The Korean Journal of Applied Statistics (2014) DOI: http://dx.doi.org/10.5351/KJAS.2014.27.4.629
27(4), 629–643

Firework Plot as a Graphical Exploratory Data Analysis Tool to Evaluate the Impact of Outliers in a Mixture Experiment

Dae-Heung Jang[a] · SoJin Ahn[a] · Youngil Kim[b,1]

[a]Department of Statistics, Pukyong National University
[b]School of Business and Economics, ChungAng University

(Received June 3, 2014; Revised July 15, 2014; Accepted July 15, 2014)

Abstract

It is common to check the validity of an assumed model with the heavy use of diagnostics tools when conducting data analysis with regression techniques; however, outliers and influential data points often distort the regression output in undesired manner. Jang and Anderson-Cook (2013) proposed a graphical method called a firework plot for exploratory analysis that could visualize the trace of the impact of possible outlying and/or influential data points on individual regression coefficients and the overall residual sum of squares(SSE) measure. They developed 3-D plot as well as pair-wise plot for the appropriate measures of interest. In this paper, the approach was extended further to tell the strength of their approach; in addition, a more meaningful interpretation was possible by adding a measure not mentioned in their paper. This approach was applied to the mixture experiment because we felt that a detailed analysis of statistical measure sensitivity is required in a small experiment.

Keywords: Outliers, influential data point, mixture experiment, 3-D firework plot, pair-wise firework plot matrix.

1. 서론

혼합물 실험계획에서의 반응변수의 측정값은 혼합물에 쓰인 구성성분(component)들의 상대적인 비율에 의존한다고 가정한다. 따라서 혼합물 실험에서 구성성분들의 개수가 q일 때, 혼합물의 i번째 구성성분의 비율을 x_i라면

$$0 \le x_i \le 1, \quad i = 1, 2, \ldots, q \tag{1.1}$$

이다. 그리고

$$\sum_{i=1}^{q} x_i = 1 \tag{1.2}$$

This work was supported by a Research Grant of Pukyong National University(Year 2014).
[1]Corresponding author: School of Business and Economics, ChungAng University, 84 Heukseok-ro, Dongjak-gu, Seoul 156-756, Korea. E-mail: yik01@cau.ac.kr

이다. 통상 이차 혼합물(quadratic mixture) 실험의 정준형(canonical form)은 다음과 같다.

$$E(y) = \sum_{i=1}^{q} \beta_i x_i + \sum \sum_{i<j}^{q} \beta_{ij} x_i x_j. \tag{1.3}$$

특수 3차 혼합물(special cubic mixture) 실험의 정준형은

$$E(y) = \sum_{i=1}^{q} \beta_i x_i + \sum \sum_{i<j}^{q} \beta_{ij} x_i x_j + \sum \sum_{i<j<k}^{q} \sum \beta_{ijk} x_i x_j x_k \tag{1.4}$$

으로 표시된다.

그리고 물론 이와 같은 모형들은 다음과 같이 행렬로 표시된다.

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1.5}$$

여기서 $\mathbf{y}$는 $n \times 1$ 반응벡터이며, X는 $n \times p\,(p < n)$ 데이터 행렬이며 $\boldsymbol{\beta}$는 $p \times 1$ 모수벡터이며 $\boldsymbol{\epsilon}$은 $n \times 1$ 확률오차 벡터이다. 모수벡터 $\boldsymbol{\beta}$의 최소제곱추정량은 $\mathbf{b} = (X'X)^{-1}X'\mathbf{y}$이다. 여기서 p는 모형에 따라 그 크기를 달리한다. 예를 들어 $q = 2$인 경우에는 p의 크기는 4가 되며 $q = 3$인 경우는 p의 크기는 6이 된다. 이러한 일반적인 모형을 선호하는 이유는 혼합물실험계획의 분석도 일반선형회귀모형과 유사하게 진행되기 때문이다.

이상점(특이값, 특이점이라고도 함) 및 영향점(영향력이 큰 관측값이라고도 함)은 자료분석을 하는데 사용되는 계량적이고 기술적인 많은 측도들(measures)을 왜곡한다. 관련된 많은 참고문헌 중에서 Beckman과 Cook (1983)의 논문은 이상점에 관한 전반적인 정보를 담고 있다. 영향점과 관련해서는 Cook의 거리통계량 (Cook's distance measure; Cook, 1977, 1979)을 위시해서 Belsley 등 (1980)의 문헌들이 대표적으로 존재한다. 그리고 요약된 숫자 통계량 뿐 아니라 이를 활용한 다양한 도시적인 방법이 존재하는데 레버리지에 대한 인덱스 그림(index plot), Cook의 거리 통계량에 대한 인덱스 그림 뿐 아니라 Emerson과 Strenio (1983)의 산포-수준 그림(spread-level plot), 그리고 Fox (2008)의 회귀 영향그림(regression influence plot) 등이 문헌에 존재한다.

기존 문헌에 추가하여 Jang과 Anderson-Cook (2013)은 이상점들이 자료 분석에 있어 어떻게 기술적인 요약정보들에 영향을 주는지 불꽃그림(firework plot)이라는 그림을 이용하여 알아보았다. 그리고 회귀모형을 이용한 자료 분석인 경우에도 개개의 자료값이 회귀모형의 계수추정에 어떠한 영향을 주는지 보았다. 이러한 방법은 개개의 관측값이 존재 하는 경우와 그렇지 않은 경우를 비교하여 통계량을 구하기 보다는 개개의 관측값에 부여된 가중치를 연속적으로 변하게 하고 그 과정을 추적하여 봄으로써 기존의 방법과는 차별적인 방법을 제시하였다. 특히 그들의 불꽃 그림은 개별적으로 다루었던 이상점과 영향점 문제를 동시에 분석자가 파악할 수 있도록 3-D그림이나 짝진 그림행렬(pairwise plot matrix)등 도시적인 방법을 통해 일차원적인 분석의 범위를 벗어나고자 하였다. 본 연구에서는 2013년의 연구에서 언급되지 않은 Cook의 거리통계량을 추가하여 이상점들과 영향점 분석을 위한 기존 통계량과 이들 그림과의 관계를 명확히 하였다. 이를 위해 본 연구에서는 혼합물 실험계획 예제를 들어 실험계획에서의 세밀한 분석의 중요성을 언급하고자 한다.

Jang과 Anderson-Cook (2013)과 유사한 접근방법은 기존 연구 문헌에 없었던 것은 아니다. Cook과 Weisberg (1989)의 논문이나 Park 등 (1992)의 논문을 들 수 있는데, 특히 Cook과 Weisberg (1989)는 모형비교를 통해 관심있는 설명변수의 가중치의 변화가 잔차분석에 어떻게 영향을 주는지 동적인 그림의 형태로 소프트웨어를 구축한 바 있다. 불꽃그림은 연구에서는 이와 유사하나 가중치 변화에 따른 추적되는 선을 모든 점들에 대해 그림으로써 분석자로 하여금 종합적인 견해를 가지게 하였다.

Table 2.1. Augmented simplex-lattice design for the etch rate experiment

NO	x_1 (Acid A)	x_2 (Acid B)	x_3 (Acid C)	y
1	1	0	0	540
2	1	0	0	560
3	0	1	0	330
4	0	1	0	350
5	0	0	1	295
6	0	0	1	260
7	1/2	1/2	0	610
8	1/2	0	1/2	425
9	0	1/2	1/2	330
10	1/3	1/3	1/3	800
11	1/3	1/3	1/3	850
12	2/3	1/6	1/6	710
13	1/6	2/3	1/6	640
14	1/6	1/6	2/3	460

제 2절에서는 Jang과 Anderson-Cook (2013)이 제안한 불꽃그림에 대해 간단한 소개를 하고 혼합물실험의 구성성분에 대한 제약 조건이 없는 혼합물 실험계획의 예제와 제약조건이 존재 하는 두 예제를 통해 그 유용성을 알아보았다. 그리고 제 3절에서 결론을 내렸다.

2. 불꽃그림 소개 및 혼합물 실험계획 예제

통상적으로 최소제곱추정량인 경우 관측값에 부여되는 가중치를 $w_i = 1,\ i = 1, 2, \ldots, n$로 가정하나 Jang과 Anderson-Cook (2013)은 i번째 자료에 부여된 가중치의 값을 1에서 0으로 변화하게 하고, 나머지 자료의 경우 $w_j = 1 (j \neq i, j = 1, 2, \ldots, n)$로 고정한 다음 가중최소제곱추정량 $\hat{\boldsymbol{\beta}}_w$ 및 가중 잔차제곱합(weighted SSE) SSE_w에 영향이 있는지를 알기 위하여 연속적으로 그 변화를 추적하고 이를 모든 관측값에 적용, 상응하는 그림을 시도하였다.

개별 관측값에 대한 가중치가 다른 경우, 가중최소제곱추정량 $\hat{\boldsymbol{\beta}}_w$와 가중 잔차제곱합 SSE_w는 다음과 같이 표기한다.

$$\hat{\boldsymbol{\beta}}_w = \mathbf{b}_w = (X'WX)^{-1}X'W\mathbf{y}, \tag{2.1}$$
$$\mathrm{SSE}_w = \sum_i^n w_i(y_i - \hat{y}_i)^2,$$

여기서 W는 대각행렬로 대각원소는 가중치, $w_1, w_2, \ldots, w_n$이다. 또한 가중치 변화에 따른 Cook의 거리 통계량 D_w도 다음과 같이 쓸 수 있다.

$$D_w = \frac{(\mathbf{b} - \mathbf{b}_w)'X'X(\mathbf{b} - \mathbf{b}_w)}{p\mathrm{MSE}}, \tag{2.2}$$

여기서 MSE는 모든 관측값에 가중치가 1이 부여된 평균잔차제곱이다. Jang과 Anderson-Cook (2013)이 제안한 불꽃그림은 $\hat{\boldsymbol{\beta}}_w$ 및 SSE_w에 어떠한 변화가 생기는지 모든 과정을 추적하여 볼 수 있는 그림이다. $\hat{\boldsymbol{\beta}}_w$에 대한 변화를 보는 것은 Belsley 등 (1980)의 DFBETAS 통계량의 연장선에 있는 개념이다. 추가로 Cook의 거리 통계량을 엮어서 각 관측값의 가중치가 1에서 0으로 변하는 경우도 필

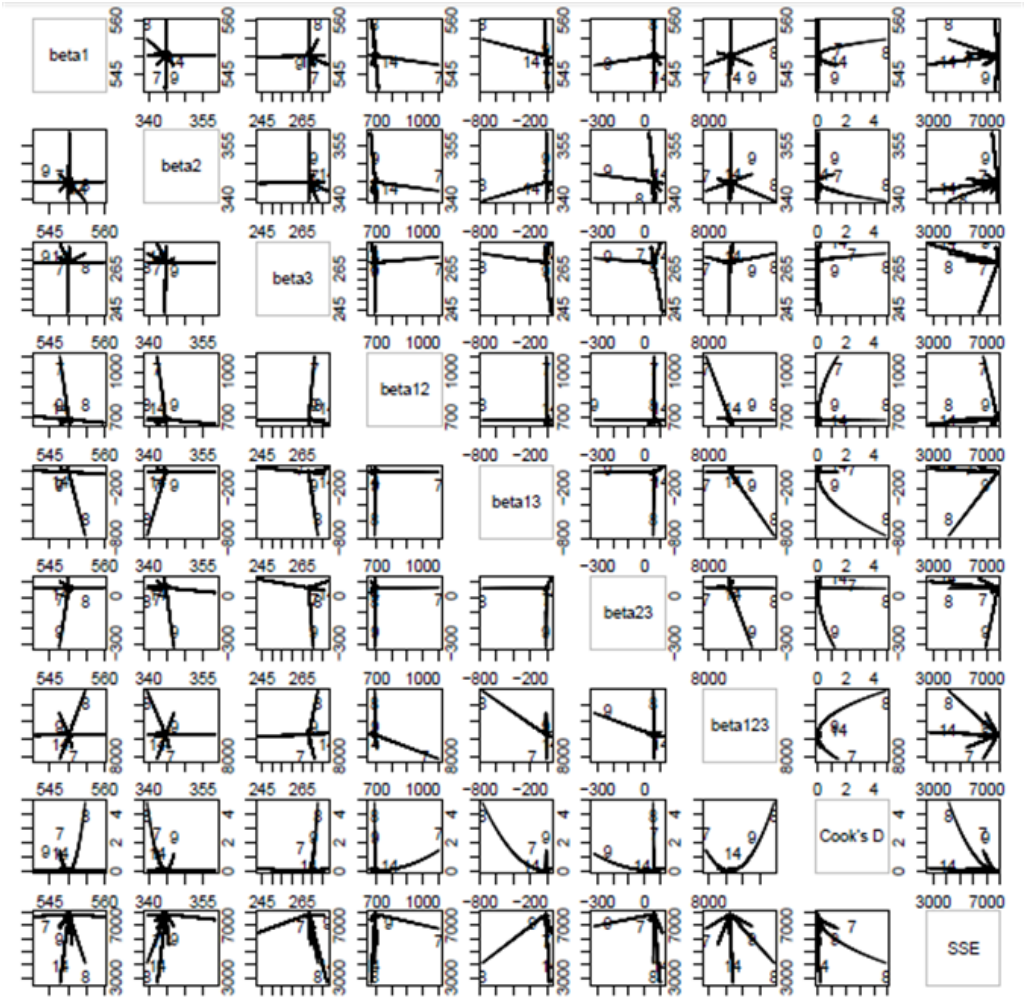


Figure 2.1. Pairwise firework plot matrix for the etch rate experiment dataset: All regression coefficients, Cook's distance and SSE

요하다고 본다. 왜냐하면 개별적인 회귀계수의 변화 뿐 만 아니라 종합적인 적합도의 변화도 필요하기 때문이다.

기존연구에서는 i번째 관측값이 제거되는 경우와 그렇지 않은 경우를 비교하고 그 차이에 대한 크기를 통계량으로 만들어 졌다면 본 연구에서 제시한 방법은 극히 도시적이라 할 수 있다. 이러한 불꽃그림은 탐색적인 연구의 목적에 따라 3차원 적으로도 만들어 질 수 있어, 보다 시각적인 효과를 뚜렷이 할 수 있다. 따라서 본 연구의 목적은 사용자로 하여금 전반적인 자료의 회귀분석계수 및 가중 잔차제곱합에 대한 영향력을 점검하는 탐색적인 목적이 있다. 2.1절 및 2.2절에서는 이러한 아이디어를 혼합물 실험계획의 예제들을 통해 구현해 보았다. 이러한 응용예제는 Jang과 Anderson-Cook (2013)의 논문의 연장선상에 있으나 본 연구에서는 Cook의 거리 통계량을 추가하고 DFBETAS의 개념을 연계하여 다양한 분석을 시도하였다.

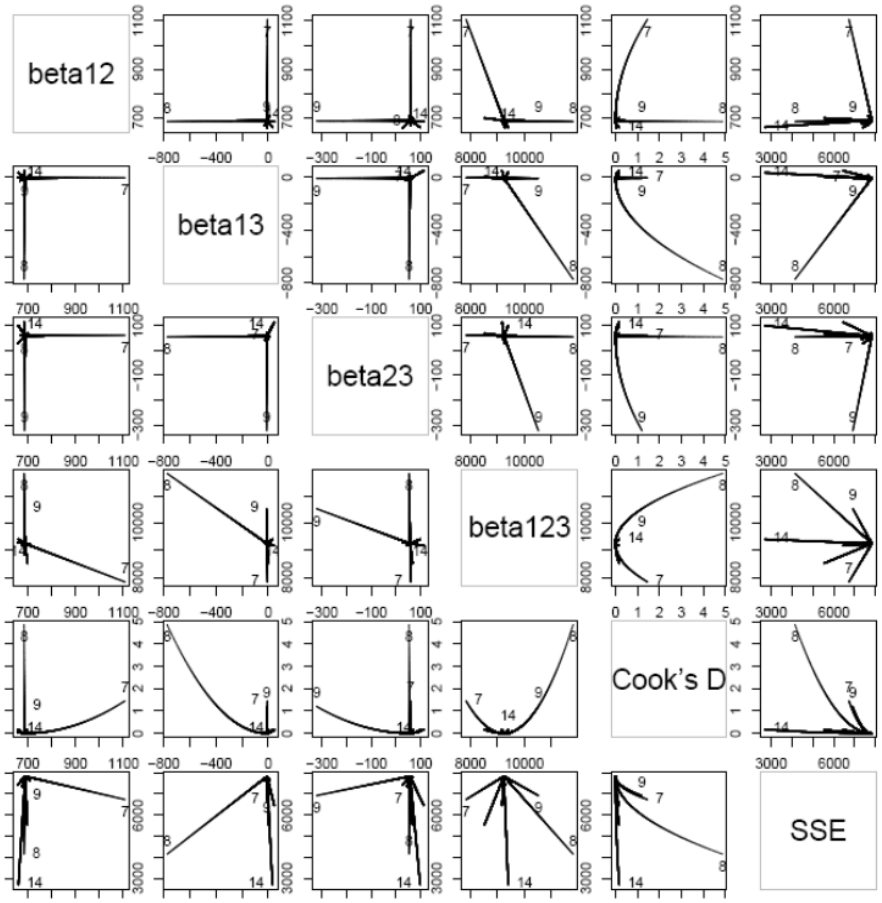
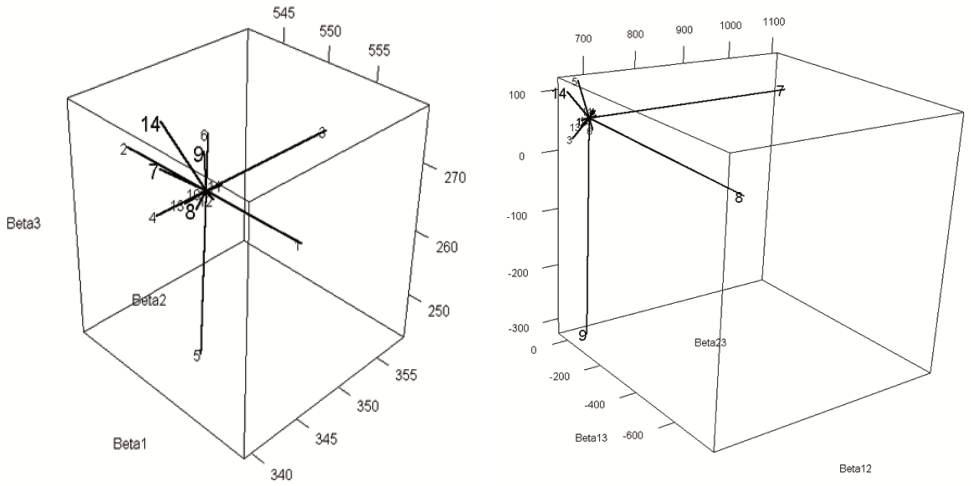


Figure 2.2. Pairwise firework plot matrix for the etch rate experiment dataset: Mixed quadratic/cubic regression coefficients, Cook's distance and SSE

대용량의 자료인 경우 개별적인 이상점이나 영향점들의 분석보다는 군집으로 존재할 수 있는 이상점이나 영향점들의 존재 여부를 파악하는 것이 중요하다. 그러나 여전히 실험계획에서 나오는 소규모의 자료인 경우 고전적인 이상점이나 영향점 분석은 유효하다고 보여 지며 특히 실험계획인 경우 특히 더 세밀한 분석을 요하는 바 예제를 혼합물 실험예제에서 취하여 보았다. 혼합물 실험인 경우 구성성분의 비율(proportion)이 x_i가 되므로 만약 이상점이나 영향점으로 판단이 되는 경우 그렇게 배합된 구성성분들의 비율로 만들어지는 제품들의 특성에 대한 추가적인 연구가 필요하기 때문에 혼합물실험계획 예제를 들어 설명하였다. 예제는 구성성분들에 대한 추가적인 제약조건이 없는 경우와 있는 경우 두 경우로 구분하였다.

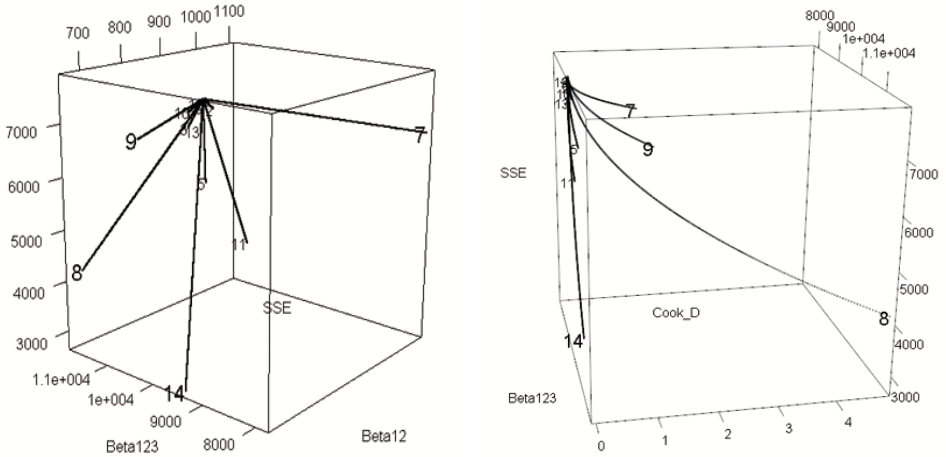
2.1. 추가적인 제약조건이 없는 예제

Table 2.1의 자료는 Myers 등 (2009)의 자료이다. 세미콘덕터(semiconductor) 산업에서는 금속화공정 전에 실리콘 웨이퍼 뒷면에 젖은 화학 에칭(chemical etching)작업을 수행한다. 에칭에 쓰이는 용



(a) Linear regression coefficients (b) Mixed quadratic regression coefficients

Figure 2.3. 3-D firework plot for the etch rate experiment dataset



(a) β_{12}, β_{123} and SSE (b) β_{123}, Cook's distance and SSE

Figure 2.4. 3-D firework plot for the etch rate experiment dataset

해물은 세 가지 물질(A, B, C) 구성성분들로 구성되어 있으며 Table 2.1은 소위 확장 심플렉스격자실험(augmented simplex lattice design)이라는 실험설계이며 반응변수는 분 시간단위당 에칭비율이다.

이 자료에 특수 3차 회귀모형(special cubic regression model)을 적합하면 다음과 같은 회귀식이 나온다.

$$\hat{y} = 550.200x_1 + 344.723x_2 + 268.295x_3 + 689.537x_1x_2 - 9.035x_1x_3 + 58.108x_2x_3 + 9242.336x_1x_2x_3.$$

그리고 각각의 추정된 회귀계수들에 대응되는 표준오차들은 순서대로 23.22, 23.22, 23.22, 146.52, 146.52, 146.52, 940.86이 된다. $\alpha = 0.05$에서는 x_1x_3와 x_2x_3의 계수 값은 유의하지 않다. 참고로 혼합

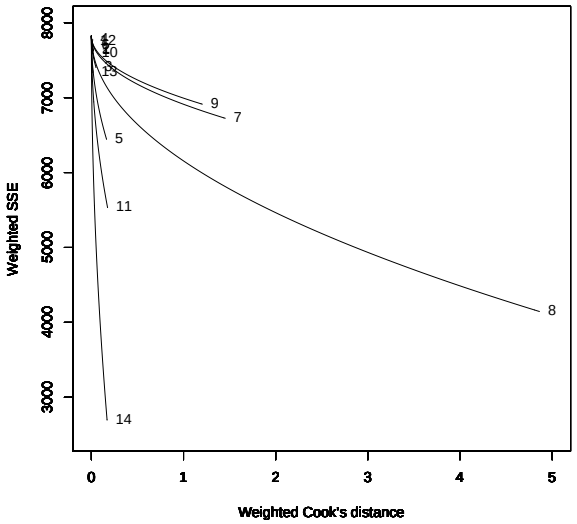


Figure 2.5. Firework plot corresponding to Cook's distance and SSE for the etch rate experiment dataset

물 실험계획의 특성상 설령 유의하게 나오지 않은 회귀계수라 하더라도 해석의 편의성을 위하여 모델 안에 포함시키는 경우가 많다. Figure 2.1은 개별회귀계수 전체와 Cook의 거리 통계량 그리고 SSE에 대하여 짝진 불꽃그림 행렬(pairwise firework plot matrix)을 보여 주고 있으며 Figure 2.2는 이들 회귀계수 중에서도 이차항과 삼차항에 대한 회귀계수에 국한하여 짝진 불꽃그림 행렬을 보여 주고 있다. 또한 Figure 2.3(a)는 일차항 회귀계수에 대한 3차원 불꽃그림, 그리고 Figure 2.3(b)는 이차항 회귀계수에 대한 3차원 불꽃그림이다. Figure 2.4(a)는 β_{12}, β_{123} 및 SSE에 대한 불꽃그림, 마지막으로 Figure 2.4(b)는 삼차항 회귀계수 및 Cook의 거리 통계량 및 SSE에 대한 3차원 불꽃그림이다. 이들 그림들로 판단하건데 7, 8, 9번째 자료값들(변점)이 회귀계수에 영향을 많이 미치는 점들로 나타난다. 특히 이러한 현상은 일차항 회귀계수 보다는 이차항과 삼차항의 회귀계수에 대하여 더 많은 영향력을 행사하는 것으로 나타난다. 14번째 자료값(내부점)은 SSE값의 변화에 많은 영향력을 행사하는 것으로 보아 이상점으로 보여 진다.

종합적으로 이상점과 영향점을 볼 수 있는 Figure 2.5는 Cook의 거리통계량과 SSE에 대한 불꽃그림이다. Fox가 제안한 회귀영향 그림(regression influence plot)과 같은 단순 산점도 형식이 아닌 가중치 변화에 따른 추적선을 이차원적으로 보여주는 점에서 참고 할 만하다. 역시 8번째는 제일 큰 영향점이고 14번째 자료값은 이상점으로 밝혀진다. 참고로 Table 2.2는 통상적인 회귀진단 통계량들을 보여 주는데 물론 본 연구가 제안한 불꽃그림이라는 도시적인 방법이 Table 2.2에서 제시된 기존의 요약된 통계량보다는 직감적인 정보를 더 많이 보여 준다.

2.2. 추가적인 제약조건이 있는 예제

혼합물 실험계획은 구성성분의 특성상 각 구성성분이 가질 수 있는 범위를 지정하는 경우가 많다. 즉, 구성성분의 상한과 하한에 대한 값을 설정하는 경우가 있다. 이 경우는 제한영역을 다음과 같이 설정하여야 한다.

$$L_i \le x_i \le U_i, \quad i = 1, 2, \ldots, q. \tag{2.3}$$

Table 2.2. $y, \hat{y}$, studentized residual, leverage(h_{ii}), Cook's distance for the etch rate experiment

NO	Actual Value, y	Predicted Value, $\hat{y}$	R-Student	leverage(h_{ii})	Cook's distance
1	540.0	550.2	−0.40	0.483	0.024
2	560.0	550.2	0.38	0.483	0.022
3	330.0	344.7	−0.58	0.483	0.050
4	350.0	344.7	0.20	0.483	0.006
5	295.0	268.3	1.13	0.483	0.164
6	260.0	268.3	−0.32	0.483	0.016
7	610.0	619.8	−0.99	0.912	1.453
8	425.0	407.0	2.31	0.912	4.862
9	330.0	321.0	0.89	0.912	1.204
10	800.0	812.2	−0.43	0.375	0.018
11	850.0	812.2	1.58	0.375	0.176
12	710.0	717.4	−0.23	0.206	0.002
13	640.0	620.2	0.64	0.206	0.016
14	460.0	523.8	−3.38	0.206	0.170

Table 2.3. Extreme vertices design for the railroad flare experiment

NO	x_1	x_2	x_3	x_4	y
1	0.4	0.1000	0.4700	0.030	75
2	0.4	0.1000	0.4200	0.080	180
3	0.6	0.1000	0.2700	0.030	195
4	0.6	0.1000	0.2200	0.080	300
5	0.4	0.4700	0.1000	0.030	145
6	0.4	0.4200	0.1000	0.080	230
7	0.6	0.2700	0.1000	0.030	220
8	0.6	0.2200	0.1000	0.080	350
9	0.5	0.1000	0.3450	0.055	220
10	0.5	0.3450	0.1000	0.055	260
11	0.4	0.2725	0.2725	0.055	190
12	0.6	0.1725	0.1725	0.055	310
13	0.5	0.2350	0.2350	0.030	260
14	0.5	0.2100	0.2100	0.080	410
15	0.5	0.2225	0.2225	0.055	425

이를 위하여 McLean과 Anderson(1966) (Myers 등 (2009)의 책 604 페이지 참조.)에 있는 예제를 들 수 있다. x_1은 마그네슘(magnesium)이고 x_2는 소디엄 니트레이트(sodium nitrate), x_3는 스트론티움(strontium)이고 x_4는 블라인더(blinder)이다. 그리고 반응변수는 1000 촛불의 크기로 정한 조도(illumination)의 양이 된다. 그리고 다음은 각 구성성분에 대한 제약조건식이다.

$$0.40 \le x_1 \le 0.60, \quad 0.10 \le x_2 \le 0.50, \quad 0.10 \le x_3 \le 0.50, \quad 0.03 \le x_4 \le 0.08.$$

이러한 제약조건식으로 인하여 제약조건 하의 실험영역은 8개의 꼭지점, 12개의 변과 6개의 면을 갖는 불규칙 다면체를 형성한다. Table 2.3은 해당하는 데이터인데 이를 혼합물 실험계획에서는 극단꼭지점 실험(extreme vertices design)이라 부른다.

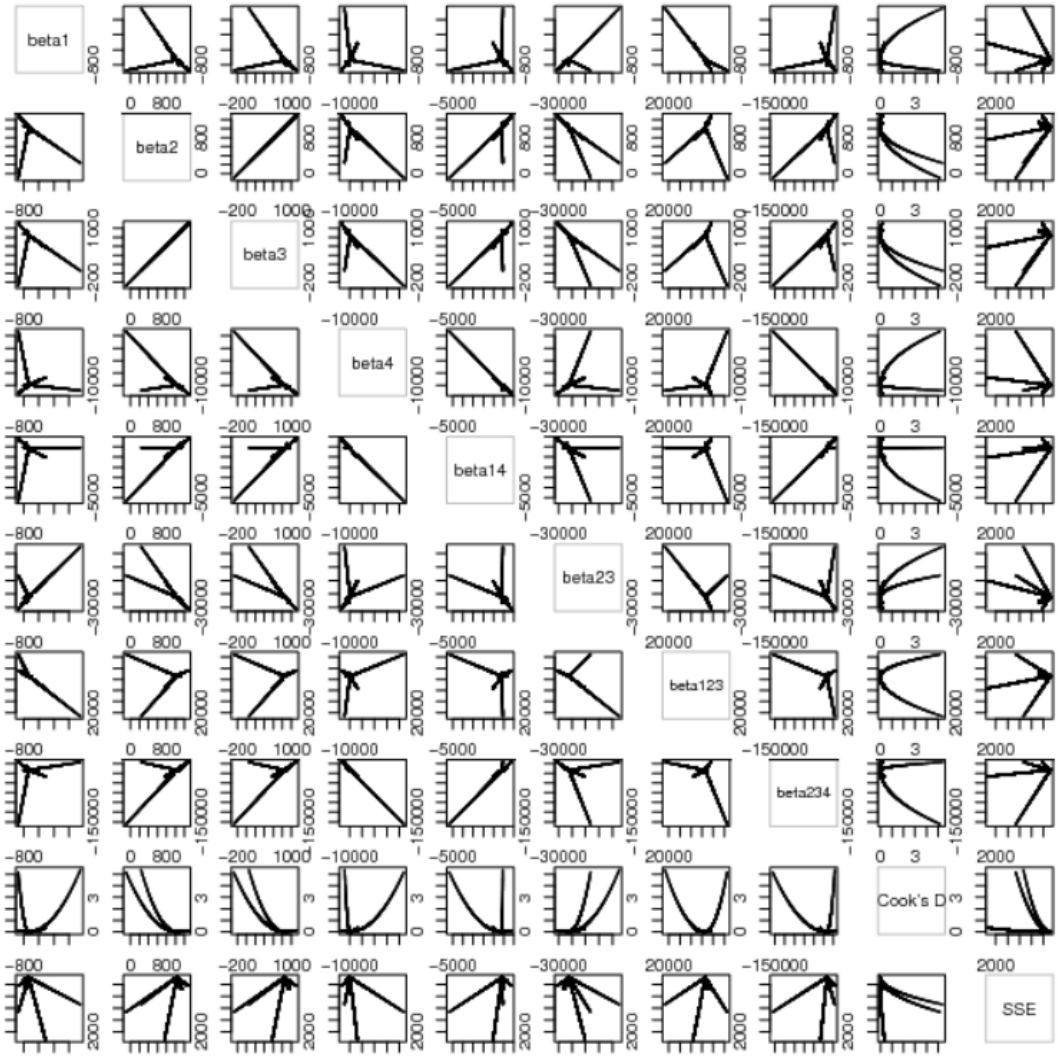


Figure 2.6. Pairwise firework plot matrix for the railroad flare experiment dataset: All regression coefficients, Cook's distance and SSE

이 자료에 특수 3차 회귀모형을 적합하면 다음과 같은 회귀식이 나온다.

$$\hat{y} = -734.307x_1 + 1029.821x_2 + 849.011x_3 - 9874.473x_4 + 19162.127x_1x_4 - 26120.150x_2x_3 + 52817.778x_1x_2x_3 + 116303.493x_2x_3x_4$$

그리고 각각의 추정된 회귀계수들에 대응되는 표준오차는 순서대로 240.3, 391.8, 391.8, 5504.0, 7661.1, 5861.1, 9497.3, 61157.9이며 $\alpha = 0.05$에서는 x_4, x_2, x_3, x_4에 해당하는 계수는 유의하지 않다.

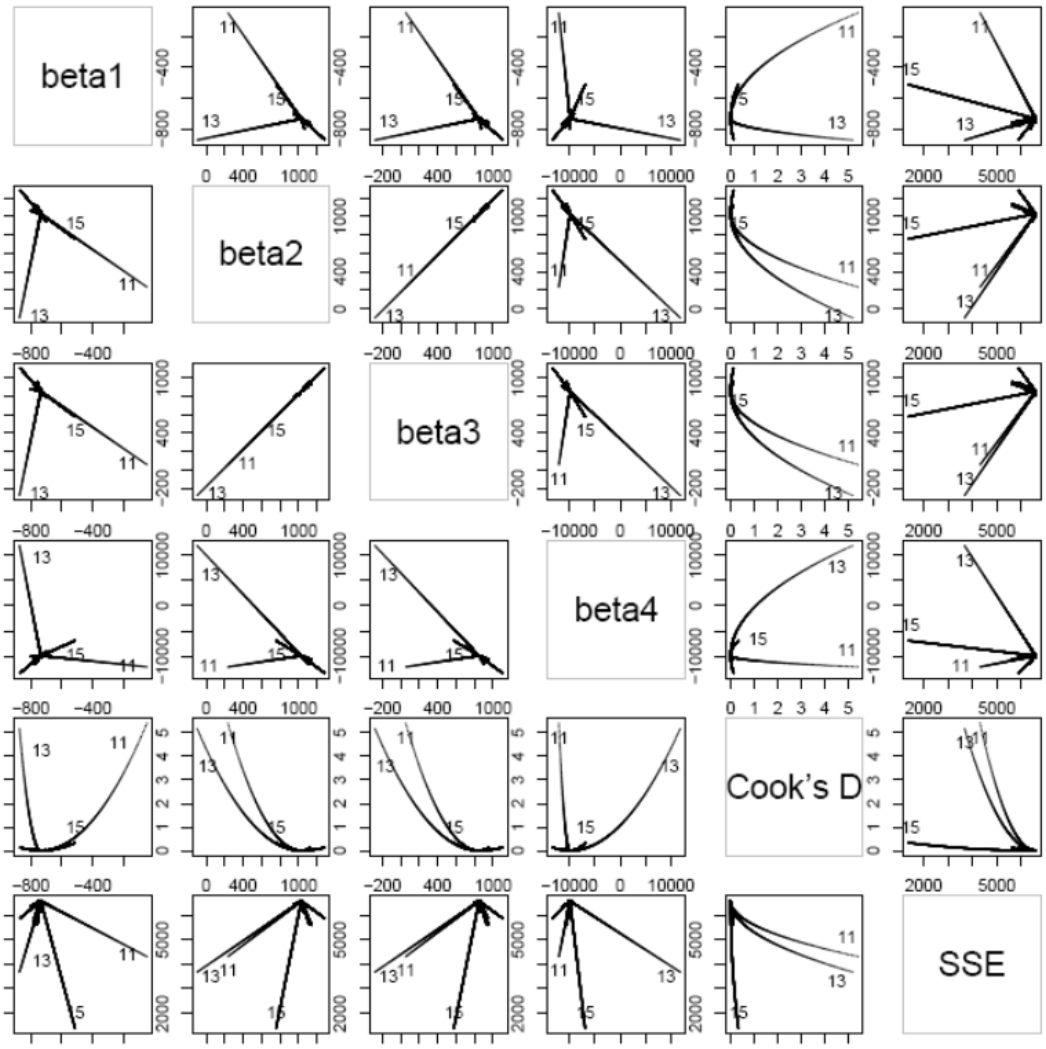


Figure 2.7. Pairwise firework plot matrix for the railroad flare experiment dataset: Linear regression coefficients, Cook's distance and SSE

Figure 2.6–2.12는 다양하게 그려본 불꽃그림들인데 2.1절 예제에서 제시된 그림들과 유사하게 배치 되어 있으므로 중복을 피하기 위하여 자세한 설명은 생략한다. 결론적으로 11, 13번째 관측값(불규칙 다면체의 면중심점)은 회귀계수에 많은 영향력을 행사하는 영향점으로 판단되고 반면 15번째 관측값(불규칙 다면체의 전체중심점)은 가중치가 1에서 0으로 변화하는 과정에서 SSE의 크기가 급속히 줄고 있어 이상점에 해당된다고 볼 수 있다.

참고로 Table 2.4는 통상적인 회귀진단 통계량들을 보여 주는데 물론 본 연구가 제안한 불꽃그림이라는 도시적인 방법이 Table 2.4에서 제시된 기존의 요약된 통계량보다는 직감적인 정보를 더 많이 보여 준다. Table 2.4의 결과가 본 연구에서 수행된 분석과 대동소이하다고 하여 종합적이고 시각적인 분석이

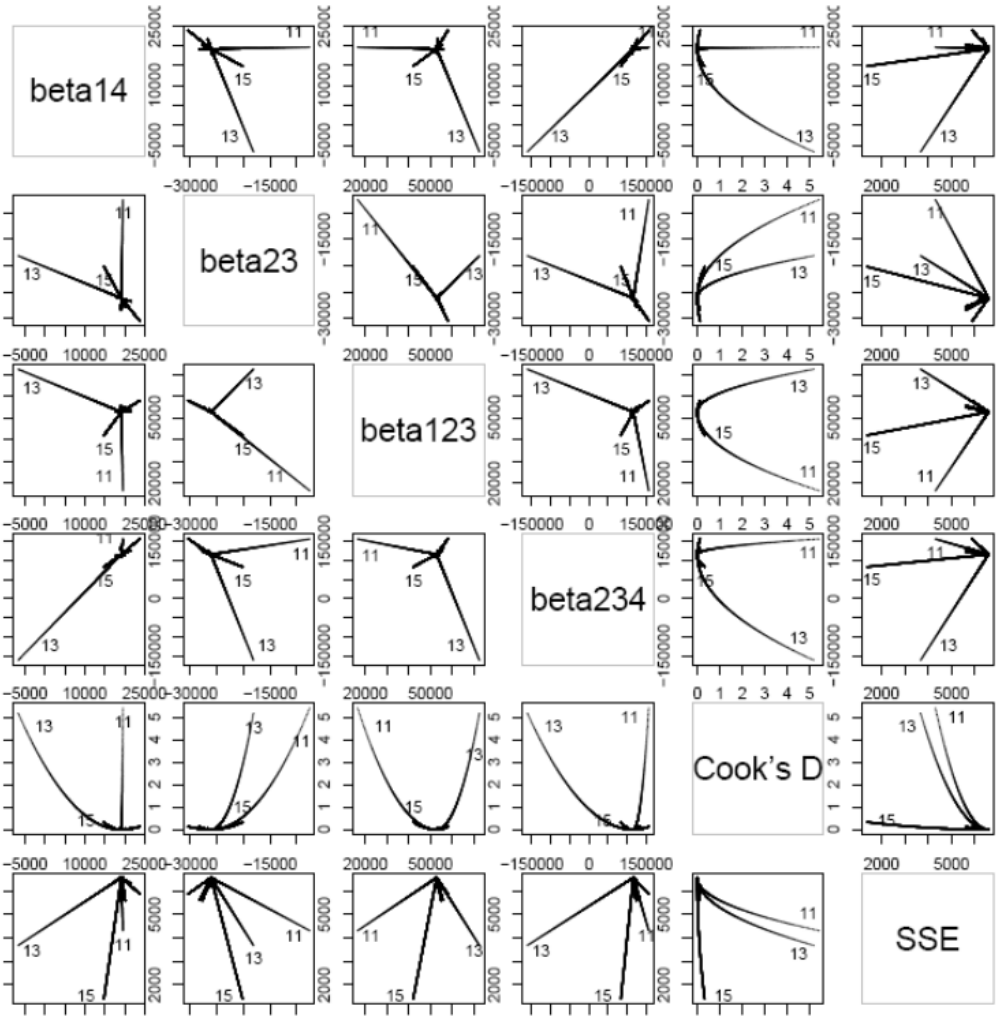
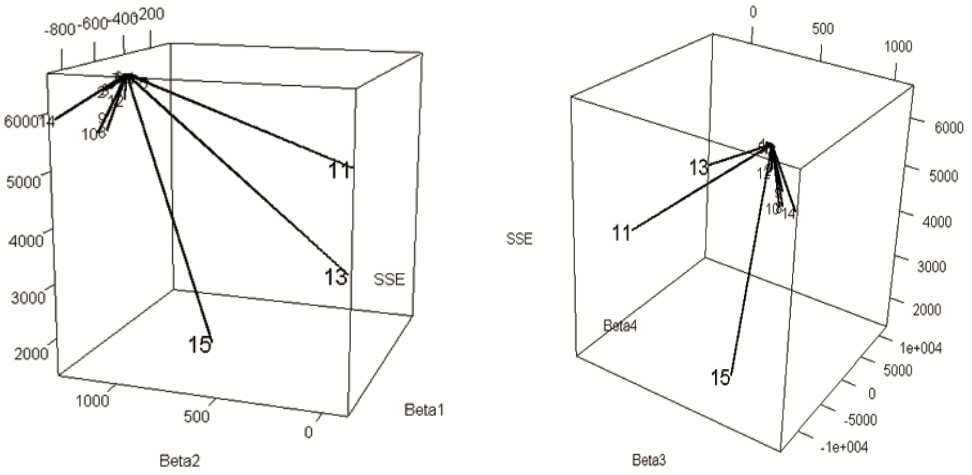


Figure 2.8. Pairwise firework plot matrix for the railroad flare experiment dataset: Mixed quadratic/cubic regression coefficients, Cook's distance and SSE

필요하지 않다는 이야기는 아니다. 어느 구성성분들의 비율이 이상점과 영향점에 해당이 되는지 종합적인 분석이 필요하다. 특히 실험에 사용되는 구성성분에 대한 제약조건이 있는 경우는 없는 경우에 비하여 극점에서 비정상적인 반응이 나올 가능성이 상대적으로 많음으로 이에 대한 점검이 세밀하게 필요하다.

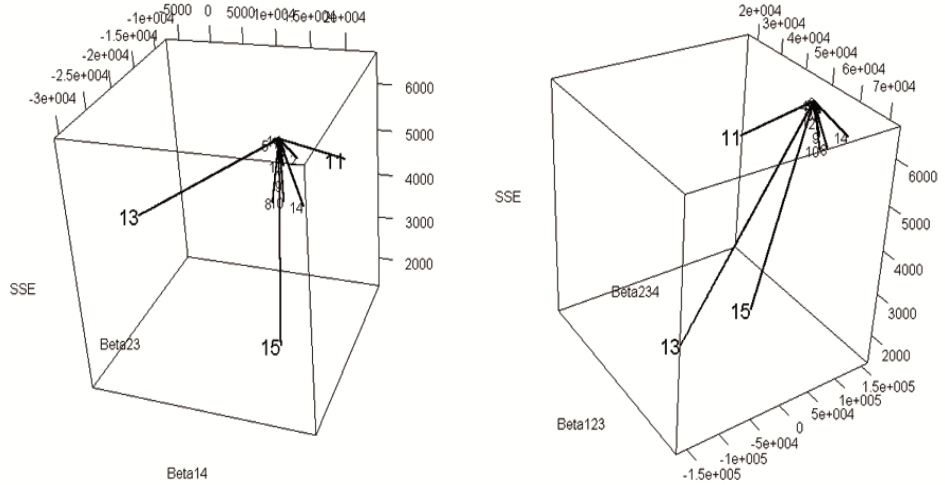
3. 결론

본 연구에서는 회귀분석에서 이상점이나 영향점을 진단하기 위한 도구로서 도시적이고 종합적인 분석이 가능한 그림 몇 가지를 제안하였다. 불꽃그림을 통하여 자칫 개별적인 결과에만 치우칠 수 있는 통



(a) Linear regression coefficients β_1, β_2 and SSE (b) Linear regression coefficients β_3, β_4 and SSE

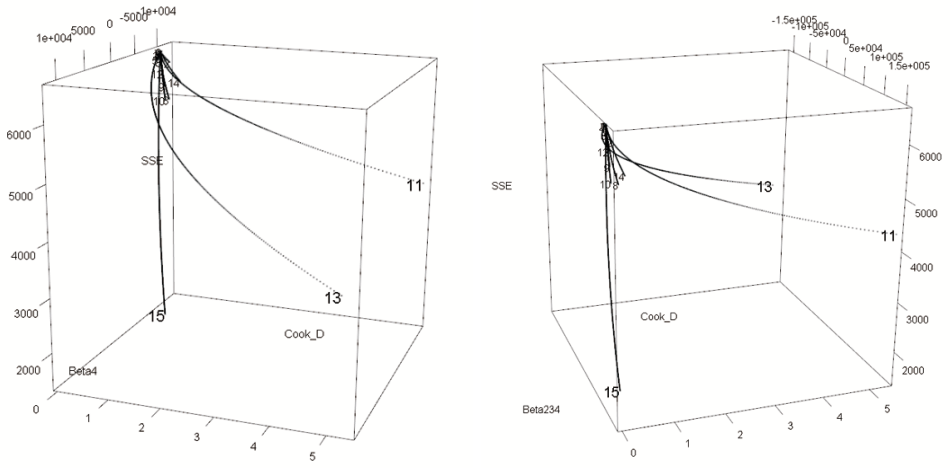
Figure 2.9. 3-D firework plot for the railroad flare experiment dataset



(a) β_{14}, β_{23} and SSE (b) β_{123}, β_{234} and SSE

Figure 2.10. 3-D firework plot for the railroad flare experiment dataset

계량들을 서로 연계하여 볼 수 있고 또한 가중치가 1에서 0으로 관측값이 완전제거 될 때까지 그 추적 동선을 판단할 수 있는 종합적인 정보를 제공한다. 분석에서 기존의 통계량의 결과와 비슷한 결론을 유도하였다 하더라도 도시적인 방법의 중요성은 간과 할 수 없다. 본 연구에서 제안한 불꽃그림은 기존의 단순 통계량을 단순 도시화하는 도구라기보다 종합적인 도구로 간주가 되어야 한다. Jang과 Cook-Anderson (2013)에서 빠진 부분 즉 Cook의 거리 통계량을 추가함으로써 더욱 영향점에 대한 분석이 용이하게 되었다. 즉 짝진 그림행렬이나 3차원 그림을 통하여 개별 회귀계수에 대한 변화량 곡선 뿐 아니라 종합적인 추적곡선을 동시에 보여 줌으로서 분석자로 하여금 더욱 더 자료에 대한 이해를 할 수 있



(a) β_4, Cook's distance and SSE (b) β_{234}, Cook's distance and SSE

Figure 2.11. 3-D firework plot for the railroad flare experiment dataset

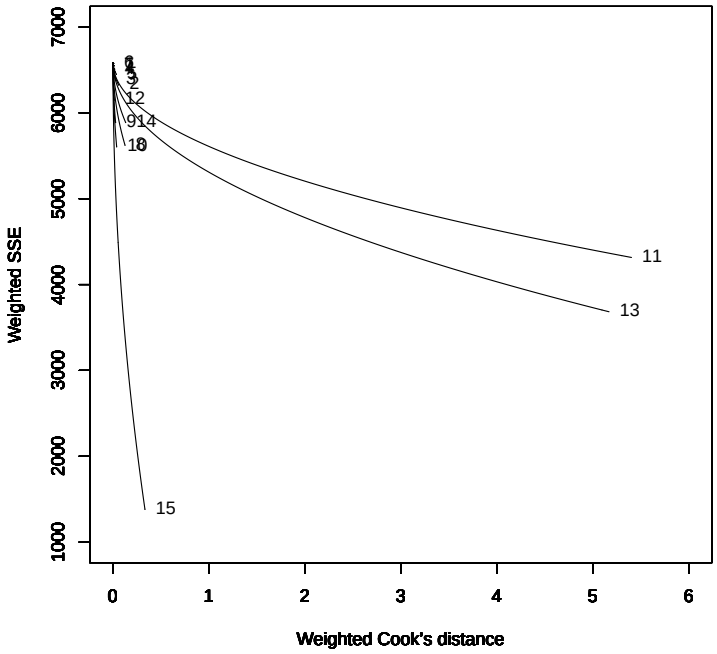


Figure 2.12. Firework plot corresponding to Cook's distance and SSE for the railroad flare experiment dataset

게 한다. 기존의 도시적인 방법에 비하여 본 연구에서 제안한 방법은 보다 종합적인 도시적 방법이라 보여지며, 기존의 방법과 보완되어 사용되면 좋을 것이다. 본 연구에서 예제로 든 혼합물 실험계획은 특히 이러한 문제가 중요시되어 연구의 예제로 사용하였다. 구성성분들에 제약조건이 있는 경우는 더 영향점

Table 2.4. $y, \hat{y}$, studentized residual, leverage(h_{ii}), Cook's distance for the railroad flare experiment

NO	Actual Value, y	Predicted Value, $\hat{y}$	R-Student	leverage(h_{ii})	Cook's distance
1	75	71.321	0.19275	0.666	0.01074
2	180	170.146	0.49825	0.629	0.05892
3	195	184.925	0.43434	0.494	0.02608
4	300	306.250	−0.26707	0.495	0.01008
5	145	138.221	0.35783	0.666	0.03646
6	230	228.005	0.09891	0.629	0.00241
7	220	215.663	0.18464	0.494	0.00483
8	350	327.947	1.01383	0.495	0.12549
9	220	243.247	−0.84445	0.227	0.02737
10	260	287.546	−1.02551	0.227	0.03842
11	190	200.960	−1.77669	0.947	5.39995
12	310	328.868	−0.65325	0.186	0.01327
13	260	274.201	−2.17475	0.931	5.16620
14	410	426.981	−0.84426	0.588	0.13242
15	425	365.720	4.75960	0.325	0.33319

이나 이상점에 주의를 요하는 상황이기 때문이다.

References

Beckman, R. J. and Cook, R. D. (1983). Outlier ... s, *Technometrics*, **25**, 119–147.

Belsley, D. A., Kuh, E. and Welch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Source of Collinearity*, Wiley, New York.

Cook, R. D. (1977). Detection of influential observation in linear regression, *Technometrics*, **19**, 15–18.

Cook, R. D. (1979). Influential observation in linear regression, *Journal of American Statistical Association*, **74**, 169–174.

Cook, R. D. and Weisberg, S. (1989). Regression diagnostics with dynamic graphics, *Technometrics*, **31**, 277–291.

Emerson, J. D. and Strenio, J. (1983). The spread-versus-level plot in Hoaglin, D. C., Mosteller, F. and Tukey, J. W.(Eds.) (1983). *Understanding Robust and Exploratory Data Analysis*, Wiley, New York.

Fox, J. (2008). *Applied Regression Analysis and Generalized Linear Models*, 2nd ed., Sage, New York.

Jang, D. H. and Anderson-Cook, C. M. (2013). Firework plot as a graphical exploratory data analysis tool for evaluating the impact of outliers in data exploration and regression, *Quality and Reliability Engineering International*, in press.

McLean, R. A. and Anderson, V. L. (1966). Extreme vertices design of mixture experiments, *Technometrics*, **8**, 447–454.

Myers, R. H., Montgomery, D. C. and Anderson-Cook, C. M. (2009). *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, 3rd ed., Wiley, New York.

Park, S. H., Kim, Y. H. and Toutenberg, H. (1992). Regression diagnostics for removing an observation with animating graphics, *Statistical Papers*, **33**, 227–240.

혼합물 실험에서 특이값의 영향을 평가하기 위한 그래픽 탐색적 자료분석 도구로서의 불꽃그림

장대흥[a] · 안소진[a] · 김영일[b,1]

[a]부경대학교 통계학과, [b]중앙대학교 경영학부

(2014년 6월 3일 접수, 2014년 7월 15일 수정, 2014년 7월 15일 채택)

요 약

회귀모형을 이용하여 자료를 분석하는 경우 이상점이나 영향점과 같은 특이값들의 유무를 검정하는 회귀진단기법은 모형의 적합성을 체크하기 위한 필수적인 도구로 잡은 지 오래이다. 이러한 점들이 존재 하는 경우 회귀분석의 결과가 왜곡되어 해석이 된다. Jang과 Anderson-Cook (2013)은 불꽃그림이란 이름을 붙인 그림도구를 발표하였는데 관측값에 부여된 가중치를 1에서 0으로 변화함에 따라 이상점이나 영향점이 회귀계수 및 잔차제곱합(SSE)에 어떠한 영향을 미치는지 3차원 그림에 추적곡선을 그려 보았을 뿐 아니라 쌍으로 대비시켜 봄으로써 분석의 시각적인 효과를 증대시켰다. 본 연구에서는 더 나아가 이러한 시도가 기존 방법과 어떤 차이점이 있는지 2013년에는 반영치 않은 통계량을 포함해서 더 많은 해석이 가능한지 혼합물 실험 계획을 통해 다양한 통계량의 민감도 분석을 실행하였다. 왜냐하면 작은 혼합물실험인 자료인 경우 더욱 세밀한 통계량에 대한 민감도 분석이 필요하기 때문이다.

주요용0어: 이상점, 영향점, 혼합물실험계획, 3-D 불꽃그림, 짝진 불꽃그림 행렬.

이 논문은 부경대학교 자율창의학술연구비(2014년)에 의하여 연구되었음.

[1]교신저자: (156-756) 서울특별시 동작구 흑석동 84, 중앙대학교 경영경제대학 경영학부.
E-mail: yik01@cau.ac.kr