를 이용하여 계산량이 작고 속도가 빠른 강건 벌점 스플라인 회귀방법을 제안하였다.

벌점 스플라인 회귀에서 $\{B_i\}_{i=1}^n$은 스플라인 기본함수 집합이며 $\boldsymbol{B}(x) = \{B_1, B_2, \ldots, B_n\}^T$라고 표기할 때 회귀함수를 일반적으로 표현하여 $m(x) = \boldsymbol{\beta}^T\boldsymbol{B}(x)$라고 하면 추정량은 다음의 함수를 최소화하는 값으로 계산된다.

$$\sum_{i=1}^{n}\left\{y_i - \boldsymbol{\beta}^T\boldsymbol{B}(x_i)\right\}^2 + \lambda\boldsymbol{\beta}^T\boldsymbol{D}\boldsymbol{\beta}$$

이때 M-추정법 벌점 스플라인 회귀추정량 $\hat{m}(x)$에 앞서 정의한 $\rho_c(x)$를 적용하면 아래와 같다.

$$\hat{m}_{robust}(x) = \arg\min_m \sum_{i=1}^{n}\rho_c\left(y_i - \boldsymbol{\beta}^T\boldsymbol{B}(x_i)\right) + \lambda\boldsymbol{\beta}^T\boldsymbol{D}\boldsymbol{\beta}.$$

위의 식에서 미분을 이용하여 $\hat{\boldsymbol{\beta}}$을 추정할 때 ρ_c와 ψ_c의 비선형적 성질로 인하여 계산이 쉽지 않지만 $m_i = m(x_i)$라고 할 때, 가상 데이터 $\tilde{y_i} = m_i + \psi_c(y_i - m_i)/2$라고 정의하고 아래와 같이 $\tilde{y_i}$값이 알려지면 상응하는 벌점 최소제곱 추정치의 폐쇄형(closed form) 해가 존재하게 된다.

$$\hat{m}_{pseudo}(x) = \arg\min_m \sum_{i=1}^{n}\left\{\tilde{y_i} - \boldsymbol{\beta}^T\boldsymbol{B}(x_i)\right\}^2 + \lambda\boldsymbol{\beta}^T\boldsymbol{D}\boldsymbol{\beta}.$$

이와 같이 추정된 $\hat{m}_{pseudo}(x)$은 $\hat{m}_{robust}(x)$에 수렴하며 (Lee와 Oh, 2007, Theorem 1) 이에 따라 Lee와 Oh (2007)는 $\hat{m}_{robust}(x)$을 계산하는 반복적 알고리즘을 제안하였다.

본 연구에서는 Hadi와 Simonoff (1993)가 선형회귀모형에서 사용한 순차적 방법을 벌점 스플라인 회귀에 적용하여 벌점 스프라인 회귀에서 이상치를 탐지하고 제거하는 직접적인 방법을 제안한다. Hadi와 Simonoff (1993)의 방법은 이상치일 가능성이 없는 관측값들의 집합을 먼저 설정하고 이를 바탕으로 추정된 회귀선으로 부터 각 관찰값의 오차를 계산하여 이상치가 될 가능성이 없는 값들을 처음에 설정한 집합에 추가하는 방식이다. M을 이상치일 가능성이 없는 관찰값들의 집합이라 하고, X_M은 집합 M에 해당하는 X의 부분행렬, $\hat{\beta}_M$은 집합 M에 의해 추정된 회귀계수, $\hat{\sigma}_M$은 집합 M에 의해 추정된 표본분산이라고 하자. 벌점 스플라인 회귀모형에 Hadi와 Simonoff (1993)의 방법을 적용하여 이상치를 탐지한 후 이를 제거하는 과정은 다음과 같이 정리될 수 있다.

(단계 0) 전체데이터에 벌점 스플라인 회귀모형 적합 :

전체데이터를 이용하여 아래와 같은 벌점 스플라인 회귀모형을 적합한다.

$$\sum_{i=1}^{n}\left\{y_i - \boldsymbol{\beta}^T\boldsymbol{B}(x_i)\right\}^2 + \lambda\boldsymbol{\beta}^T\boldsymbol{D}\boldsymbol{\beta}.$$

먼저 λ의 범위를 설정하고 각 λ에 따라 위의 식을 최소화하는 $\hat{\beta}_\lambda$를 구하며 GCV(generalized cross-validation)를 이용하여 최적의 $\hat{\beta^*}_\lambda$과 λ^*을 구한다. 이때 $\hat{\beta}_\lambda$의 적합값은 $\hat{\beta}_\lambda = X(X^TX + \lambda D)^{-1}X^Ty$이다.

(단계 1) 초기 정상데이터군 M생성 :

1. 우선적으로 데이터를 '기초 군(Basic set)'과 '비기초 군(Non-basic set)'으로 나눈다. 처음의 기초 군은 $(k+1)$개의 관측값으로 구성하고, 비기초 군은 나머지 $(n-k-1)$개의 관측값으로 구성한다. (단계 0)에서 전체 데이터에 의해 계산된 회귀계수 $\hat{\beta}_\lambda$를 이용하여 잔차를 구한 후 잔차의 절대값 $|e_i|$, $(i = 1, 2, \ldots, n)$를 오름차순으로 정렬하여 그 크기가 작은 $(k+1)$개의 관측값을 기초군으로 구성한다. 이때 k는 모형의 회귀계수의 개수이다.

2. 기초 군에 해당하는 $(k+1)$개의 관측값을 B라 하고, 기초 군 B에 의한 독립변수 행렬인 X_B로 벌점 스플라인 회귀를 적용하여 회귀계수 $\hat{\beta}_{B_\lambda}$를 추정한 후 잔차 $y_i - x_i^T\hat{\beta}_{B_\lambda}$를 구한다. 이때의 $\hat{\beta}_{B_\lambda}$도 GCV에 의해 결정된다.
3. 각 군에 대하여 아래와 같은 수정된 잔차 (adjusted residual) d_i'를 계산하여 값이 작은 순서를 기준으로 기초 군과 비기초 군 조정 작업을 반복한다. 아래 식에서 x_i는 행렬 X의 i번째 행벡터이다.

$$d_i' = \begin{cases} \dfrac{\left|y_i - x_i^T\hat{\beta}_{B_\lambda}\right|}{\sqrt{1 - x_i^T\left(X_B^TX_B + \lambda D\right)^{-1}x_i}}, & \text{if } i \in B, \\ \dfrac{\left|y_i - x_i^T\hat{\beta}_{B_\lambda}\right|}{\sqrt{1 + x_i^T\left(X_B^TX_B + \lambda D\right)^{-1}x_i}}, & \text{if } i \notin B. \end{cases}$$

4. 위의 과정을 한번 반복할 때 마다 기초 군 B의 개수를 1개씩 증가시켜 초기 정상 데이터군 M의 개수인 $h = [(n+k-1)/2]$가 될 때까지 계속한다.

(단계 2) 이상치 탐색 :

1. 기초 군 B의 크기가 h개가 되면 이를 초기 정상데이터군 M으로 할당하고, 나머지 $n-h$개의 관측값은 비정상 데이터군에 포함시킨다.
2. 정상 데이터군과 비정상 데이터군으로 나누어진 데이터로 표준화된 잔차 d_i를 아래와 같이 계산한다:

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

3. 이상치 여부를 판단하기 위해 d_i의 절대값 $|d_i|$의 크기가 작은 순서대로 데이터를 정렬한다. 주어진 정상 데이터군 M의 크기를 s라고 하고 $d_{(s+1)}$를 $(s+1)$번째 $|d_i|$값 이라고 하자.
   (a) 만약 $d_{(s+1)} \geq t_{(\alpha/2(s+1),s-k)}$이면, $(s+1)$번째와 그 이후의 관찰값을 이상치로 간주하고 과정을 멈춘다.
   (b) 만약 $d_{(s+1)}$가 $d_{(s+1)} < t_{(\alpha/2(s+1),s-k)}$이면 $(s+1)$번째 관찰값까지 정상 데이터군 M을 포함시켜 위의 과정을 반복 시행한다.

(단계 3) 정상 데이터군에 벌점 스플라인 회귀모형 적합 :

위의 단계에 의해 판정된 정상 데이터군을 이용하여 벌점 스플라인 회귀모형을 적합하며 각 λ에 따라 기준식을 최소화하는 $\hat{\beta}_\lambda$를 구하고 GCV를 이용하여 최적의 $\hat{\beta}_\lambda^*$와 λ^*을 구한다.

3. 모의실험 및 예제

3.1. 모의실험

강건 벌점 스플라인 회귀방법과 본 논문에서 제안한 이상치 탐지법을 비교하기 위하여 Lee와 Oh (2007), Cantoni와 Ronchetti (2001) 등에서 사용한 모형과 동일하게 아래의 함수를 이용하여 임의의

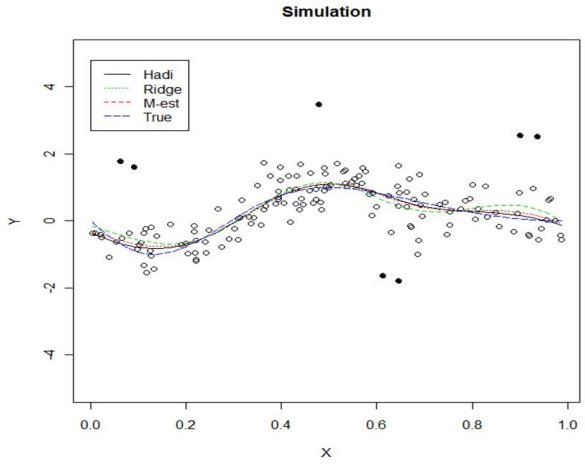

Figure 3.1. A simulated data set with outliers: Fitted curves and a curve representing the true model

자료를 발생 시켰다.

$$y_i = m(x_i) + 0.5\epsilon_i,$$
$$m(x_i) = \sin\left\{2\pi(1-x_i)^2\right\},$$

여기서 x_i는 구간 $(0,1)$의 균등분포에서, ϵ_i는 표준정규 분포에서 각각 150개씩 생성하였으며 이상치는 x_i의 작은 순서를 기준으로 10번째 이내에서 2개, 75번째에서 105번째 사이에 3개 그리고 135번째 이후에서 2개를 임의로 생성하였다. 모의실험의 자료는 100번 반복 생성되었으며 매번의 실험마다 모형오차인 $n^{-1}\sum\{m(x_i)-\hat{y}_i\}$을 계산하였다. 모의실험에서는 이상치가 포함되지 않은 경우에도 각 방법을 적합하여 그 결과를 비교하였다.

실험에서 사용한 quadratic 스플라인 모형은 아래와 같다.

$$m(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \sum_{h=1}^{H} \beta_{2+h}(x-K_h)_+^2.$$

등간격이 각각 $(1/20, 1/30, 1/40, 1/50)$인 네 개의 각 knot 집합에 대해, λ값은 0과 0.01사이에서 간격을 0.001로 하는 범위를 설정하여 CGV에 의해 결정되었다 (Ruppert와 Wand, 2003, Chapter 5). 여기서 구한 λ는 이상치를 고려한 경우에서는 최적의 값이 아니지만 어느 정도 신뢰성을 가지면서도 간편하고 빠르게 적용할 수 있다.

(1) 이상치가 있는 경우

Figure 3.1은 모의실험의 결과 중 한 가지 예를 나타낸다. 논문에서 제안된 방법의 적합선을 Hadi, 이상치를 고려하지 않은 벌점 스플라인 회귀의 적합선을 Ridge, 강건 벌점 스플라인 회귀의 적합선을 M-est, 모의실험 함수의 실제선을 True로 표기하였다. 검은색으로 채워진 점들은 제안된 방법에 의하여 이상치로 선택된 값이다. Figure 3.1을 살펴보면 능형회귀 적합선은 다른 두 적합선 보다 이상치가 있는 방향으로 가깝게 그려진 것을 확인할 수 있다.

Figure 3.2는 각 방법에 따른 실제값과 적합값의 편차를 제곱하여 평균한 모형오차의 상자그림이다. 상자그림에서 사분위편차로 비교해 볼 때, 산포는 일반 스플라인 회귀모형이 가장 크고, 그 다음이

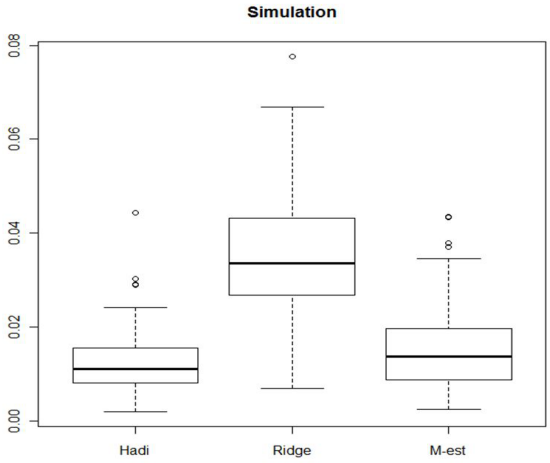

Figure 3.2. Box-plots of model-errors estimated from simulated data sets with outliers

강건 벌점 스플라인 회귀방법, 제안된 방법이 가장 작다. 모형오차의 중앙값으로 비교하면 일반 스플라인 회귀의 모형 오차가 가장 크고, 그 다음이 강건 벌점 스플라인 회귀방법, 제안된 방법이 가장 작으며 사분위편차로 비교한 산포도 동일한 결과를 보인다. 따라서 전반적으로 일반 벌점 스플라인의 모형오차값이 크다는 것을 알 수 있으며, 제안된 방법이 강건 스플라인 회귀 방법보다 상대적으로 모형오차값이 작은 것을 알 수 있다.

반복된 모의실험의 각 결과에서 고려된 방법들의 모형오차값을 비교한 결과, 제안된 방법은 강건 벌점 스플라인 회귀방법보다 100번 중 69번 작았으며 일반적 벌점 스플라인 회귀방법보다 99번 작았다. 제안된 방법은 100번의 반복실험에서 실제 이상치를 이상치로 탐지하지 못했던 경우는 없었고, 두 번의 실험에서 한 개씩의 정상 관찰값을 이상치로 추가로 잘못 선택하였다.

(2) 이상치가 없는 경우

Figure 3.3은 이상치가 없는 경우의 모의실험 결과중 하나이다. 이상치가 없는 경우 세 가지 방법 모두 큰 차이가 없는 것으로 보이며 모형오차의 상자그림인 Figure 3.4에서도 동일한 결론을 내릴 수 있다. 모형오차의 상자그림인 Figure 3.4에서도 세 가지 방법은 큰 차이가 없음을 알 수 있다. 각 모의실험에서 모형오차값에 의하여 세 방법들을 비교할 때 제안된 방법과 능형회귀의 모형오차값이 같았던 경우는 100번 중 95번이었고, 제안된 방법은 100번의 실험 중 2번을 한 개의 정상값을 이상치로 선택하였다.

이상의 모의실험로 부터 이상치가 있는 경우 다른 두 방법에 비해 제안된 방법의 적합결과가 더 좋으며 이상치가 없는 경우 세 방법에 큰 차이가 없음을 확인할 수 있었다.

3.2. 예제

제안한 방법의 효율성을 검증하기 위하여 사용하는 실제자료는 Balloon 자료와 Liver surgery 자료이다. Balloon 자료는 시간별(X)로 태양으로 부터의 방사선량을 측정한 데이터 (Davies와 Gather, 1993; Kovar와 Silverman, 2000)이다. 총 데이터의 개수는 4984개 이지만 대략 1/10정도인 500개로 실험을 하였다.

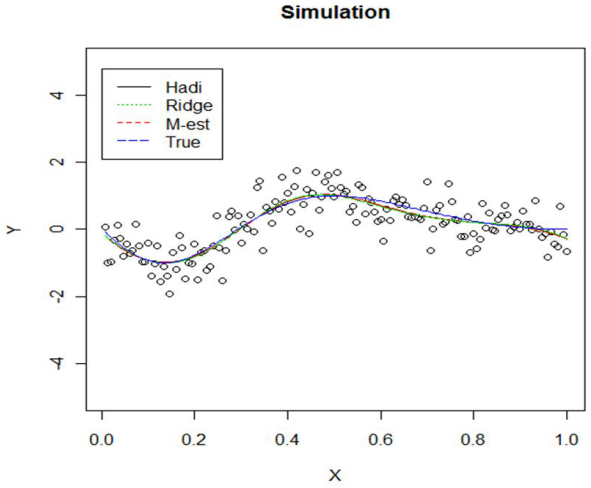

Figure 3.3. A simulated data set without outlier: Fitted curves and a curve representing the true model

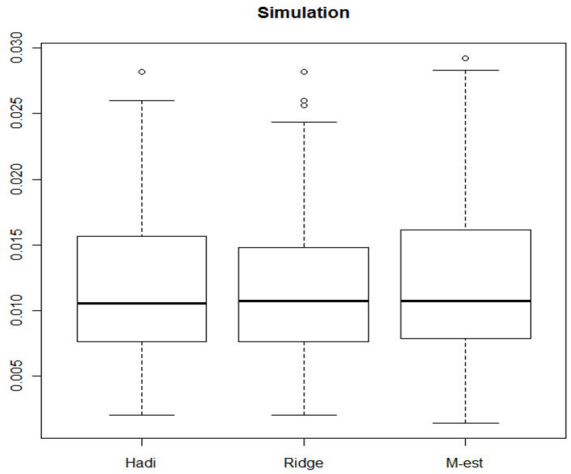

Figure 3.4. Box-plots of model-errors estimated from simulated data sets without outlier

λ는 $(0.1, 1, 10, 100)$값을 각각 사용하였으며 knot 간격은 $(1/10, 1/20, 1/30, 1/40, 1/50)$의 값을 등간격으로 각각 설정하고 그 중 GCV에 의해 최적값을 선택하였다. 각 방법의 적용 결과를 나타내는 Figure 3.5에서 별모양으로 표시된 점이 제안된 방법에 의해 이상치로 선택된 것이다. 데이터의 크기가 큼에도 불구하고 Figure 3.5에서 보듯이 이상치의 영향을 받은 일반적인 벌점 스플라인의 적합선은 다른 방법에 의한 적합선 보다 낮게 추정되며, 제안된 방법과 강건 벌점 스플라인 회귀방법의 적합선은 비슷하게 추정된다.

Liver surgery 자료는 특정유형의 간 수술을 받은 108명의 환자의 생존을 예측하는 데이터 (Neter 등, 1990)이다. 원래의 설명변수(X)는 네 개(X_1: blood clotting score, X_2: prognostic index, X_3: enzyme function score, X_4: liver function test score)이고, 각 환자의 수술 전 평가기록에서 얻은 것이며 반응 변수(Y)는 환자의 생존 시간이다. 간 수술 데이터중 1번부터 54번까지의 관찰값을 이용하여 원래

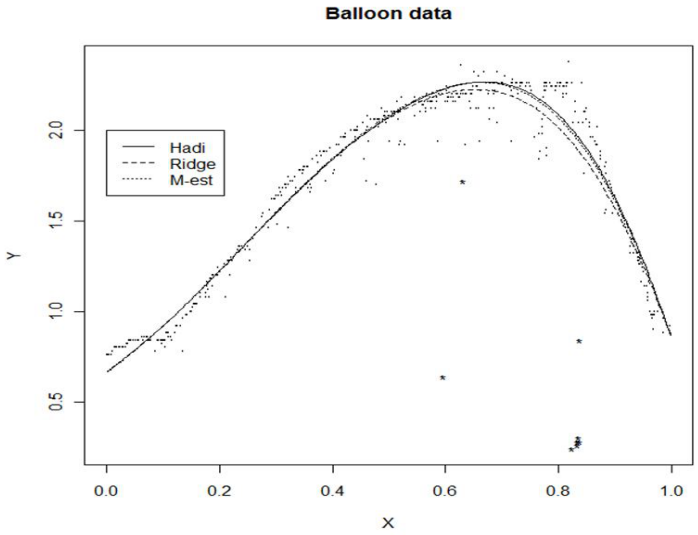

Figure 3.5. Ballon data: Fitted curves

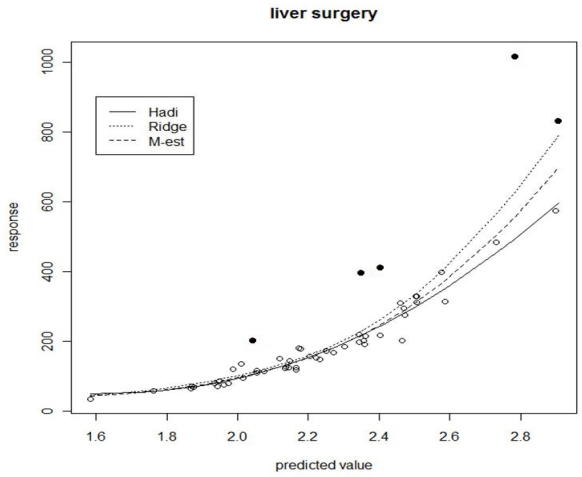

Figure 3.6. Liver surgery data: Fitted curves

의 반응변수 Y에 상용로그를 취한 $\log_{10} Y$를 네 개의 설명변수에 회귀분석을 실시하여 $\log_{10} \hat{Y}$를 추정한 후 $\log_{10} \hat{Y}$을 예측변수로, 원래의 반응변수 Y를 종속변수로 간주하여 모형추정을 시행하였다. 이 과정에서 1번부터 5번까지의 관찰 값들의 Y값은 실제값 보다 2배 증가된 값을 사용하여 이상치로 설정하였다. λ로 지정된 값은 $(0.1, 1, 10, 100)$이고, knot 간격은 $(1/5, 1/10, 1/20, 1/30)$ 등의 각각 등간격 값이며 GCV에 의해 최적값이 선택되었다.

Figure 3.6의 그림에서 검은색의 관찰치는 실제 이상치들이며 제안된 방법은 이들 값 이외에 13번 관찰값을 이상치로 탐지하였다. 일반적인 벌점 스플라인 회귀방법은 이상치의 영향을 많이 받는 것을 알 수 있고 강건 벌점 스플라인 회귀방법도 어느 정도 이상치의 영향을 받는 것으로 확인된다.

4. 결론

본 논문에서는 벌점 스플라인 회귀모형에서 이상치를 탐지하는 방법을 제안하고, 강건 벌점 스플라인 회귀방법, 일반적인 능형회귀모형과 비교하였다. 논문에서 제안한 제안된 방법은 이상치로 판단되는 관찰값을 제거하고 모형적합을 하는 반면에 강건 벌점 스플라인 회귀방법은 이상치를 제거하지 않고 강건한 모형적합을 위하여 M-추정법을 사용한다. 모의실험에서 한 가지의 이상치 모형을 사용했다는 점에서 제한적인 결론이라고 할 수 있지만, 모의실험과 예제의 결과에 따르면, 데이터의 숫자가 많을 경우에는 제안된 방법과 강건 스플라인 회귀방법에 큰 차이가 없으나 적정한 크기의 데이터의 경우 강건 벌점 스플라인 회귀방법보다 제안된 방법인, 이상치를 탐지한 후 벌점 스플라인 회귀모형을 사용하는것이 더 효율적인 것을 확인할 수 있었다. 다만 강건 스플라인 회귀방법에 비해 제안된 방법은 반복적인 검정을 수행하여 상대적으로 계산시간이 길다는 단점이 있으므로 예제에서 사용한 Balloon 자료 데이터 크기인 $n = 500$을 기준으로 두 방법의 적용 여부를 선택하는 것이 현실적이라고 할 수 있다.

References

Cantoni, E. and Ronchetti, E. (2001). Resistant selection of the smoothing parameter for smoothing splines, *Statistics and Computing*, **11**, 141–146.

Davies, L. and Gather, U. (1993). The identification of multiple outliers (with discussion), *Journal of the American Statistical Association*, **88**, 782–801.

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 Simonoff, J. S. (1993). Procedures for the Identification of Multiple Outliers in Linear Models, *Journal of the American Statistical Association*, **88**, 1264–1272.

Hoeting, J., Raftery, A. E. and Madigan, D. (1996). A method for simultaneous variable selection and outlier identification in linear regression, *Computational Statistics & Data Analysis*, **22**, 251–270.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo, *Annals of Statistics*, **1**, 799–821.

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.

Kovac, A. and Silverman, B. W. (2000). Extending the scope of wavelet regression methods by coefficient dependent thresholding, *Journal of the American Statistical Association*, **95**, 172–183.

Lee, T. C. M. and Oh, H.-S. (2007). Robust penalized regression spline fitting with application to additive mixed modeling, *Computational Statistics*, **22**, 159–171.

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

Neter, J., Wasserman, W. and Kutner, M. H. (1990). *Applied Linear Statistical Models (3rd ed.)*, Richard D. Irwin, Homewood.

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

Ruppert, D. and Wand, M. P. (2003). *Semiparametric Regression*, Cambridge University Press, Cambridge.

벌점 스플라인 회귀모형에서의 이상치 탐지방법

서한손[a] · 송지은[a] · 윤민[b,1]

[a]건국대학교 응용통계학과, [b]부경대학교 통계학과

(2013년 7월 15일 접수, 2013년 8월 13일 수정, 2013년 8월 21일 채택)

요 약

이상치가 존재하는 경우 모형 적합의 결과가 왜곡될 수 있기 때문에 이상치 탐색은 데이터분석에 있어서 매우 중요하다. 이상치 탐지 방법은 많은 학자들에 의해 연구되어 왔다. 본 논문에서는 Hadi와 Simonoff (1993)가 제안한 직접적 이상치 탐지 방법을 벌점 스플라인 회귀모형에 적용하여 이상치를 탐지하는 과정을 제안하며 모의실험과 실제 데이터에 적용을 통하여 스플라인 회귀모형, 강건 벌점 스플라인 회귀모형과 효율성을 비교한다.

주요용어: 강건 방법, 벌점 스플라인 회귀, 이상치.

[1]교신저자: (608-737) 부산신 용서로 남구 45, 부경대학교 통계학과, 부교수. E-mail: myoon@pknu.ar.kr