The Korean Communications
in Statistics Vol. 7, No. 3, 2000
pp. 633-641

Outlier Identification in Regression Analysis using Projection Pursuit

Hyojung Kim[1] and Chongsun Park[2]

Abstract

In this paper, we propose a method to identify multiple outliers in regression analysis with only assumption of smoothness on the regression function. Our method uses single-linkage clustering algorithm and Projection Pursuit Regression(PPR). It was compared with existing methods using several simulated and real examples and turned out to be very useful in regression problem with the regression function which is far from linear.

Keywords : Outliers, Regression analysis, Projection pursuit

1. 서 론

회귀분석에서의 이상값의 존재 여부는 분석의 결과에 많은 영향을 미치며 이러한 이유로 회귀분석에서의 이상값 식별에 대한 많은 연구가 이루어져 왔다. 하나의 이상값이 있을 경우에는 적절하게 이상값을 식별할 수 있는 방법들이 개발되었으나, 실제에 있어서는 하나의 이상값만이 존재한다고 할 수 없다. 경우에 따라 다중 이상값을 적절하게 식별할 수 있는 경우도 있지만, 대부분의 경우 인접해 있는 관측값의 영향에 의해 이상값을 이상값이 아닌 값으로 식별하거나, 이상값이 아닌 값을 이상값으로 식별하는 문제가 발생하게 된다. 전자의 경우를 가장효과(masking effect)라 하며, 후자의 경우를 편승효과(swamping effect)라 한다. 이처럼 둘 이상의 이상값이 존재하는 경우 발생하는 문제들을 해결하기 위한 여러 연구들에 대하여 살펴보기로 하자.

선형 회귀분석에서의 다중 이상값 식별방법들의 대부분은 주어진 자료를 이상값이 포함되지 않은 순수집합(clean set)과 비-순수집합으로 구분하는 방법에 핵심을 두고 있다. 순수집합에 대한 정의는 Gentleman과 Wilk (1975)에 의해 처음 제시되었으며, Hawkins와 Bradu 그리고 Kass (1984)는 이러한 순수집합으로 이루어진 기본집합을 이용하여 정규 오차 구조를 갖는 선형 모형 자료에서 이상값들을 식별하는 문제를 고려하였다. 또한 중앙값과 같은 로버스트한 위치 추정량들을 이용하여 가장효과와 편승효과를 최소화하는 방법들이 제시되었으며, Bradu와 Hawkins (1991)

1) Graduate student, Department of Statistics, Sungkyunkwan University, 110-745, Korea.
E-mail : nomad2000@orgio.net
2) Associate Professor, Department of Statistics, Sungkyunkwan Univeristy, 110-745, Korea.
E-mail : cspark@skku.ac.kr

는 Rousseeuw (1984)의 최소 중앙값 제곱을 이용하는 방법에 기본집합의 개념을 적용하였다. 또한 Hadi와 Simonoff (1993)등에 의해 제시된 단일 이상값(single outlier)식별을 이용하는 방법, Krasker와 Welsch (1982)의 로버스트 회귀 추정 방법 등이 있다. Sebert 등 (1998)은 선형회귀에서 얻어진 예측값과 잔차에 단일 연결 군집방법을 적용하여 이상값을 식별하였다.

Marasinghe (1985)는 기존의 방법들을 여러 번 적용하는 다단계(multi-stage) 접근 방법을 제시하였는데, 이 방법은 너무 많은 계산을 요구하는 단점이 있다. 이 외에도 Paul과 Fung (1991)은 2단계 방법을 제시하였다.

비-선형 회귀분석에 있어서의 다중 이상값 식별문제는 Manski (1984)에 의하여 시작되었으며, Hardle과 Gasser (1985)가 이상값에 영향을 받지 않는 핵-추정량(kernel estimator)을 제시하였다. Naes (1986)는 고정 성분과 임의 성분을 갖는 선형 혼합 모형에 대한 이상값 식별을 고려하였으며, Gasser, Sroka 그리고 Jennen-Steinmetz (1986) 등은 관심있는 관측값들의 한쪽에 있는 두 점과 적합 직선까지의 잔차인 유사잔차(pseudo-residual)를 이용하여 이상값들과 이-분산을 검정하였다. Simonoff와 Tsai (1986)는 일반적인 로버스트 척도와 이상값에 관심을 가지고 표본 재-사용 접근법(구체적으로 재크나이프 접근법)을 제시하였다. 또한 일반화 선형 모형, 그 중에서도 특히 로지스틱 모형에 대한 많은 연구들이 이루어져 왔으며, 여기에는 Jennings (1986), Follman과 Lambert (1989), Bedrick과 Hill (1990) 등의 연구가 있다.

그러나 기존의 모든 연구들은 선형 또는 비-선형 회귀함수를 사전에 가정하고 이상값을 식별하는 방법들이며 따라서 이상값이란 가정된 회귀함수에서 상대적으로 멀리 떨어져 있는 관측값들을 의미하고 있다.

본 논문에서는 회귀함수의 형태에 대한 가정을 하지 않고 단지 평활한 함수라는 가정하에서 이상값들의 식별 가능성을 고려하였다. 회귀함수의 구체적인 형태에 대한 가정, 예를 들면 선형함수나 정준 연결함수(canonical link function)를 갖는 포아송 모형 등의 가정이 없이 단지 평활하다는 가정하에서의 이상값이란 평활한 진실회귀함수에서 상대적으로 멀리 떨어진 관측값들을 의미한다고 할 수 있다.

먼저 제2절에서는 이상값에 대한 기존의 정의를 살펴보고 본 연구에서 사용된 정의를 다루었으며, 제3절에서는 본 논문에서 고려하고 있는 모형과 알고리즘에 대해 살펴보았다. 마지막으로 제4절에서는 본 논문에서 고려한 방법들의 효율성을 모의자료와 실제자료들을 통하여 살펴보았으며 결론을 제5절에 포함하였다.

2. 모형 및 이상값에 대한 정의

서론에서 살펴본 바와 같이 회귀분석에서 이상값 식별에 관한 연구들은 대부분 사전에 선형 또는 비-선형 회귀함수 또는 회귀모형을 가정하고 이 모형이나 회귀함수의 적합을 통하여 구해진 잔차들을 이용하여 이상값을 식별한다. 따라서 이상값은 가정한 회귀함수에서 상대적으로 멀리 떨어진 관측값들 즉, 잔차가 큰 관측값들이라고 정의할 수 있다. 그러나 회귀함수의 형태에 대한 가정이 없는 경우에는 위와 같은 기존의 정의는 더 이상 무의미하게 된다. 우선 다음과 같은 회귀모형을 고려하자.

$$y = f(\beta^T x) + \varepsilon$$

여기서, f는 회귀함수이다. ε은 평균이 0이며, 분산 σ^2을 가진다. 또한 y는 반응변수 이고 x

는 $k\times1$인 설명 변수 벡터이다. 따라서 β는 $k\times1$인 미지의 모수 벡터가 된다. 각 각의 관측치는 아래첨자 i를 사용하여 표기하며 표본수는 n이라고 가정하자. 이 때 이상값을 정의하기 위하여 회귀모형에서 필요한 가정은 다음과 같다.

첫째, 회귀함수 f는 평활(smooth)하다.
둘째, 이상값의 수는 정상값의 수와 비교했을 때 상대적으로 적다.

첫 번째 제약은 회귀함수에 대한 제약조건이다. 회귀함수를 사전에 알지 못한다는 가정하에서 회귀함수가 평활하지 않은 경우 이상값에 대한 정의는 사실상 불가능하게 된다. 따라서 회귀함수에 필요한 최소한의 가정은 평활성이 되며, 이 경우 각 관측값들은 평활한 경향(smooth trend)을 나타내는 회귀함수의 주변에서 관측될 것이다. 그러므로 이상값들은 평활한 회귀함수에서 벗어난 하나 이상의 관측값들로 생각할 수 있다. 평활한 회귀함수는 단절이 없이 매끄럽게 변하는 특성을 가지게 된다는 것을 고려하면, 함수를 급격하게 변하게 하는 범위에서 관측되는 관측값들을 이상값들로 생각할 수 있을 것이다. 본 논문에서의 이상값은 위와 같은 개념의 이상값을 사용할 것이다.

[정의] 이상값 - 평활한 진실 회귀함수에서 상대적으로 멀리 떨어진 관측값으로 회귀함수의 평활성을 해치는 관측값

두 번째 제약조건은 이상값의 크기에 관한 제약조건이다. 이상값의 수와 정상값의 수를 비교했을 때 이상값의 수가 정상값의 수보다 많다면, 회귀함수의 추정은 정상값에 대한 회귀함수 추정이 아닌, 이상값의 영향에 의한 이상값에 대한 회귀함수 추정이 될 것이다. 따라서 이상값의 수가 몇 개인지는 가정할 수 없지만 정상값에 비해 상대적으로 적다는 가정이 필요하며 이는 일반적인 이상값 식별방법들에서 대부분 가정하고 있는 것이다.

3. 알고리즘

위의 절에서 정의한 이상값들을 찾아내기 위한 알고리즘을 단계적으로 살펴보고 각 단계에 대한 구체적인 내용을 다루기로 한다.

단계1. 이상값이 없는 순수집합 M을 찾는다.
단계2. 사영추적회귀 (Projection Pursuit Regression - Friedman과 Stuetzle, 1981)를 이용하여 순수집합으로부터 모형에 필요한 설명변수들의 선형결합 $\hat{\beta}^T x$를 찾는다.
단계3. 단계2에서 구해진 $\hat{\beta}^T x$와 y에 가중 Loess (Cleveland, 1979)를 적용하여 다음과 같은 잔차를 계산하고 잔차의 절대값을 오름차순으로 나열한다.

$$d_i=\frac{y_i-\text{Loess}(\hat{\beta}^T x)}{\hat{\sigma}_M\sqrt{1-\text{diag}(S)}},\quad i=1,\ \ldots,\ n$$

$$\text{Loess의 가중값} = \begin{cases} 1, & i \in M \\ 0.001, & i \notin M \end{cases}$$

$$\widehat{\sigma_M}^2 = \frac{RSS_{\text{Loess}}}{n-1}$$

RSS_{Loess} : Loess의 잔차제곱합

S : 평활기 행렬, h : M의 크기

단계4. $(h+1)$ 번째 크기의 잔차가 임계값 $t_{(\alpha/2(h+1),\, h-k-\mathrm{tr}(S))}$보다 크다면 $(h+1)$ 번째 관측값을 포함한 나머지 관측값들을 이상값으로 식별하고, 작다면 순수집합에 $(h+1)$번째 관측값을 추가하여 단계2부터 다시 반복한다.

각 단계를 구체적으로 살펴보면,

단계1에서는 단일 연결 군집방법을 이용하여 순수 집합(clean set)과 비-순수 집합(non-clean set)을 찾게 된다. 우선 x변수들과 y변수를 모든 관측치에 대하여 함께 고려한 집합을 $Z = (X: Y)$라하고 이를 표준화한 것을 $W = Z\,\hat{\Sigma}^{-1}$라고 하자. 이 때 X는 $n \times k$ 행렬이고 Y는 $n \times 1$ 벡터가 되며 $\hat{\Sigma} = \mathrm{var}(Z)$이다. W에 대해서 단일 연결 군집방법을 이용하여 초기 순수 집합 M을 찾게 된다. 초기 순수 집합 M의 크기는 $(n+k-1)/2$ 의 정수 부분인 h이다. 이 방법은 Simonoff와 Hadi(1993)가 선형 회귀모형에 대한 이상값의 식별에서 사용한 방법이지만, 회귀모형에 대한 가정이 필요하지 않으므로 본 논문에서도 수정 없이 사용될 수 있다.

단계2에서는 단계1에서 찾아진 초기순수 집합을 이용하여 회귀함수에 필요한 선형결합을 사영추적회귀(Friedman과 Stuetzle, 1981)를 이용하여 찾게 된다. 사영추적회귀에서는 일반적으로 한 개 이상의 선형결합들(설명변수들의)이 모형에 포함되지만 본 논문에서는 하나의 선형결합만이 모형에 필요하므로 이것만을 추정하게 된다.

단계3에서는 단계2에서 찾아진 선형결합에 대한 국소 회귀모형(Local Regression Model : Loess-Cleveland, 1979)의 적합에 의해서 구한 예측값과 관측값 y_i의 차이를 구하고 표준화시킨다. 국소회귀모형은 주어진 설명변수의 관측값을 중심으로 가장 가까운 일정한 수의 관측값들에 가중회귀모형을 적용하는 과정을 모든 관측값에 반복 적용하여 비모수적 회귀선을 구하게 된다. 일반적으로 Loess를 포함하는 많은 종류의 평활기들은 적당한 행렬 S에 반응변수를 곱한 Sy의 형태를 갖게 되며 이는 일반적인 선형회귀의 $H=X(X^TX)^{-1}X^T$에 대응하는 것으로 생각할 수 있다. 표준화된 d_i는 선형회귀의 모자행렬 H를 평활기 행렬인 S로 대치하여 계산되었다.

일반적인 평활기(smoother)는 이상값들에 의해 급격하게 함수의 형태가 변경될 수 있다. 함수형태의 급격한 변경은 이상값을 적절하게 식별하지 못하게 하는 단점을 가진다. 이러한 평활기의 단점은 순수 집합과 비-순수집합에 다른 가중값을 주어 해결하였다. 즉 순수집합에 포함된 관측값에 대해서는 큰 가중값을 주고, 비-순수집합에 대해서는 상대적으로 작은 가중값을 주게 된다. 가중값은 비-순수집합에 포함되어 있을 지도 모르는 이상값에 의하여 회귀함수의 추정함수인 평활함수(smooth function)가 급격히 변화하는 것을 방지하게 되며, 동시에 모든 설명변수의 범위에 대한 예측값(prediction value)을 제공하게 된다.

마지막으로 단계4에서는 단계3에서 구해진 d_i들을 t분포의 임계값과 비교하여 이상값을 식별하게 된다. d_i들은 평활기의 편의(bias)가 무시할 정도로 작은 경우 점근적으로 t분포를 따르는 것으로 알려져 있으며 (Härdle, 1990) t분포의 자유도는 일반적인 평활기의 자유도인 $\mathrm{tr}(S)$와 설명변수의 수를 고려하여 계산하였다. 또한 마지막 단계에서 순수 집합의 크기가 관측값의 크기 n과 같게 되면 이상값이 없는 것으로 판단한다.

3.1 가중값과 대역폭의 결정

단계3에서 사용되는 국소 회귀모형(Loess)에서 고려되어야할 중요한 문제는 가중값과 대역폭(bandwidth)의 문제이다.

우선 가중값은 진실 회귀함수 추정에 이상값들이 영향을 미치지 않으면서 모든 관측값들에 대한 예측값을 제공하도록 결정되어야 한다. 따라서 가중값은 평활법에 의한 함수 추정에 영향을 줄 수도 있는 비-순수집합의 관측값들의 영향을 최소화하는 방향으로 즉, 비-순수집합에 포함되어 있을 지도 모르는 이상값들에 의해 평활 함수가 급격하게 변화하는 것을 방지하는 방향으로 결정되어야 한다. 또한 평활법들은 함수의 추정에 사용되는 관측값들의 영역에서는 관측값들에 대한 예측값은 제공해 주지만, 영역 밖에서는 관측값에 대한 예측값을 제공해주지 않는다. 그러나 순수집합과 비-순수집합의 모든 관측값에 대한 잔차를 구하기 위해서는 모든 관측값들에 대한 예측값이 필요하다. 간단한 모의실험 결과, 순수집합에 포함되는 관측값들에 대한 가중값을 1로 하였을 경우 비-순수집합에 포함된 관측값들에 대한 가중값이 0.01이하이면 위의 두 가지 조건을 충족시키는 것으로 나타났다. 본 논문의 예제들에서는 비-순수집합에 대한 가중값을 0.001로 고정하여 사용하였으나, 0.01이하의 어떠한 값을 사용하여도 결과는 비슷하게 된다.

일반적으로 적절한 대역폭의 결정은 표본수, 진실 회귀함수의 형태 등에 영향을 받는 것으로 알려져 있다. 회귀함수 추정을 위한 대역폭을 어떻게 설정하느냐에 따라 추정되는 회귀함수의 형태가 달라지게 된다. 이상값을 식별하기 위해 사용되는 평활법에 의해 추정되는 회귀함수가 진실 회귀함수라고 생각하면, 추정 회귀함수가 매끄럽고 평활한 형태를 보이는 대역폭을 설정하는 것이 적절할 것이다. 대역폭의 변화에도 불구하고 추정 회귀함수의 형태가 매끄럽고 평활한 형태를 보이게 되면, 식별되는 이상값들은 큰 차이를 보이지 않는 것을 살펴볼 수 있다. 따라서 변화 범위 내의 임의의 대역폭을 사용하여도 이상값은 적절하게 식별될 것이다. 본 논문에서는 대역폭으로 전체 자료에 대한 비율(span)을 사용하였다. 따라서 대역폭이 1에 가까울수록 선형인 회귀함수의 형태를 가지게 된다. 대역폭은 자료에 따라 결정되었으며 자세한 내용은 각 예제들에 따라 다시 언급하기로 한다.

4. 모의자료 및 실제자료 분석

본 장에서는 회귀분석에서의 이상값 식별을 위하여 본 논문에서 제시된 방법과 기존의 LMS(Least Median of Squares) 방법, HS(Hadi와 Simonoff) 방법을 모의자료와 실제자료를 통하여 비교하여 보기로 한다. 분석에는 Hadi와 Simonoff가 S-plus를 이용하여 작성한 프로그램을 본 논문의 방법에 맞도록 수정하여 사용하였다.

4.1 모의자료 1

모의자료 1에서는 선형회귀모형의 가정이 타당한 경우에 기존의 방법들과의 비교를 위하여 Hawkins, Bradu 그리고 Kass (1984, Table 4)의 가상 선형자료를 사용하였다. 이 자료에는 가장 효과를 나타내는 10개의 이상값이 포함되어 있으며 이상값이 아닌 4개의 자료는 편승효과를 나타내도록 되어 있어 M-추정법, 축차제거 (sequential deletion) 방법 등 많은 방법들이 관측치 1-10의 이상값을 식별하는 데 실패하는 것으로 알려져 있다.

이 자료에 포함되어 있는 10개의 이상값을 성공적으로 찾는 방법들로는 Hawkins, Bradu 그리고 Kass가 제시한 방법, HS 방법, LMS 방법 등이며 본 논문에서 제시한 방법 (가중치는 0.001을 대역폭은 0.6) 또한 10개의 이상값을 정확히 구별하였다.

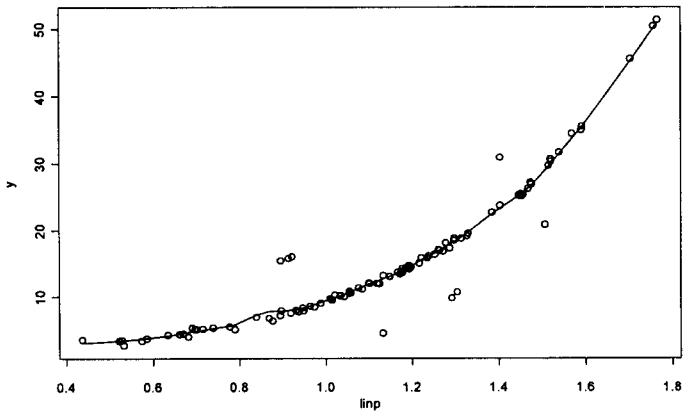
4.2 모의자료 2

모의자료 2에서는 비-선형모형의 경우로 사용한 모형은

$$y = \exp(x_1 + x_2 + x_3 + x_4 + x_5) + \varepsilon$$

이며, $x_i \sim U(0, 1)$, $\varepsilon \sim N(0, (0.4)^2)$ 이며 100개의 관측치를 모의추출 하였다. 그리고 임의로 17, 20, 33, 40, 46, 77, 82, 83번째 값들에 8을 가감하여 이상값을 생성시켰다. <그림 1>은 추정된 설명변수들의 선형결합값과 반응변수값들의 산점도에 제시된 방법에 의해서 추정된 회귀함수를 포함하고 있다. 이상값들은 곡선의 회귀선에서 상대적으로 떨어진 8개의 관측값으로 그림에서 쉽게 판단할 수 있다.

LMS 방법은 8개의 이상값 이외에 20개의 이상값이 아닌 관측값들을 이상값으로 식별하고, HS 방법 또한 28개의 관측값을 이상값으로 식별하였다. 제시된 방법은 17, 20, 33, 37, 40, 46, 77, 82, 83번째 관측값을 이상값으로 식별하였다. LMS 방법과 HS 방법은 이상값을 과대 식별하였으며 이는 선형모형의 가정이 타당하지 않은 자료에 선형모형을 적합시킨 결과로 생각된다. 제시된 방법은 37번째 관측값 만을 잘못 식별하였으며 대역폭은 0.3을 사용하였다.



<그림 1> 수평축: 추정된 선형결합값, 수직축: 반응변수값 곡선: 추정회귀선

4.3 실제자료

사용된 자료는 Brownlee의 stack-loss자료이다. M추정법과 축차 방법은 관측값 1, 3, 4, 21을 이상값으로 판단하였다 (Andrews, 1974; Hawkins, 1980). 반면, LMS 방법은 관측값 1, 2, 3, 4, 21을 HS 방법은 관측값 1, 3, 4, 21을 이상값으로 판단하였다. 본 논문에서 제시된 방법은 대역폭에 따라 결과에 차이가 있다. 우선 대역폭이 0.5-0.7인 경우에는 관측값 4, 21을 0.8인 경우는 관측값 3, 4, 21을 그리고 0.9인 경우에는 LMS 방법과 같이 관측값 1, 2, 3, 4, 21을 이상값으로 판단하였다.

제시된 방법은 대역폭에 따라 회귀선이 변하게 되고 그 값이 작은 경우에는 굴곡이 상대적으로 심한 곡선이 되어 이상값의 수가 상대적으로 적어지게 되며 값이 커질수록 직선에 가까워지게 되며 이상값의 수도 많아지게 된다.

5. 결 론

기존의 이상값 식별방법들은 회귀 모형을 가정하고, 가정된 모형하에서의 추정회귀함수의 잔차들을 이용하여 이상값들을 식별하였다. 따라서, 모형에 대한 가정이 적절하지 않게 되면 효율적으로 이상값을 식별할 수 없게 된다.

제시된 방법은 사영추적회귀와 평활법에 의해 추정된 회귀함수를 이용하여 이상값을 식별하고자 하였다. 따라서 회귀 모형에 대한 가정에서 발생할 수 있는 문제를 최소화할 수 있었다. 그러나 평활법을 사용하기 때문에 가중값과 대역폭의 선택이 문제가 된다. 우리는 이러한 문제들을 주어진 대역폭에 따라 추정된 회귀함수가 진실 회귀함수라고 가정하고, 가중값은 비-순수집합이 추정된 회귀함수에 미치는 영향을 최소화하면서 모든 관측값에 대한 예측값를 제공하도록 결정하였다.

일반적인 사영추적회귀는 2개 이상의 선형결합을 포함하는 모형에 적용할 수 있으나 본 논문에서는 하나의 선형결합이 필요한 경우만 고려하였으나 몇가지의 문제점을 해결한다면 하나 이상의 선형결합을 갖는 경우로 확장할 수 있다. 따라서 두 개 이상의 선형결합이 모형에 필요한 경우에 대한 연구가 필요할 것으로 생각된다. 또한 앞에서 구해진 d_i들의 절대값의 크기와 모형에 미치는 영향사이의 관계를 알 수 있다면 영향점(leverage point)들을 식별하는 데 사용될 수 있을 것이다.

참 고 문 헌

[1] Andrews, D.F. (1974), A Robust Method for Multiple Linear Regression, *Technometrics*, Vol. 16, 523-531.

[2] Barnett, V., and Lewis, T. (1994), *Outliers in Statistical Data,* John Wiley & Sons, New York.

[3] Bedrick, E.J., and Hill, J.R. (1990), Outlier Tests for Logistic Regression: A Conditional Approach, *Biometrika*, Vol. 77, 815-827.

[4] Bradu, D., and Hawkins, D.M. (1991), Sample Size Requirements for Multiple Outlier Location Techniques Based on Elemental Sets, *Research Report*, 1-17.

[5] Cleveland, W.S. (1979), Locally Weighted Regression and Smoothing Scatterplots, *Journal of the American Statistical Association*, Vol. 74, 829-836.

[6] Follmann, D.A., and Lambert, D. (1989), Generalizing Logistic Regression by Nonparametric Mixing, *Journal of the American Statistical Association*, Vol. 84, 295-300.

[7] Friedman, J.H., and Stuetzle, W. (1981), Projection Pursuit Regression, *Journal of the American Statistical Association,* Vol. 76, 817-823.

[8] Gasser, T., Sroka, L., and Jennen-Steinmetz, C. (1986), Residual Variance and Residual Pattern in Nonlinear Regression and for the Detection of Outlier, *Biometrika,* Vol. 73, 625-633.

[9] Gentleman, J.F. and Wilk, M.B. (1975), Detecting Outliers: II Supplementing the Direct Analysis of Residuals, *Biometrics,* Vol. 31, 387-410.

[10] Hadi, A.S., and Simonoff, J.S. (1993), Procedures for the Identification of Multiple Outliers in Linear Models, *Journal of the American Statistical Association,* Vol. 75, 1264-1272.

[11] Härdle, W. (1990), Smoothing Techniques With Implementation in S.

[12] Härdle, W., and Gasser, T. (1985), On Robust Kernel Estimation of Derivatives of Regression Functions, *Scandinavian Journal of Statistics- Theory and Applications,* Vol. 12, 233-240.

[13] Hastie, T., and Tibshirani, R. (1990), *Generalized Additive Models*, Chapman & Hall, London.

[14] Hawkins, D.M. (1980), *Identification of Outliers*, London: Chapman and Hall.

[15] Hawkins, D.M., Bradu, D., and Kass, G.V. (1984), Location of Several Outliers in Multiple-Regression Data Using Elemental Sets, *Technometrics*, Vol. 26, 197-208.

[16] Jennings, D.E. (1986), Outliers and Residual Distributions in Logistic Regression, *Journal of the American Statistical Association*, Vol. 81, 987-990.

[17] Krasker, W.S., and Welsch, R.E. (1982), Efficient Bounded-Influence Regression Estimation, *Journal of the American Statistical Association,* Vol. 77, 595-604.

[18] Manski, C.F. (1987), Adaptive Estimation of Nonlinear Regression Models, *Economic Review*, 3, 145-210.

[19] Marasinghe, M.G. (1985), A Multistage Procedure for Detecting Several Outliers in Linear Regression, *Technometrics,* Vol. 27, 395-399.

[20] Naes, T. (1986), Detection of Multivariate Outliers in Linear Mixed Models, *Communications in Statistics- Theory and Methods,* Vol. 15, 33-47.

[21] Paul, S.R., and Fung, K.Y. (1991), A Generalized Extreme Studentized Residual Multiple Outlier Detection Procedure in Linear Regression, *Technometrics,* Vol. 33, 339-348.

[22] Rousseeuw, P.J. (1984), Least Median of Squares Regression, *Journal of the American Statistical Association*, Vol. 79, 871-880.

[23] Sebert, D.M., Montgomery, D.C., and Rollier, D.A. (1998), A Clustering Algorithm for Identifying Multiple Outliers in Linear Regression, *Computational Statistics & Data Analysis,* Vol. 27, 461-484.

[24] Sockett, E.B., Daneman, D., Clarson, C., and Erich, R.M. (1987), Factors Affecting and Patterns of Residual Insulin Secretion During the First Year of Type I (Insulin Dependent) Diabetes Mellitus in Children, *Diabet*, Vol. 30, 453-459.