⊠ 응용논문

회귀 분석에서 이상치가 미치는 영향

김광수 · 배영주
충주산업대학교 산업공학과
이진규
동국대학교 산업공학과

The Effect of Outliers in Regression Analysis

Kwang-Soo Kim, Young-Ju Bae
Dept. of Industrial Engineering, Chung Ju National University
Jin-Gue Lee
Dept. of Industrial Engineering, Dongguk University

Abstract

Outlier is one that appears to deviate extremely from other data in collected data. Thus treatment of outlier is very important work, because it is to distort the meaning of whole data in its analysis and to reduce the accuracy and validity for adequate models. The aim of this paper is to present some ways of handling outliers in given data and to investigate the effect of the analysis result before and after outlier reject. As a variety of methods has been proposed, we sellect the linear regression analysis and two linear programming techniques and compare to each result.

1. 서론

일반적으로 회귀분석(regression analysis)이 자료(data)들의 관계를 비교적 정확히 반영하여 설명이나 예측하고 있지만, 그러면서도 문제시 되는 것은 보다 정확한 방법이 없겠는가와 이상치(outlier) 및 다공선성(multicolinearity) 등에 의해 어떤 영향을 받을 것인가 하는 것이다.

특히 이상치의 발생은 현실적으로 자료를 구할 때 흔히 발생하는 것으로 기록의 오류, 환경등 치명적인 여러 요인이 원인인 것으로 생각 된다. 단순회귀 분석(simple regression analysis)인 경우를 제외하고는, 단순히 자료의 검사만으로는 이상치가 포함되어 있는지의 여부를 판별하기가 어렵다. 따라서 이상치를 다루는 한가지 방법으로는 이상치로 생각되는 자료를 제거하고 분석하는 것을 생각할 수 있다.

본 연구는 예에서 언급되겠지만 선형적인 관계를 가진 것으로 생각되는 단일 변수에 대한 단순선형 회귀분석을 실시하고 정확성 측면과 이상치의 제거 전·후의 결과를 비교하면서 이상치가 미친 영향정도를 분석하게 된다.

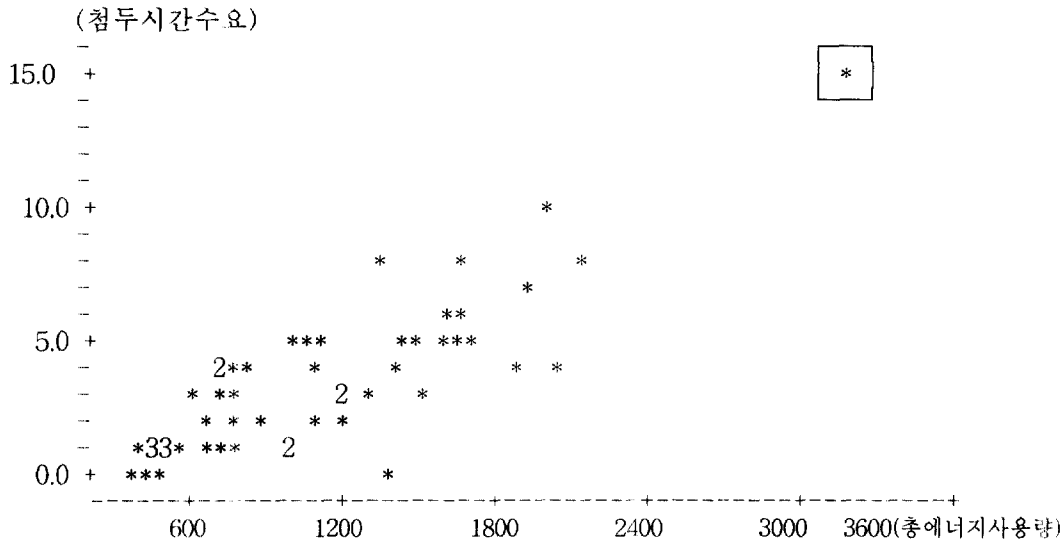
또한 회귀분석에 선형계획법(linear programming)을 적용할 수 있다는 것은 Wagner (1959), Ignizio(1982) 등에서 볼 수 있고, 예측의 정확성 면에서는 회귀분석보다는 선형계획법이 우수하다고 언급되어 있다. 그에 대한 실제 문제의 적용을 통하여 선형계획법에 의한 분석이 회귀분석에서 이상치의 영향에 대한 차이 등의 비교도 동시에 이루어 진다. 그 외에 회귀분석에 의해 얻은 결과와 선형계획법을 이용하여 얻은 결과들 간의 정확도등 몇가지 점에 대해서 논한다.

결과를 분석하는 방법으로는 이상치 제거 전과 후에 대하여 Montgomery와 Peck(1982)의 예에대해서 3장의 LP모형화와 함께 간단히 알아보고, 또한 실제 현장의 데이타인 운수회사의 이동거리 대 이동시간의 관계에 대해서는 4장에서 이상치로 생각되는 자료를 제거하기 전과 후에 얻은 각각의 결과들에 대해서 관측치와 각 방법들의 예측 편차를 비교하는 것과 편차의 산포에 대해 다중 산포도(multiple scatter diagram)로서 그 정도를 알아보고자 한다.

2. 회귀분석과 이상치

하나의 변수가 관측치와 선형적인 관계가 있다고 할 때의 회귀모형을 단순선형 회귀모형 (simple linear regression model)이라하며, 이때 이상치가 자료에 들어 있다면 회귀분석에 의한 예측 등 분석 결과를 상당히 왜곡 시키게 된다. 그러나 자료만을 보고서는 확실하게 이상치가 있다고 말하기는 어려우며, 이것을 보다 간편하게 알 수 있는 방법으로는 산포도를 보는 것이다.

물론 이상치라고 생각되는 데이터를 찾는데에는 몇가지 설명이 가능하겠지만 그 중의 한가지는 수집된 자료에서 특이한 값(unusual behavior data)으로 볼 수도 있고 또 다른 방법으로는 회귀식을 추정하고 난 후에 잔차의 평균과 표준편차의 관계를 비교하는 것도 생각해 볼 수 있다. 그러나 여기에서는 전자의 입장에서 즉, 산포도로 확인하여 보니 특이한 값으로 생각되는 점을 이상치로 간주한다. 이에 대해서는 Montgomery 와 Peck(1982)의 자료에 대해 <그림2-1>과 같이 산포도로서 대략적인 좌표 (3600, 15.0)의 한점이 이상치라는 것을 예상하고 연구를 수행한다.



〈그림 2-1〉 산포도에 의한 이상치의 예측

본 연구에서는 Montgomery와 Peck의 예로서, 간단히 설명하면 전기이용에 있어서 월간 총 에너지 사용량(X1)에 대한 첨두 시간 수요량(Y)과의 관계에 대한 회귀분석을 적용하며, 또한 이상치의 영향을 알아보기 위해 이상치 제거 전과 후의 예측식을 계산한다.

총에너지 사용이라는 단일변수에 대한 단순회귀분석 이므로 예측식은,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$$

이 된다.

이 문제에 대한 Montgomery 와 Peck의 추정치는 β_0=- 0.8313, β_1=0.00368, 결정계수가 0.7046(70.46%)이였다. Minitab을 이용했을 경우의 추정치는 $\hat{\beta}_0$=-0.8293, $\hat{\beta}_1$= 0.0036826이고 결정계수가 0.704(70.4%) 로서 소수점 아래 자리에서 차이가 나는데, 본 연구에서는 적합도를 알아보는 결정계수가 거의 비슷하게 나타났으므로 원래의 예에 주어진 추정치를 사용하기로 한다. 이상치 제거 전의 예측식은,

$$\hat{y} = - 0.8313 + 0.00368\ x_1$$

이며,

이상치 제거 후의 결과는

$$\hat{y} = - 0.44050 + 0.00328\ x_1$$

이다.

3. 회귀문제의 LP 모형화

회귀분석에 선형계획법을 적용할 수 있다는 것은 Wagner(1959), Ignizio(1982) 등이 그 개념을 설명한 바 있다. 이때, 두 가지 개념으로 선형계획법을 적용하기 위한 모형을 구축 할 수 있는데, 이 두가지 기법이 그 중 하나는 절대편차합의 최소화(minimize the sum of absolute deviation : MSD)와 다른 하나는 최대편차의 최소화(minimize the maximum deviation : MMD)이다. 이에대해 자세한 내용은 상기의 참고문헌과 또한 김광수 외 2인(1995)의 문헌을 참고할 수 있으며, 간단히 LP로 모형화한 결과만을 나타내면 다음과 같다.

3.1 MSD에 의한 모형화

절대편차의 합을 최소화하는 MSD를 LP로 모형화하면

$$\text{Min} \quad \sum_{i=1}^{m} (n_i - p_i)$$

$$\text{s.t.,} y_i - (\hat{\beta}_0^{+} - \hat{\beta}_0^{-} + \hat{\beta}_1^{+} x_{i1} - \hat{\beta}_1^{-} x_{i1}) + n_i - p_i = y_i, \quad \text{for all i}$$

$$\hat{\beta}_j^{+}, \ \hat{\beta}_j^{-} \geq 0 \quad \text{for } j = 0, 1$$

$$n_i, p_i \geq 0, \quad \text{for all i}$$

와 같다. 이때 ni는 음의 편차(negative deviation), pi는 양의 편차(positive deviation)를 나타내는 것으로 이들 편차를 최소화하는 것이다. 이렇게 MSD를 모형화하여 적용한 이상치 제거 전의 예측식은

$$\hat{y} = -1.09491 + 0.00382\, x_1$$

이고, 이상치 제거 후의 예측식은

$$\hat{y} = -0.75906 + 0.00307\, x_1$$

이다.

3.2 MMD에 의한 모형화

최대편차를 최소화하는 MMD에 의한 LP 모형은 최대편차 ε을 최소화 시키는 것으로 다음과 같이 모형화 할 수 있다. 즉,

$$\text{Min} \quad \varepsilon$$

$$\text{s.t.,} \quad \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \varepsilon \geq y_i, \quad \text{for all i}$$

$$\hat{\beta}_0 + \hat{\beta}_1 x_{i1} - \varepsilon \geq y_i, \quad \text{for all i}$$

$$\varepsilon \geq 0$$

와 같다. 이때의 ε도 물론 편차를 나타내는 것으로 이를 최소화하는 모형이다. 이

방법에 의해 계산된 이상치 제거 전의 예측식은

$$\hat{y} = 0.00305\ x_1$$

이고, 이상치 제거 후의 예측식은

$$\hat{y} = -0.33459 + 0.00252\ x_1$$

이다.

이들을 이상치 제거 전과 후의 경우로 간략히 설명해 보면 다음과 같다. 즉, 이상치 제거 전의 경우 잔차의 값이 적으므로 보다 정확하게 예측되었다고 생각할 수 있는 비율을 계산해 볼 수 있는데, 회귀분석의 예측치에 비해서 MSD는 58.5%(31/53)가 되고, MMD는 41.5%(22/53)이였다. 이 결과로는 MSD가 다른 방법에 비해 상대적으로 정확하게 예측된 것으로 보여지며 또한 편차의 절대값의 합에 의해서도 MSD가 다소 정확하게 예측된 것으로 볼 수 있다. 이상치 제거 후에는 MSD가 53.8%(28/52), MMD가 44.2%(23/52)로 정확성 면에서 회귀분석에 비해 MSD가 낮긴 하지만 이상치를 제거하고 난 후의 결과가 상대적으로 낮아졌고, MMD는 낮긴 하지만 반대로 더 정확해 졌다고 볼 수 있다. 편차의 절대값의 합을 비교한 경우 큰 차이가 있는건 아니지만, 이 경우에는 회귀분석이 MSD 보다 더 작은 값을 나타냈다.

다음으로는 편차의 산포를 비교하는 것으로 이상치 제거 전에는 대체적으로 회귀분석과 MSD에서 분산의 안정성이 다소 좋은 것으로 나타났다. 그러나 이상치 제거 후의 분산의 안정성 면에서는 MMD가 상당한 정도 우수한 것으로 보여지지만 그렇다고 일반적으로 MMD가 우수하다고 볼 수는 없으며 다만 본 예의 경우에 대한 결과이다.

적합성을 설명하기 위해서는 분산분석이나 결정계수의 검토가 필요하다[김광수 외 2인, 1993]. 여기에서는 그 보다는 이상치 제거 전과 후의 산포 만으로 설명을 한다. 참고로 회귀분석의 경우 이상치 제거 전의 결정계수는 0.7046, 제거 후 0.601 인데, 이 결과를 보아도 중대한 영향을 미쳤을 것으로 생각할 수 있다. 각 방법에 대한 이상치 제거 전과 후의 산포로서 판단해 보면 회귀분석에서 이상치 제거 전과 후의 경우 이상치 제거 후에 분산의 안정성이 더 좋아졌다고 볼 수 있다. 그리고 MSD에서는 이상치 제거 전과 후의 차이로서 분산의 안정성이 좋아졌다고 판단하기어려운 결과를 얻었고, MMD에서는 이상치 제거 전과 후의 산포를 비교해본 결과는 분산의 안정성이 매우 좋아졌다고 볼 수 있다.

이들 결과는 Montgomery 와 Peck(1982)의 예를 분석한 것으로 다음장에서는 이와 같은 순서의 내용으로 모 운수회사에서 얻은 실제 데이타에 대하여 분석을 하고자 한다.

4. 이상치의 영향

분석 대상 자료중에서 다른 것과 비교하여 특히 편차가 큰것을 일반적으로 이상치라고 부르는데, 즉 관측자가 생각하기에 대부분의 데이타들과 멀리 떨어져 있는 관측치 이다. 단순회귀 분석인 경우를 제외하고는 자료의 검사만으로는 이상치가 포함되어 있는지의 여부를 알기 어렵다. 그리고 이상치를 처리하는 방법은 여러가지가 있지

만, 본 연구에서는 이상치로 생각되는 자료를 제거하기 전과 후의 결과를 분석하고 앞에서 제시한 각 방법들을 비교하고자 한다.

한가지 주의할 점은 이상치의 제거가 반드시 필요한가 하는 것은 매우 중요한 일이다. 어떤 경우 하나의 이상치가 나머지 모든 자료에서 얻는 정보 보다 귀중한 경우가 있을 수 있는데, 이것은 분석하는 사람의 판단에 관한 문제로서 Barnett 과 Lewis(1977)를 참고하면 되는데, 여기에서는 더 이상의 언급은 하지 않고 이상치가 미치는 영향에 대해서 정확성 등 몇가지 비교를 실시한다.

4.1 이상치 제거 전과 후의 정확성의 비교

이상치를 제거하기 전의 회귀분석과 선형계획법을 적용한 결과에 대하여 관측치와 각각의 예측치에 의한 계산의 정확성을 확인하는 것이다. 확인하기 전에 회귀분석과 MSD, MMD에 의해 구한 추정식을 이상치 제거 전과 후로 정리하여 나타낸 것이 <표4-1> 이다. 물론 계산에 필요한 방법은 앞에서 간단히 언급했고, 필요한 경우 문헌을 참조하면 되겠다.

〈표 4-1〉 이상치 제거 전과 후의 각 방법에 따른 예측식

구 분	이상치 제거 전의 예측식	이상치 제거 후의 예측식
회귀분석	$\hat{y} = 53.2 + 1.034\ x_1$	$\hat{y} = 49.9 + 1.030\ x_1$
MSD	$\hat{y} = 32.3 + 1.090\ x_1$	$\hat{y} = 32.7 + 1.086\ x_1$
MMD	$\hat{y} = 217.0 + 0.767\ x_1$	$\hat{y} = 151.6 + 0.767\ x_1$

<표 4-2>에는 모 운수회사에서 운행하는 11톤 화물차량의 영차상태인 경우를 기준으로 하여 이동거리 대 이동시간 간의 관계를 갖는 자료와 관측치 그리고 각 방법에 따른 예측치와 편차의 일부를 나타내고 있다.

<표 4-2>를 보면 편차의 절대값이 작은 경우를 보다 정확하게 예측하였다고 했을 때, 그 비율로서 회귀분석에 대한 예측치에 비해 MSD는 51.6%(32/62), MMD는 21.0%(13/62)인데, 이것으로 판단 한다면 회귀분석과 MSD는 거의 비슷한 정도의 정확성을 갖는 다고 보여지나 MMD에 비해 서는 상대적으로 정확하게 예측된 것으로 보여진다. 또한 편차의 절대값의 합에 의해서도 알 수 있는 것으로 큰 차이라고 보기는 어려우나 MSD가 다소 정확하게 예측된 것으로 볼 수 있다.

<표 4-3>은 앞에서 설명된 회귀분석과 선형계획법을 적용하여 이상치 즉, <표 4-2>의 점선부분을 제거한 후의 예측식을 통하여 얻은 결과의 일부이다. 이 결과를 정확성 측면에서 검토하면 <표 4-2>에서는 MSD가 51.6%, MMD가 21.0%로 회귀분석과 MSD는 정확성 면에서 거의 같은 결과를 나타내고, MMD는 덜 정확한 것으로 나타났다. 이상치를 제거한 후에는 MSD가 50.8%(31/61), MMD가 36.1%(22/61)로 정확성 면에서 회귀분석에 비해 MSD가 좋다고 할 수 없을 정도로 같은 결과를 얻었

고 오히려 이상치를 제거하고 난 후의 결과가 상대적으로 낮아졌고, MMD는 낮긴 하지만 반대로 이상치 제거 전 보다는 보다 더 정확해 졌다고 볼 수 있다.

또한 참고할 만한 내용은 <표 4-2> 와 <표 4-3>의 편차 절대값의 합을 비교할 경우, 이상치 제거 전과 후에 있어서 모두 큰 차이는 없지만 회귀분석보다 MSD가 보다 더 작은 값을 나타내고 MMD에 비해서는 상당히 작다는 것을 알 수 있다.

〈표 4-2〉 이상치 제거 전의 회귀분석과 선형계획법의 결과비교(일부)

이동거리 (km)	이동시간 (분)	회귀분석		MSD		MMD	
		예측치	편차	예측치	편차	예측치	편차
416	390	483.292	93.292	485.594	95.594	536.801	146.801
399	390	465.716	75.716	467.070	77.070	523.731	133.731
299	330	362.326	32.326	358.110	28.110*	446.851	116.851
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:
287	620	349.919	-270.081	345.035	-274.965	437 626	-182.374**
128	150	185.529	35.529	171.789	21.789	315.386	165.386
$\Sigma \lvert \varepsilon_i \rvert$ for all i			2945.7		2928.0		6335.6

* MSD의 편차가 회귀분석의 편차보다 적은것(62개 관측치중 32개)
** MMD의 편차가 회귀분석의 편차보다 적은것(62개 관측치중 13개)

이상치로 간주한 자료의 결과

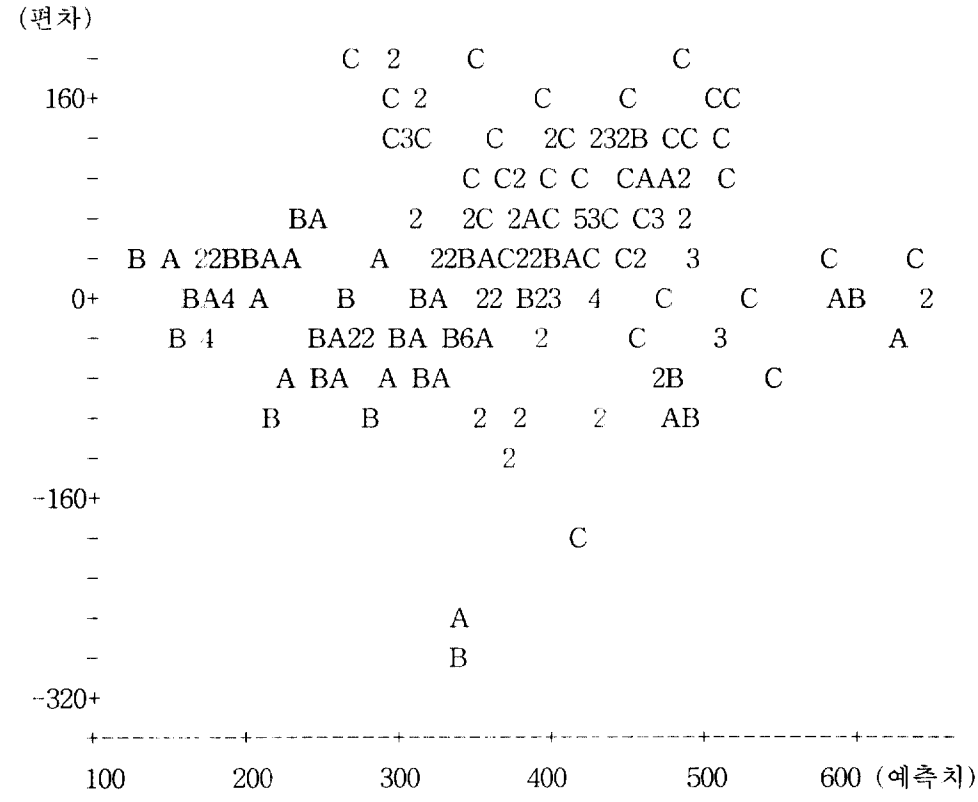
〈표 4-3〉 이상치 제거 후의 회귀분석과 선형계획법의 결과비교(일부)

이동거리 (km)	이동시간 (분)	회귀분석		MSD		MMD	
		예측치	편차	예측치	편차	예측치	편차
210	300	266.157	-33.843	260.775	-39.225	312.673	12.673**
289	360	347.503	-12.497	346.569	-13.431	373.251	13.251
128	150	181.722	31.722	171.723	21.723*	249.796	99.796
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:
357	420	417.523	-2.477	420.417	0.417	425.393	5.393
441	540	504.018	-35.982	511.641	-28.359	489.804	-50.196
$\Sigma \lvert \varepsilon_i \rvert$ for all i			2671.1		2652.8		3499.5

* MSD의 편차가 회귀분석의 편차보다 적은것(61개 관측치중 30개)
** MMD의 편차가 회귀분석의 편차보다 적은것(61개 관측치중 22개)

4.2 이상치 제거 전과 후의 편차의 산포 비교

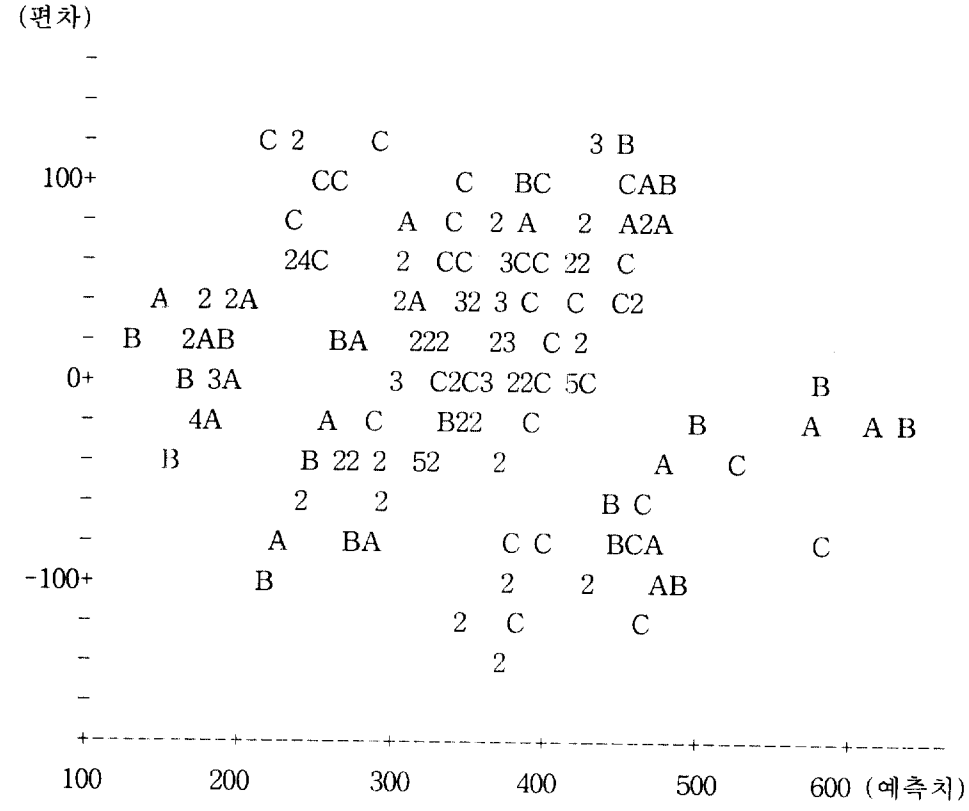
다음으로는 편차의 산포를 비교하는 것으로 단순회귀에서는 자료의 산포로서 그들의 관계를 대체로나마 알 수 있고 또한 편차에 대한 산포로서는 분산의 안정성을 검사할 수 있다. 이상치 제거 전의 자료는 <그림 4-1>에 그리고 제거 후의 경우는 <그림 4-2>에 각각의 방법들에대한 편차를 나타낸다.



〈그림 4-1〉 이상치 제거 전의 예측치와 편차의 다중 산포도
A = 회귀분석의 예측치대 편차의 산포
B = MSD의 예측치대 편차의 산포
C = MMD의 예측치대 편차의 산포

<그림 4-1>로 산포를 설명하기는 쉽지 않지만 대략적으로 문자 A와 B가 중앙에 모여 잇고, 문자 C가 양의 방향에 매우 많이 나타남을 알 수 잇다. 이 결과로 보면 회귀분석과 MSD가 MMD에 비해 분산의 안정성이 다소 좋은 것으로 볼 수 있다. 한가지 재미있는 결과는 이상치로 보이는 한점이 - 180 이하에서 나타났는데 이는 이상치에 대해서 MSD가 가장 민감하게 그리고 MMD가 비교적 둔감하게 반응한다고 생각할 수 있다.

<그림 4-2>는 이상치 제거 후의 예측치와 편차의 다중 산포도이다. 이 또한 <그림 4-1>에서와 마찬가지로 설명이 쉽지 않고 데이타들 간에 중복되어 나타나 보이지만 대략 문자 C가 중앙선을 경계로 상하로 많이 벗어나 있음을 알 수 있고 대략적인 편차로 보면 문자 A와 B가 좁게 잇는데 이는 회귀분석과 MSD가 분산의 안정성 면에서는 MMD보다 우수한 것으로 보여진다. 그렇지만 일반적으로 회귀분석과 MSD가 우수하다고 볼 수는 없으며 다만 본 예의 경우에 대한 결과로 생각 할 수 있다.



〈그림 4-2〉 이상치 제거 후의 예측치와 편차의 다중 산포도

A = 회귀분석의 예측치대 편차의 산포

B = MSD의 예측치대 편차의 산포

C = MMD의 예측치대 편차의 산포

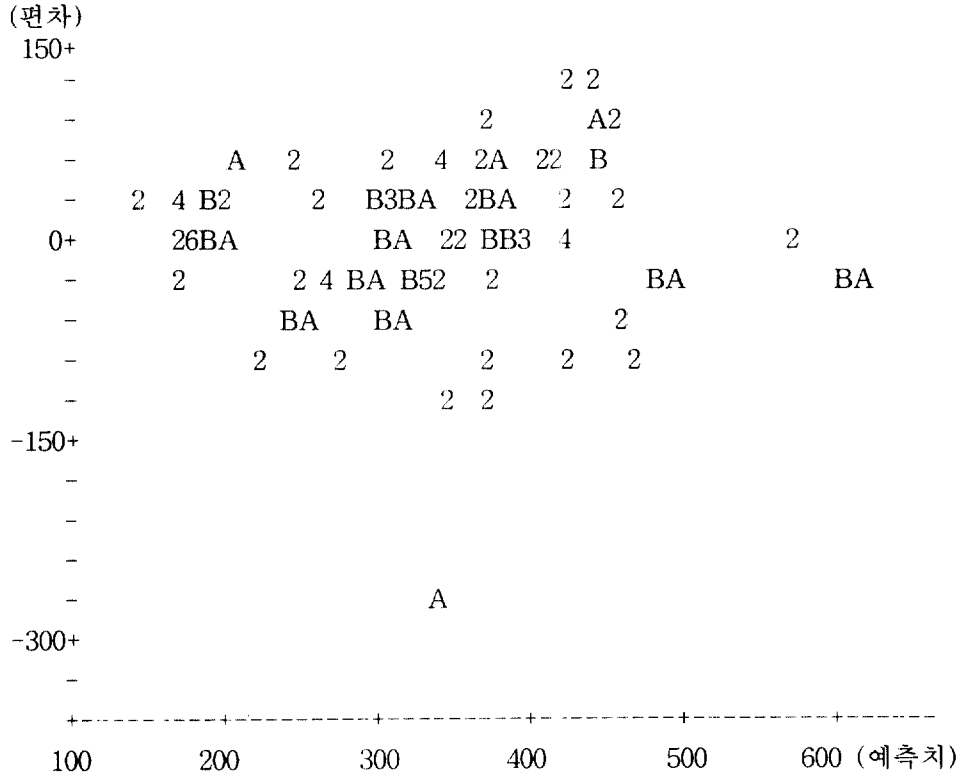
또한 위의 그림들은 매우 복잡하게 나타났기 때문에 보다 일반적인 설명이 어렵다는 점을 감안하여 두 데이타 쌍 간에 차이가 있는지 없는지를 검사해 보는 것도 의미가 있을 것으로 생각된다. 그러므로 이들 편차들 간의 차이를 D라고 했을 경우에 이들이 차이가 있는지를 검정하기 위해 흔히 사용되는 방법으로 다음과 같은 가설을

세워서 검정한 결과를 보인다.

즉, 귀무가설 $H_0 : D = 0$, **대립가설** $H_1 : D \neq 0$으로 놓고 검정해본 결과 이상치 제거 전의 회귀분석과 MSD의 경우 |t0| = 7.179 > t(0.01) = 2.389, 회귀분석과 MMD의 경우 |t0| = 24.87 > t(0.01) = 2.389 이였으며, 이상치 제거 후의 회귀분석과 MSD의 경우 |t0| = 7.131 > t(0.01) = 2.390, 회귀분석과 MMD의 경우 |t0| = 7.952 > t(0.01) = 2.390 으로 유의수준 1%에서 모두 기각되어 차이가 있는 것을 알 수 있다.

4.3 이상치의 영향에 따른 방법의 비교

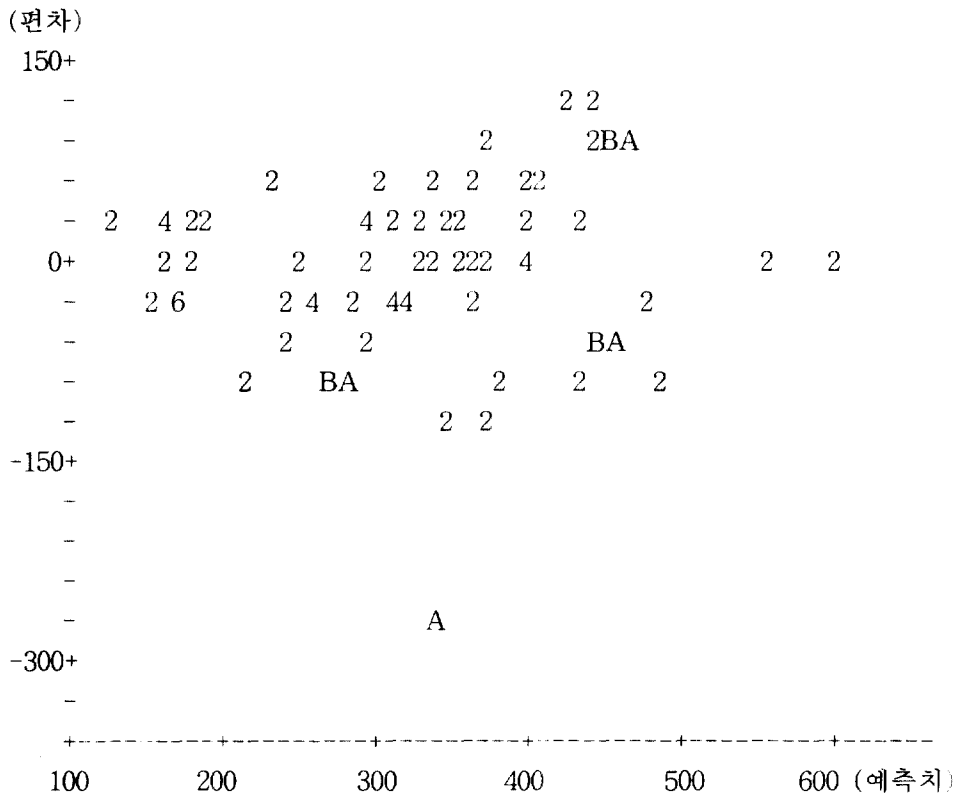
적합성을 설명하기 위해서는 분산분석이나 결정계수의 검토가 필요하다[김광수외 2인, 1993]. 여기에서는 그 보다는 이상치 제거 전과 후의 산포 만으로 설명을 한다. 참고로 회귀분석의 경우



〈그림 4-3〉 이상치 제거 전과 후의 회귀분석의 예측치와 편차의 산포도

A = 이상치 제거 전의 회귀분석의 예측치대 편차

B = 이상치 제거 후의 회귀분석의 예측치대 편차



〈그림 4-4〉 이상치 제거 전과 후의 MSD의 예측치와 편차의 산포도

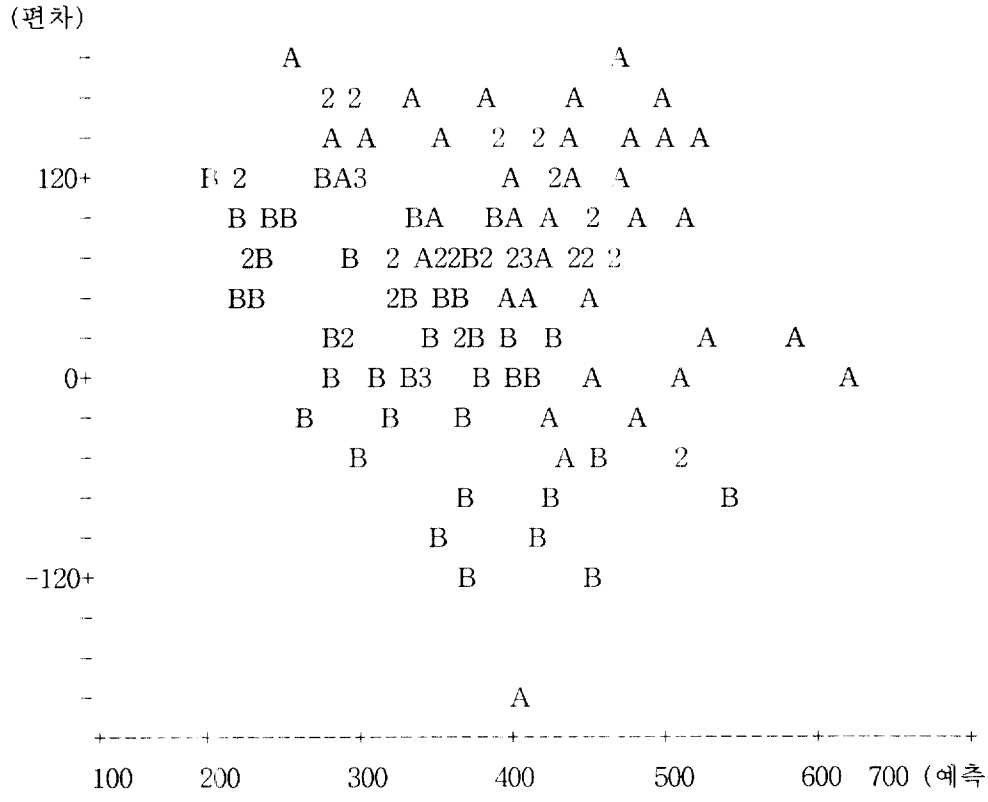
A = 이상치 제거 전의 MSD의 예측치대 편차

B = 이상치 제거 후의 MSD의 예측치대 편차

이상치 제거 전의 결정계수는 0.75, 제거 후 0.808 인데, 이 결과를 보아도 이상치가 중대한 영향을 미쳤을 것으로 생각할 수 있다. 각 방법에 대한 이상치 제거 전과 후의 산포도는 <그림 4-3 ~ 4-5>에 나타냈다.

회귀분석에서 이상치 제거 전과 후의 산포도는 <그림 4-3>로서 표준편차를 보면 이상치 제거 전에는 65, 제거 후에는 55정도로 나타 났으며, 이런 수치로 대략적인 관계를 보게되면 문자 A보다 문자 B가 평균에 보다 집중된 것을 알 수 있다. 이는 이상치 제거 후의 분산이 안정성 면에서 더 좋아졌다고 볼 수 있다.

<그림 4-4>는 MSD에서 이상치 제거 전과 후의 산포도로서 제거 전의 표준편차 65, 제거 후의 표준편차 55로 회귀분석과 거의 같은 결과를 나타내고 있으며 이 또한 위와 같은 결론을 내릴 수 있다.



〈그림 4-5〉 이상치 제거 전과 후의 MMD의 예측치와 편차의 산포도
A = 이상치 제거 전의 MMD의 예측치대 편차
B = 이상치 제거 후의 MMD의 예측치대 편차

MMD에서의 이상치 제거 전과 후의 산포도는 <그림 4-5>으로 제거 전의 표준편차는 71, 제거 후의 편차는 62 정도이며, 그림에서 볼 수 있듯 이상치 제거 전에 비해 제거 후가 분산의 안정성 면에서 매우 좋아졌다고 볼 수 있다.

이들 결과를 종합해보면 회귀분석과 MSD는 이상치 제거 전에 비해 후가 보다 안정성이 있다고 판단되지만 이들 간에는 별 차이가 없는 것으로 나타났고, MMD의 경우에서는 제거 전에 비해 제거 후가 분산의 안정성에서는 비교적 우수한 것으로 생각된다.

5. 결론

이상치가 자료를 처리하는데 있어서 영향을 미친다는 것은 정확성과 적합성 등 몇 가지 결과로서 알 수 있었다. 이상치 제거 전의 정확성 면에서는 MSD가 다른 방법

보다 비교적 우수한 것으로 나타났다. 또한 MSD가 이상치 제거 후에 회귀분석에 비해 우수하다고 하기는 어려울지 모르지만 최소한 다른 방법에 비해 나쁘지 않다고 하는 점은 확실해 보인다. 그리고 전반적으로 이상치 제거 전에 분산의 안정성 면에서 보면 회귀분석과 MSD가 좋은 것으로 보이나 제거 후가 전반적으로 더 좋아졌다고 볼 수 있으며, 특히 MMD에 대해서만 보면 이상치 제거 전에 비해 이상치 제거 후가 매우 좋아진 것으로 나타났다.

또한 참고해야 할 것은 결정계수로 설명되는 모형과의 적합도가 회귀분석의 경우만 비교한 경우에 Montgomery 와 Peck(1982)의 예에서는 이상치 제거 전은 70.46%이였으나 제거 후에는 60.1%로 낮아 졌고, 운수회사의 예에서는 이상치 제거 전에는 75.0%에서 제거 후에는 80.8%로 올라갔다는 점을 알수 있다. 그러므로 이상치의 제거가 자료를 분석하는데 반드시 필요한 것인지에 대해서는 보다 신중한 판단이 필요할 것으로 생각된다.

한편 위의 결과와는 달리 선형계획법이 통계적인 방법 보다 간단하다고 하는 견해도 있으나 Glorfeld 와 Gaither(1982)는 여러 요인 중에서 개인의 교육적 경험과 방법에 대해서 어떤 기법을 사용하는데 있어서 개인적인 선호에 따르기 때문인 것으로 보고있다. 물론 본인의 지식이나 경험으로 좋은 방법이라고 생각되는 것은 있겠지만 한가지 방법에 대해 단정적으로 우수하다고 판단하는 것은 바람직 스럽지 않다고 생각된다.

참고문헌

[1] Barnett, V. and Lewis, T. (1977), "Outliers in Statistical Data," Wiley.

[2] Cavalier, T. M. and Ignizio, J. P. and Soyster, A. L. (1989), "Discriminant analysis via mathematical programming : Certain problems and their causes," *Computers and Operations Research*, Vol. 16, No. 4, pp. 353-362.

[3] Charnes, A. and Cooper, W. W. and Sueyoshi, T. (1986), "Least squares/ridge regression and goal programming/constrained regression alternatives," *European Journal of Operational Research*, Vol. 27, pp. 146-157.

[4] Freed, N. and Glover, F. (1986), "Evaluating alternative linear programming models to solve the two-group discriminant problem," *Decision Sciences*, Vol. 17, pp. 151-162.

[5] Glorfeld, L. W. and Gaither, N. (1982), "On using linear programming in discriminant problems," *Decision Sciences*, Vol. 13, pp. 167-171.

[6] Ignigio, J. P.(1982), "Linear Programming in Single- & Multiple- Objective Systems," Prentice-Hall, pp. 243-253.

[7] Montgomery, D. C. and Peck. E. A. (1982), "Introduction to Linear Regression Analysis," Wiley, pp. 90-91.

[8] Wagner, H. M.(1959), "Linear programming techniques for regression analysis," *Journal of American Statistical Association*, Vol. 54, pp. 206-212.

[9] Weisberg, S.(1980), "Applied Linear Regression," Wiley, pp. 31-57.

[10] 강창욱(1994), "회귀모형진단", 「품질경영학회지」, 제22권, 제1호, pp. 229-230.

[11] 김광수, 정지안, 이진규(1993), "선형계획법을 이용한 회귀분석 결과의 비교연구", 「품질관리학회지」, 제21권, 제 1호, 한국품질관리학회, pp. 161-170.

[12] 김광수, 배영주, 이진규(1995), "회귀분석에서 이상치가 미치는 영향", 「제9회 아시아 품질 경영 심포지엄 발표문집」, 대한품질경영학회, pp. 662-673.