응용통계연구 제10권 2호
1997년 9월 pp. 241-251

동적그래픽스에 의한 회귀진단시스템(REDS)의 구현*

유 종 영[1], 안 기 수[2], 허 문 열[3]

요 약

기존의 회귀진단 방법은 자료의 구조를 변화시키거나 회귀모형의 형식을 변화시킬 때 이것이 잔차에 미치는 영향을 분석하는 것이 주를 이루었다. 그러나 역으로 잔차를 변화시킬 때 이것이 회귀모형에 미치는 영향을 분석하는 것을 생각할 수 있다. 이것은 현실적으로 몇 개의 특정한 자료에 더욱 가중치를 부여하여 회귀모형을 만들거나 또는 잔차의 패턴을 정상화하고자 할 때 유용한 방법이 될 수 있다. 본 연구팀은 기존의 회귀진단방법과 더불어 잔차패턴을 변화시킴으로서 회귀진단을 실시하는 방법을 동적그라픽스기법에 의해 회귀진단시스템(REgression Diagnostics System - REDS)으로 구현하였으며 본 논문을 통해 이를 소개하고자 한다.

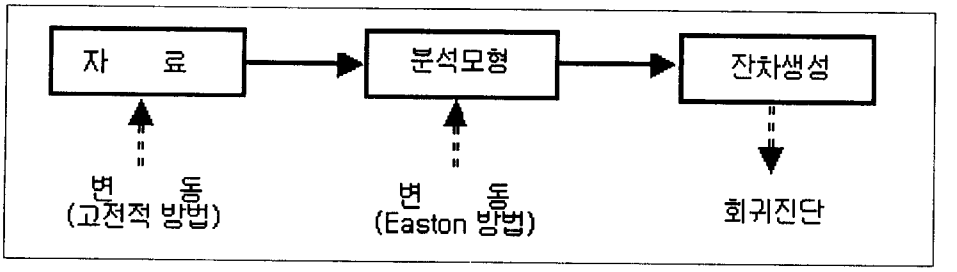
1. 회귀진단 방법에 대한 고찰

회귀진단이라고 하면 자료의 변동이 모형에 미치는 영향분석과 회귀계수의 변동이 잔차에 미치는 영향분석을 주로 다루고 있다. 자료의 변동에 의한 회귀진단은 자료를 1개 혹은 그 이상을 삭제 또는 변동시켰을 때 회귀직선의 변화폭과 잔차의 변동을 탐색하므로서 이상값을 식별하는 것을 의미한다. 회귀진단 통계량으로는 최소제곱법에 의한 잔차와 표준화된 잔차, Belsley 등이 제안한 표준화된 잔차, Cook 통계량 및 리버리지값, 잔차에 대한 상자도형 등이 있다. 그러나 이러한 통계량은 다중이상값이 존재하는 경우 가장효과(masking effect)와 편승효과(swamping effect)에 영향을 받는 것으로 분석되어 왔다(Rousseeuw와 Leory (1987)).

Easton(1994)이 제안한 회귀계수 변동에 의한 회귀진단방법은 선형회귀에서의 회귀계수 $\boldsymbol{b}$가 $N(\boldsymbol{\beta}, (\mathbf{X}'\mathbf{X})^{-1}\sigma^2)$에 따른다는 가정 하에서 회귀계수의 99% 신뢰구간을 산정하고, 이 구간 안에서 회귀계수를 변동시켜가면서 새로운 모형을 만들어 이상값을 탐색하는 방법이다. 이 방법에 의하면 설명변수가 p개인 경우 정규분포의 가정 하에서 각 회귀계수의 99%의 신뢰구간 $(b_1 \pm \delta_1, b_2 \pm \delta_2, \ldots b_p \pm \delta_p)$을 계산하고, 이 구간 안에서 잔차의 MAD(Median Absolute Deviation)를 최소로 하는 회귀모형을 탐색하여 이상값을 식별하게 된다. <그림 1>은 자료의 변동 및 회귀계수 변동에 의한 회귀진단 방법의 흐름도이다.

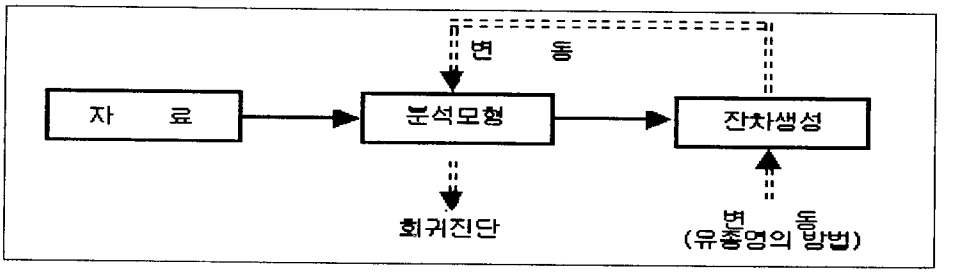
1) (449-714) 경기도 용인시 삼가동 117-6 용인대학교 전산통계학과, 조교수
2) (440-714) 경기도 수원시 장안구 동남보건전문대학 사무자동학과, 전임강사
3) (110-745) 서울시 종로구 명륜동 3가 53번지 성균관대학교 통계학과, 교수

* 본 논문의 프로그램이 필요한 경우 http://stat.skku.ac.kr/research/programs/ REDS.html을 접속하면 된다.



<그림 1> 일반적인 회귀진단방법의 흐름도

<그림 1>과 같은 회귀진단 방법과는 달리 유종영과 김현철(1996)은 잔차의 크기를 변동시킬 때 이것이 모형에 미치는 영향을 분석하는 방법을 제안하였다(<그림 2> 참조). 잔차의 크기를 일정한 신뢰구간안에서 변동시키는 것은 여러 가지 통계적 의미를 지니고 있다. 예를 들어 연도별 경제성장률을 가지고 회귀분석을 할 경우 과거의 경제성장률에 대한 정보와 최근의 성장률의 정보가 같은 가중값을 갖게 된다. 그러나 과거에는 경제가 급격히 성장하였고 최근에는 완만한 추세를 보였다면 향후 예측값은 과거의 급격한 성장률 자료에 의해 과대 계산되는 오류를 보이게 된다. 이러한 경우 최근 년도의 잔차의 크기를 일정한 범위 안에서 제어하므로서 최근 자료에 더욱 가중값을 부여할 수 있다. 또한 특정한 년도들에 더 관심이 있는 경우 이 년도들의 잔차를 제어할 필요성이 있다. 이와 같이 잔차의 크기를 제어하는 방법은 현실적인 자료분석에서 매우 유용하며 가중회귀분석과 맥을 같이 한다.



<그림 2> 잔차의 변동에 의한 회귀진단방법의 흐름

통계적 자료분석에서 동적그라픽스의 적용은 Tukey 등(1973)에 의해 이루어진 PRIM-9이 시발점이 되었다. 이들은 다차원자료의 구조를 파악하는데 투사, 회전, 분리 등의 동적그라픽스 방법을 적용하여 9차원까지의 자료를 그라프로 표현하였고, Huh (1995)는 꺾은누적분포함수(Flipped Empirical Distribution Function)의 개념을 제시하여 자료를 분석하였다. 이러한 동적그라픽스에 의한 통계적 자료분석은 S-PLUS 통계팩키지의 빗질(Brushing)을 이용하여 다차원 자료를 분석하는 방법과 Cook과 Weisberg(1994)에서 예시된 자료의 변동과 동시에 회귀직선이 움직이는 방법 등을 예로 볼 수 있다.

본 연구에서는 자료변동, 회귀계수변동, 잔차변동 등의 회귀진단방법을 동적그라픽스 기법으로 처리할수 있는 회귀진단시스템(REDS)을 구현하였으며 2절에서는 REDS의 구현을 설명하고, 3절에서는 REDS를 사용하여 실제자료를 처리하는 예를 보여주고 있다.

2. REDS의 구조

이 절에서는 자료의 변동에 의한 회귀진단과 Easton의 회귀계수 변동에 의한 회귀진단, 유종영과 김현철의 잔차변동에 의한 회귀진단 등을 동적그래픽스 방법을 이용하여 REDS를 설명하고자 한다. REDS를 실행하면 먼저 <그림 3>과 같이 주창이 나타나고 필요에 따라 주창의 메뉴를 선택하면 보조창들이 나타나게 된다. 주창의 화면과 주창의 메뉴에 대해 살펴보면 다음과 같다.

2.1 주창의 화면

주창의 화면은 잔차그라프, LMS(Least Median of Squares Regression)에 의한 잔차그라프, 추정값과 반응값의 그라프, 꺽은누적분포함수와 같은 4가지의 그라프, 그리고 4개의 메뉴단추, MAD, SD, LMS표준편차의 추정값을 나타내는 MED 등의 진단통계량막대, 회귀계수를 나타내는 회귀계수막대 등의 정보를 갖고 있다. 각 부분에 대해 설명하면 다음과 같다.

(a) 잔차그라프
횡축은 자료의 순서를 나타내고 종축은 잔차를 나타낸다.

(b) LMS에 의한 잔차그라프
로버스트 추정법중에서 이상값에 로버스트한 추정법의 하나가 LMS로 이 그라프는 LMS 표 준편차 (σ_{LMS}))의 추정값을 구하여 잔차(e_i)를 나눈 값을 그린 것으로, Rousseeuw와 Leory는 $|e_i/\hat{\sigma}_{LMS}| \geq 2.5$인 경우 이상값으로 판단하였다. 따라서 REDS에서는 LMS의 잔차그라프에 $e_i/\hat{\sigma}_{LMS} = \pm 2.5$ 선을 그어 이상값의 판단기준을 제공한다.

(c) 추정값($\hat{\mathbf{y}}$) 및 반응값($\mathbf{y}$)의 그라프
그라프에서 반응값은 실선, 추정값은 " · "으로 표시되어 있다.

(d) 꺽은누적분포함수(Flipped Empirical Distribution Function)
Huh는 EDF를 수정하여 다음과 같은 꺽은누적분포함수(FEDF)를 제안하였다.

$$\text{FEDF} = \begin{cases} \text{EDF}, & \text{if } x \leq \text{median} \\ 1-\text{EDF}, & \text{if } x > \text{median} \end{cases}$$

단, $\text{EDF} = (i-1/3)/(n+1/3),\ i = 1, 2, \ldots, n$

(i 는 i 번째의 순서화된 번호를 말하고 n은 표본의 크기임)

EDF는 단순증가함수로서 단순한 구조를 가지고 있으며 순서통계량의 함수이므로 여러 가지 효율적인 통계적 특성들을 지니고 있다. 따라서 FEDF도 여러 가지 효율적인 통계적 특징들을 갖게 된다. FEDF를 이용하면 그라프로부터 해당변수의 분포형태가 대칭인가를 즉시에 판단할 수 있고 간단한 통계량을 직접 얻을 수도 있으며 동적그래픽스에서 매우 유용하게 이용된다.

(e) 메뉴 단추

위에서 설명한 그라프들을 동적그래픽스로 효율성 있게 나타내는 기능으로 다음과 같은 4개의 메뉴 단추를 제공한다.

- 원시자료 : 이 메뉴를 택하면 모든 그라프와 분석 자료에 원시자료를 적용한다.
- 이 동 : 이는 잔차를 변동시키는 메뉴이다. 주창의 잔차그라프에서 특정한 잔차를 선택하여 움직이면 다른 그라프상에서 이 잔차에 해당하는 관측값이 동시에 변화되고 주창의 왼쪽에 나타나 있는 여러 가지 통계량들이 변화한다.
- 연 결 : 하나의 도형에 나타나 있는 점(하나 또는 몇 개의 점들의 집합)이 다른 도형에서 어떤 위치에 있는 가를 나타내주고 각 그라프에 연결된 자료에 라벨을 붙여준다.
- 선 택 : 관심있는 자료만을 선택하여 분석할 때 사용한다.

(f) 진단통계량막대

여기서는 3가지의 진단통계량이 막대로 나타난다. 첫 번째는 회귀계수 추정에서 로버스트의 척도를 나타내는 중앙값절대편차(MAD)이고 두 번째는 최소제곱 추정에 의한 잔차의 표준편차(SD), 세 번째는 LMS 회귀분석에서의 표준편차(σ_{LMS})의 추정값을 나타내는 MED이다. 진단통계량막대에서 가운데 점은 최초의 최소제곱법을 수행하였을 때 얻어지는 통계량을 나타내며, 왼쪽의 끝은 최초의 통계량을 5로 나눈 값이며, 오른쪽 끝은 최초 통계량의 5배이다.

(g) 회귀계수막대

각 설명변수에 해당하는 회귀계수를 제공한다. 이 막대의 중앙은 원시자료로 부터 최소제곱법을 적용하였을 때 구해지는 회귀계수의 추정값이며 막대의 양끝은 해당 회귀계수 추정량의 99% 신뢰구간의 하한과 상한이다. REDS에서는 회귀계수막대를 스크롤바형태로 회귀계수를 직접 변동시킬 수 있다.

2.2 주창의 메뉴

주창의 메뉴는 'File', 'Edit', '진단 도형', '보조 도형', 'Window'를 제공한다. 여기서 'File', 'Edit', 'Window' 등은 일반적인 Window에서 제공하는 기능과 같고, '진단 도형', '보조 도형' 등은 주창에서 제공하지 못하는 여러 가지 통계량을 보조창으로 제공한다.

(a) 진단 도형 메뉴

메뉴에서 [진단 도형]을 선택하면 현 모형에서의 잔차, 표준화된 잔차, Belsley 등이 제안한

표준화된 잔차, Cook 통계량 및 리버리지값, 잔차에 대한 상자도형 등으로 구성된 보조창이 나타난다. <그림 4>는 Hawkins 등(1984)의 자료를 분석하는 REDS에서 [진단 도형]을 선택한 그림이다.

(b) 보조 도형 메뉴

보조 도형은 <그림 5>에서와 같이 주창에서 [보조 도형] 메뉴를 선택하면 나타난다. 이 메뉴를 이용하므로서 회귀진단에 관한 여러 가지 통계량에 대한 정보를 얻을 수 있다. [보조 도형]에서 얻을 수 있는 정보는 다음과 같다.

- 현재 모형의 기초통계량
- 독립변수에 대한 잔차의 그라프 등 여러 가지 형태의 잔차도형
- 이상값을 판정할 수 있는 기각역을 제공하는 표준화된 잔차도형
- Belsley 등에 의한 이상값에 민감한 표준화된 잔차도형
- 잔차의 사후분포에 ±2SD 의 바차트를 나타낸 베이즈 잔차도형
- 적합된 추정값을 각 독립변수와 비교한 추정값 그라프
- Cook 통계량
- 각 독립변수에 대한 리버리지값
- 잔차 및 독립변수, 종속변수 등의 히스토그램
- 각 변수의 상자도형
- 산점도행렬

여기서 제공하는 잔차도형은 주창의 잔차도형에 비해 선택폭이 넓다. 즉, 주창에서 나타난 잔차도형의 경우 횡축에는 항상 관측값의 순서로 지정되어 있으나 이 경우 횡축에는 <그림 6>과 같이 사용자의 기호에 따라 적절한 통계량을 선택할 수 있다. <그림 6>의 경우 잔차도형의 횡축에 사용할 통계량으로 추정값을 선택하였으며 이 결과로 나타난 잔차도형이 <그림 7>과 같다.

3. 실제자료를 사용한 REDS의 실행 예

여기서는 Hawkins 등의 자료를 사용하여 REDS를 실행하는 과정을 살펴보기로 한다. 이 자료는 독립변수가 3개이며 관측값의 수는 총 75개(0-74번)로 0-9번의 자료는 가장된 이상값이고, 10-13번은 리버리지효과를 지니고 있으나 나머지 14-75번 자료와 적합이 잘되는 '좋은 리버리지점'들로 인위적으로 구성되어 있다. 이 자료를 REDS로 실행한 결과 주창이 <그림 8>로 가장효과와 리버리지효과에 의하여 이상값에 해당되는 0-9번의 자료는 이상값으로 판정되지 않고 '좋은 리버리지점'에 해당되는 10-13번의 자료가 이상값으로 판정되는 경향이 있는 것으로 나타났으며 이때의 MAD는 0.728이다. 메뉴바에서 [진단 도형]을 선택하면 <그림 4>에서 알 수 있듯이 10-13번의 자료가 잔차그라프에서 이상값으로 나타나고, 영향값을 나타내는 Cook통계량에서도 10-13번의 자료를 나타내고 있다. 또한 13번의 자료는 매우 강한 리버리지 효과를

나타내고 있으며 10-12번의 자료는 비교적 약한 리버리지 효과를 나타내고 있어 적절히 이상값을 탐색하지 못하는 것으로 나타났다.

<그림 8>의 잔차그라프에서 11번의 잔차를 +5.65 변동시키면, 이것이 잔차그라프 전체에 영향을 미쳐 <그림 9>의 새로운 진단그라프가 나타난다. <그림 9>를 참고하면 서로 가장된 효과를 지니고 있는 0-9번의 자료가 1개의 군집을 이루고 또한 리버리지효과를 지니고 있는 10-13번 자료가 다른 군집을 이루고 있는 것으로 나타났으며, 가장효과를 지닌 0-9번 자료를 이상값으로 판정할 수 있다. <그림 9>는 <그림 8>의 회귀계수막대에서 b_2 회귀계수를 0.131, b_3 회귀계수를 -0.003으로 직접 변동시켜서생성할 수도 있다.

<표 1>은 11번 잔차의 변동크기를 여러 가지 값으로 만들 때 MAD와 SD가 변화하는 과정으로, 11번 잔차를 양의 방향으로 5.65 이동시켰을 때 MAD가 0.728에서 0.554로 줄어들고 SD는 2.189에서 2.698로 늘어나는 것을 알 수 있다.

<표 1> Hawkins 자료에서 11번 잔차의 이동에 따른 SD, MAD

이동크기	MAD	SD	이동크기	MAD	SD
-7.00	1.267	3.050	2.84	0.641	2.352
-5.59	1.165	2.770	4.24	0.585	2.540
-4.19	1.045	2.531	5.65	0.554	2.698
-2.78	0.939	2.346	7.05	0.654	3.062
-1.38	0.911	2.228	8.46	0.698	3.375
0.00	0.728	2.189	9.87	0.773	3.709
1.43	0.659	2.232	11.27	1.042	4.061

참 고 문 헌

[1] 유종영, 김현철(1996). Hadi와 Simonoff의 다중이상점 식별방법의 개선과 여러 다중이상점 식별방법의 효율성 비교, 「한국통계학회 논문집」 제 3권 3호, 11-23.

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

[3] Becker, R. A., Cleveland, W. S., and Wilks, A. R.(1988), "Dynamic Graphics for Data Analysis (With Discussion)", *Statistical Science*, 2, 355-395.

[4] Cook, R. D. and Weisberg, S.(1994), An Introduction to Regression Graphics, JOHN WILEY & SONS.

[5] Easton, G. S.(1994),"A Simple Dynamic Graphical Diagnostics Method for Almost Any Model", *Journal of American Statistical Association,* 89, 201-207.

[6] Hawkins, D, M., Bradu, D., and Kass, G.V. (1984), "Location of several outliers in multiple regression data using elemental sets", *Technometrics,* 26, 197-208.

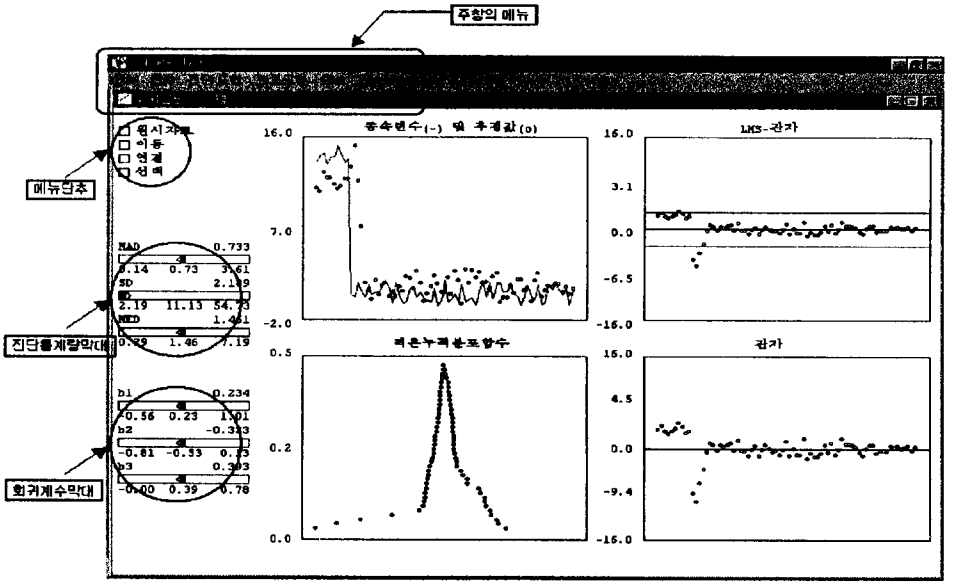
[7] Huh, Moon Yul(1995), "Exploring Multidimentional Data With the flipped empirical distribution function", *The Journal of Computational and Graphical Statistics,* Vol. 4, No. 4, 1-9.

[8] Tukey, J. (1973), *PRIM-9,* Produced by Stanford linear Accelerator Center, Stanford, Ca. Bin 88 Productions. (Film).

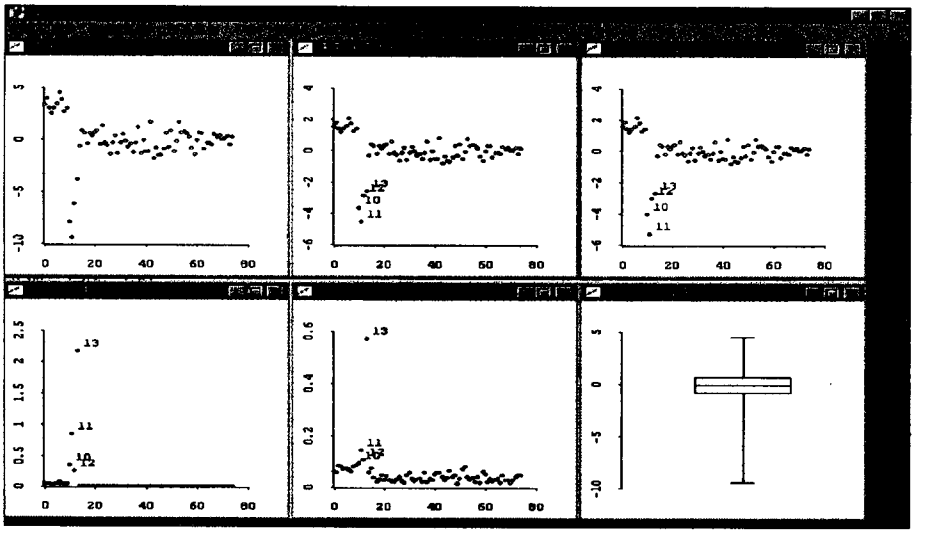
[9] Tierney, L.(1990), *LISP-STAT,* John Wiley & Sons

[10] Rousseeuw, P. J. and Leroy, A. M.(1987), *"Robust Regression and Outlier Detection",* JOHN WILEY & SONS.

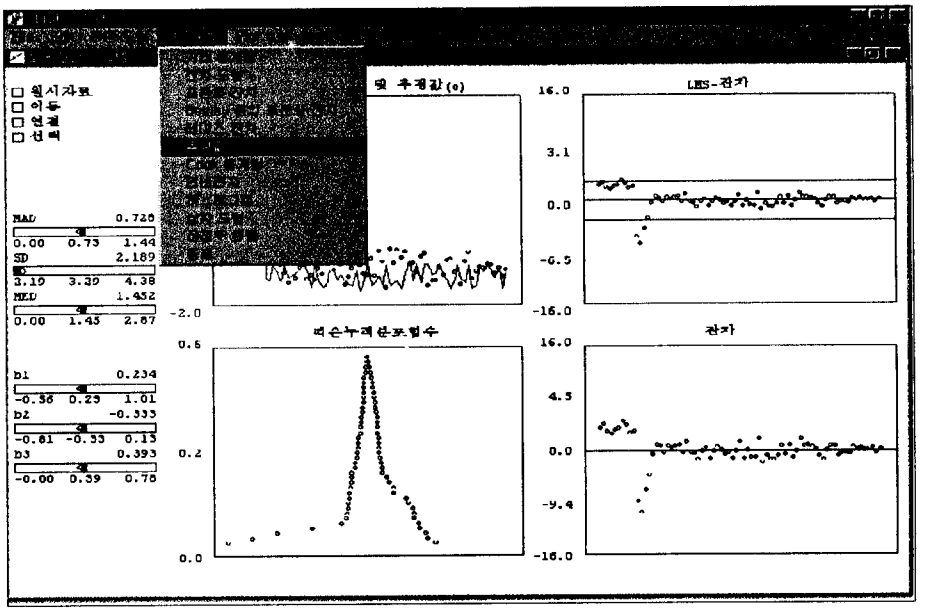
[11] Rousseeuw, P. J.(1984) "Least Median of Squares Regression" , *Journal of American Statistical Association,* 79, 871-880.



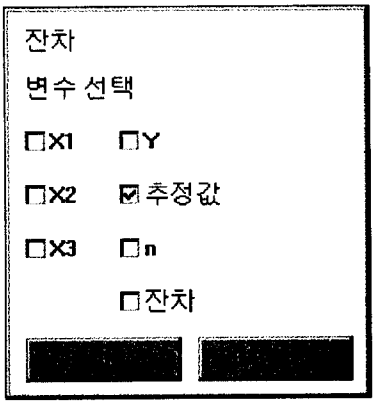
<그림 3> REDS 시스템의 구성



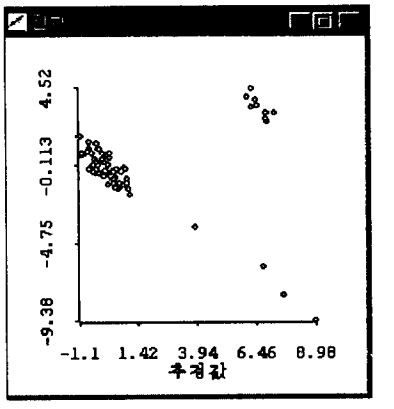
<그림 4> 메뉴바에서 [진단 도형]을 선택하였을 경우의 진단그라프



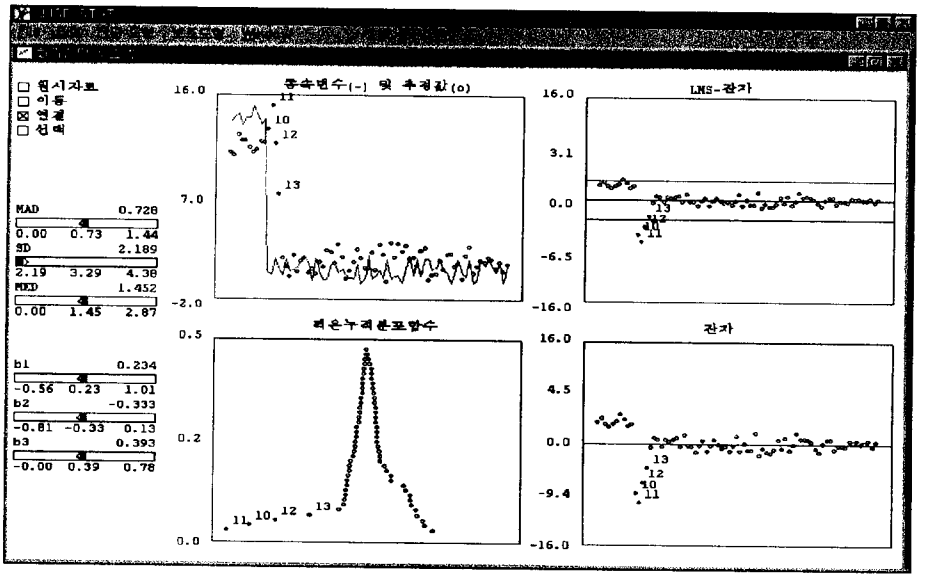
<그림 5> 메뉴바에서 [보조 도형]을 선택하였을 경우의 진단그라프



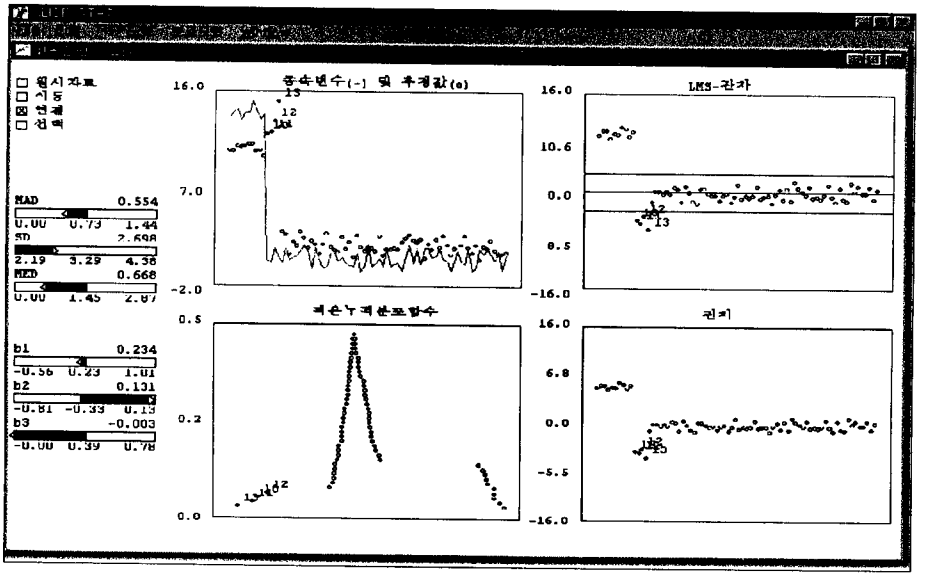
<그림 6> 변수선택상자



<그림 7> 추정값에 대한 잔차그라프



<그림 8> Hawkins 자료에서 최소제곱법에 의한 진단그라프



<그림 9> <그림 8>에서 11번 잔차를 +5.65 변동후의 진단그라프

Dynamic Graphics Approach for Regression Diagnostics System (REDS)

Jongyoung Yoo[4] · Kisoo Ahn[5] · Mun Yul Huh[6]

Abstract

Several studies have been down on the work of dynamic graphical methods for regression diagnostics. The main proposes of the methods were to investigate (1) the effects of change of data, or (2) the effects of change of regression coefficients on the regression models. But, by contrast, we can also investigate the effects of change of regression residuals on the regression model. This method can be used in fitting better a certain set of observations to a regression model than the other observations. Our researach team approaches regression diagnostics by using dynamic graphics(REDS), and we introduce REDS in this thesis.

4) Assistant Professor, Department of Computer Science and Statistics, Yongin University, Yongin, 449-714, Korea

5) Full-time Lecture, Department of Office Automation, Dongnam Health Junior College, Suwon, 440-714, Korea

6) Professor, Department of Statistics, Sung Kyun Kwan University, Seoul, 110-745, Korea