혼합물실험에서 능형추정량에 대한 붓스트랩 신뢰구간

장대흥*†

* 부경대학교 수리과학부 통계학전공

Bootstrap Confidence Intervals of Ridge Estimators in Mixture Experiments

Dae-Heung Jang*†

* Division of Mathematical Sciences, Pukyong National University

Key Words : Mixture Experiments, Ridge Estimators, Bootstrap Confidence Intervals

Abstract

We can use the ridge regression as a means for stabilizing the coefficient estimators in the fitted model when performing experiments in highly constrained regions causes collinearity problems in mixture experiments. But there is no theory available on which to base statistical inference of ridge estimators. The bootstrap could be used to seek the confidence intervals of ridge estimators.

1. 서 론

제품이 여러 개의 성분의 혼합으로 구성되어 있고 각 성분의 혼합비율이 문제가 되는 경우가 있다. 이처럼 여러 개의 성분들의 혼합물에 관한 실험에서 반응값에 유의한 영향을 끼치는 성분들을 찾고 반응을 최대 또는 최소로 만드는 최적혼합비율을 구하고자 하는 실험계획을 혼합물 실험계획이라 한다. q개의 성분들의 혼합물 실험에서 x_i를 i번째 성분의 혼합비율이라고 하면 다음과 같은 관계식을 만족하여야 한다.

$$x_1+x_2+\cdots+x_q=1,\ x_i\geq 0,\ i=1,2,\cdots,q \tag{1}$$

식 (1)을 만족하는 점 $\boldsymbol{x}=(x_1, x_2, \cdots, x_q)$의 집합을 $(q-1)$차원 심플렉스라 부른다. 혼합물실험에 대한 통계적 모형은 다음과 같은 2차 회귀모형을 주로 사용한다.

$$y=\sum_{i=1}^{q}\beta_i x_i+\sum_{i<j}^{q}\beta_{ij}x_i x_j+\epsilon \tag{2}$$

† 교신저자 dhjang@pknu.ac.kr

식 (2)를 행렬 형태로 표현하면 다음과 같다.

$$\boldsymbol{y}=X\beta+\boldsymbol{\epsilon} \tag{3}$$

산업현장에서는 단체의 전 영역이 아니고 제한된 영역에서만 실험을 하여야 하는 경우가 많다. 즉 각 성분이 $0\leq x_i\leq 1$의 모든 값을 취할 수 있는 것이 아니라 제한된 구간

$$0\leq l_i\leq x_i\leq u_i\leq 1,\ i=1,2,\cdots,q \tag{4}$$

안에서만 값을 취하는 경우이다. 이 때 제한된 영역 때문에 공선성(collinearity) 문제가 매우 자주 일어난다(Cornell, 2002). 공선성문제가 발생하면 식 (2)의 회귀계수에 대한 추정값이 매우 불안정하게 되어 심한 경우는 반응표면을 찾는 것이 무의미하게 될 수도 있다. 우리는 이러한 공선성문제를 해결하기 위하여 능형회귀(ridge regression), 주성분회귀(principal regression), 잠재근회귀(latent root regression) 등을 사용한다. 그러나 이러한 방법들을 사용하면 추정된 회귀계수들은 편의추정량들이 되고 우리는 이 편의추정량들에 대하여 통계적 추론을 할 수가 없게 된다. 예를 들어 추정된 회귀계수

들의 신뢰구간을 구할 수가 없다. 이 때 추정된 회귀계수들의 분포를 구하기 위하여 우리는 재표본 기법 중의 하나인 붓스트랩 기법(Politis 외, 1999; Montgomery 외, 2001)을 사용할 수 있다. 이 기법을 사용하면 편의 추정량에 대한 분포를 구할 수 있고 신뢰구간을 구할 수 있다. 편의추정량 중 가장 많이 사용하는 것이 능형추정량이므로 이를 중심으로 논의를 전개하고자 한다.

2. 능형추정량에 대한 붓스트랩 신뢰구간

혼합물실험에서 식 (4)와 같은 제한된 영역 때문에 공선성문제가 발생하면 식 (2)의 회귀계수에 대한 추정값이 매우 불안정하게 되므로 이를 해결하기 위하여 우리는 주로 능형추정량을 사용한다. 능형추정량은 BLUE(best linear unbiased estimator)인

$$b=(X'X)^{-1}Xy \tag{5}$$

대신

$$b_R=(X'X+kI)^{-1}Xy \tag{6}$$

을 사용한다. 여기서 k는 능형상수이고 I는 단위행렬이다. 이 능형상수를 구하는 방법들로서는 다양한 방법들이 있다. 그 중 그래픽 방법으로서 능형트레이스(ridge trace)를 자주 사용한다. 이 능형트레이스를 사용하여 적당한 능형상수를 구한 후 식 (6)을 이용하여 능형추정량을 구한다. St. John(1984)는 여러 예들을 사용하여 능형상수 k가 $0.004 \le k \le 0.005$ 정도면 적당하다고 주장하였다. 그런데 이 능형추정량은 추정된 회귀계수의 분산은 작게 하여 주나 편의 추정량이 된다. 더 큰 문제는 이 능형추정량의 분포를 모르기 때문에 우리는 이 능형추정량에 대한 통계적 추론을 행할 수 없게 된다. 우리가 이 능형추정량을 사용하려면 이 능형추정량에 대한 신뢰구간이 필요하다. 이 때 우리는 능형회귀추정량의 분포를 구하기 위하여 재표집 기법 중의 하나인 붓스트랩 기법을 사용할 수 있다. 이 기법을 사용하면 능형추정량에 대한 분포를 구할 수 있고 붓스트랩 신뢰구간을 구할 수 있다.

회귀분석에서 쓰이는 붓스트랩 방법은 크게 붓스트랩 잔차(bootstrap residuals) 방법과 붓스트랩 쌍(bootstrap pairs) 방법 두 가지가 있다. 혼합물실험에서 i번째 실험점을 $x_i=(x_{1i}, x_{2i}, \cdots, x_{qi})$라 하고 i번째 반응값을 y_i라 하자. 붓스트랩 잔차 방법은 식 (4)와 같은 제한된 영역 하에서의 혼합물실험에서 원래의 n개의 실험자료 쌍 (x_i, y_i), $i=1, 2, \cdots, n$을 이용하여 능형추정량 b_R을 구하고 이를 이용하여 잔차 $e_R=[e_{R1}, e_{R2}, \cdots, e_{Rn}]$을 구한 후 복원추출을 이용하여 붓스트랩 잔차벡터 e^*를 계산한다. 이 붓스트랩 잔차를 다음과 같이 능형회귀추정량으로 추정된 반응값에 더한다.

$$y^*=Xb_R+e^* \tag{7}$$

이렇게 구한 붓스트랩 표본 (x_i, y_i^*), $i=1, 2, \cdots, n$을 이용하여 능형회귀추정량을 구한다. 이러한 능형추정량을 m번 반복하여 구한다.

붓스트랩 쌍 방법을 소개하면 식 (4)와 같은 제한된 영역 하에서의 혼합물실험에서 원래의 n개의 실험자료 쌍 (x_i, y_i), $i=1, 2, \cdots, n$에서 복원추출을 이용하여 붓스트랩 표본 (x_i^*, y_i^*), $i=1,2,\cdots,n$을 m번 구한다.

m의 값은 능형추정량에 대한 분포를 구할 수 있을 정도의 값이면 되는 데 주로 200에서 1,000 사이의 값을 이용한다. 이렇게 구한 붓스트랩 표본을 이용하여 각 붓스트랩 표본에 대한 능형회귀추정량을 m번 구하면 능형회귀추정량에 대한 붓스트랩 신뢰구간을 구할 수 있다. 하나의 회귀계수 β에 대한 능형회귀추정량을 b_R이라 하자. 그리고 m번의 붓스트랩 표본을 이용한 능형회귀추정량 b_R^*에 대한 $100\times\alpha/2\%$ 백분위수와 $100(1-\alpha/2)\%$ 백분위수를 각각 $b_R^*(\alpha/2)$와 $b_R^*(1-\alpha/2)$라 하자. 그러면 β에 대한 $100(1-\alpha)\%$ 붓스트랩 신뢰구간은 다음과 같다.

$$2b_R-b_R^*(1-\frac{\alpha}{2}) \le \beta \le 2b_R-b_R^*(\frac{\alpha}{2}). \tag{8}$$

3. 수치 예

Snee(1975) 논문에 나타나는 윤활유 문제에서 4개의 성분이 사용되는 데 x_1은 첨가제, x_2는 성분 A, x_3는 성분 B, x_4는 성분 C이다. 반응값은 물리적 성질을 나타낸다. 4가지 성분들에 대한 제한된 영역은 다음과 같다.

$$\begin{aligned} &0.07 \le x_1 \le 0.18,\ 0.00 \le x_2 \le 0.30 \\ &0.37 \le x_3 \le 0.70,\ 0.00 \le x_4 \le 0.15 \end{aligned} \tag{9}$$

이러한 제한영역을 이루는 도형은 10개의 꼭지점, 7개의 면, 15개의 변으로 이루어진 도형이 되고 실험계획을 만들기 위한 후보점들로서 총 33개의 후보점들이 만들어진다. <표 1>은 33개의 후보점들 중 18개를 선택하여 만든 실험계획점들과 반응값을 나타낸다.

<표 1> 18개 실험계획점들과 반응값

x_1	x_2	x_3	x_4	y
0.150	0.0000	0.7000	0.150	13.89
0.180	0.3000	0.3700	0.150	7.60
0.070	0.2300	0.7000	0.000	9.45
0.070	0.0800	0.7000	0.150	12.93
0.180	0.1200	0.7000	0.000	7.38
0.070	0.3000	0.6300	0.000	8.58
0.070	0.3000	0.4800	0.150	15.65
0.180	0.0000	0.6700	0.150	11.94
0.180	0.3000	0.5200	0.000	13.99
0.180	0.0000	0.7000	0.120	15.24
0.070	0.2275	0.6275	0.075	8.24
0.180	0.1440	0.5920	0.084	13.84
0.125	0.3000	0.5000	0.075	10.08
0.130	0.0860	0.7000	0.084	11.48
0.125	0.2375	0.6375	0.000	9.64
0.130	0.1360	0.5840	0.150	11.94
0.133	0.1630	0.6170	0.087	11.25
0.180	0.1500	0.5200	0.150	14.65

이 자료를 이용하여 회귀계수를 구하면 다음 <표 2>와 같다. 그런데 $(X'X)^{-1}$의 행렬식(determinant)과 트레이스(trace)를 구하면 각각 $0.3729454699\times10^{25}$과 161378.980이 되고 조건수(condition number)를 구하면 943240.911로 공선성문제를 심각하게 일으킴을 알 수 있다. 통상 조건수가 100 이상이면 공선성문제를 일으킨다고 볼 수 있다. 이 공선성 문제를 해결하기 위하여 능형트레이스 등을 이용하면 능형상수를 $k=0.005$로 정해도 큰 무리가 없음을 알 수 있다. $k=0.005$일 때 능형회귀추정량을 구하면 <표 3>과 같다. <표 2>의 추정값과는 상당히 다름을 알 수 있다. 이러한 능형회귀를 적용하는 경우 $(X'X+kI)^{-1}(X'X)(X'X+kI)^{-1}$의 행렬식과 트레이스를 구하면 각각 0.1608147816×10^{7}과 101.076이 되고 조건수를 구하면 188.130으로 공선성문제를 상당히 완화시켰음을 알 수 있다.

<표 2> 추정된 회귀계수

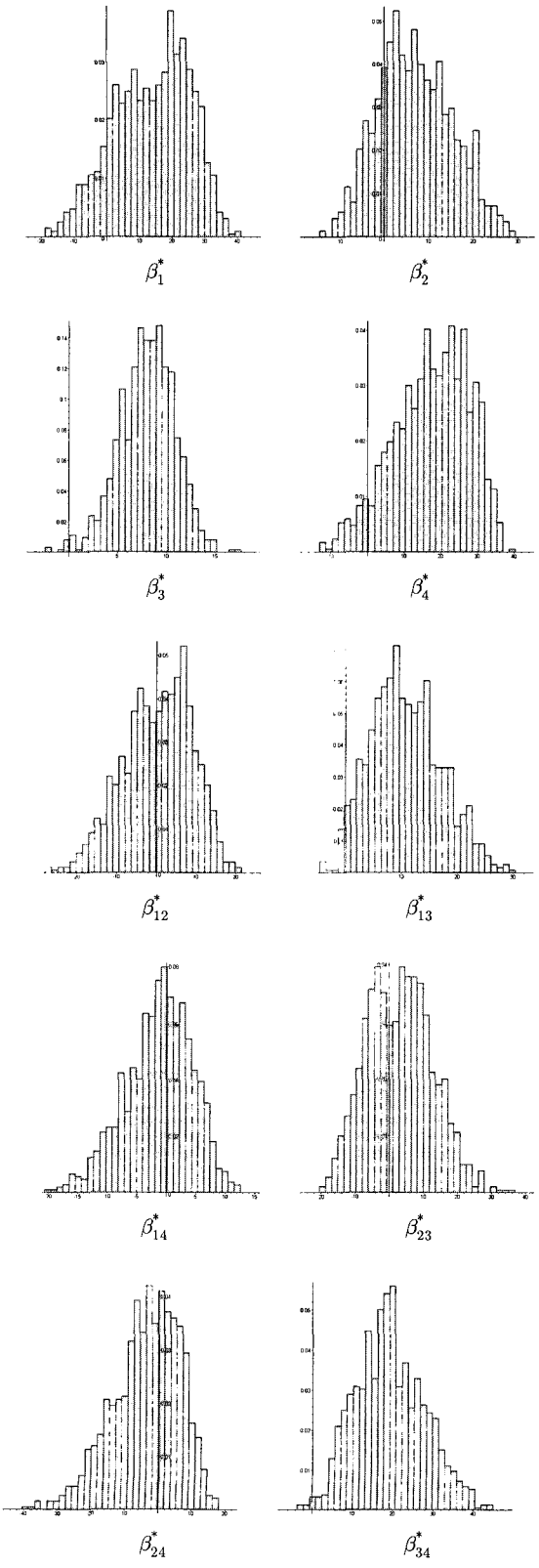
회귀계수	추정값
β_1	219.190
β_2	10.797
β_3	-10.644
β_4	47.402
β_{12}	-346.848
β_{13}	-171.426
β_{14}	-551.724
β_{23}	66.637
β_{24}	-121.746
β_{34}	107.075

<표 3> 능형회귀추정량

회귀계수	추정값
β_1	10.442
β_2	7.364
β_3	7.688
β_4	16.972
β_{12}	-2.100
β_{13}	13.432
β_{14}	-4.172
β_{23}	5.387
β_{24}	-7.017
β_{34}	24.917

$k=0.005$로 정한 후 원래의 18개의 실험자료 쌍 $(\boldsymbol{x}_i, y_i)$, $i=1,2,\cdots,18$에서 복원추출을 이용하여 붓스트랩 표본 $(\boldsymbol{x}_i^*, y_i^*)$, $i=1,2,\cdots,18$을 1,000번 구하였다. 이렇게 구한 붓스트랩 표본을 이용하여 각 붓스트랩 표본에 대한 능형회귀추정량을 1,000번 구하면 능형회귀추정량에 대한 붓스트랩 신뢰구간을 구할 수 있다. 각 붓스트랩 능형추정량의 분포를 보면 다음 <그림 1>과 같다. 하나의 회귀계수 β에 대한 능형회귀추정량을 b_R이라 하자. 그리고 1,000번의 붓스트랩 표본을 이용한 능형회귀추정량 b_R^*에 대한 2.5% 백분위수와 97.5% 백분위수를 각각 $b_R^*(0.025)$와 $b_R^*(0.975)$라 하자. 그러면 β에 대한 95% 붓스트랩 신뢰구간은 다음과 같다.

$$2b_R - b_R^*(0.975) \le \beta \le 2b_R - b_R^*(0.025) \qquad (10)$$



<그림 1> 붓스트랩 능형회귀추정량의 분포

식 (10)를 이용하여 능형회귀추정량에 대한 신뢰구간을 구하면 다음 <표 4>와 같다.

<표 4> 능형회귀추정량에 대한 95% 신뢰구간

회귀계수	신뢰구간
β_1	(-11.873, 30.988)
β_2	(-9.003, 22.764)
β_3	(2.348, 13.238)
β_4	(0.086, 39.143)
β_{12}	(-18.995, 13.125)
β_{13}	(3.063, 27.013)
β_{14}	(-16.604, 5.112)
β_{23}	(-11.136, 25.476)
β_{24}	(-26.568, 11.164)
β_{34}	(12.753, 45.328)

4. 결 론

혼합물실험에서 제한된 영역 때문에 공선성문제가 발생하면 회귀계수에 대한 추정값이 매우 불안정하게 되므로 이를 해결하기 위하여 우리는 주로 능형추정량을 사용한다. 이 때 붓스트랩 기법을 사용하면 능형추정량에 대한 붓스트랩 신뢰구간을 구할 수 있다.

참 고 문 헌

[1] Cornell, J.(2002), *Experiments with Mixture*, 3rd ed., John Wiley & Sons, Inc., New York.

[2] Montgomery, D. C., Peck, E. A., and Vining, G. G.(2001), *Introduction to Linear Regression Analysis*, John Wiley & Sons, Inc., New York.

[3] Politis, D. N., Romano, J. P., and Wolf, M. (1999), *Subsampling*, Springer, New York.

[4] Snee, R. D.(1975), "Experimental Designs for Quadratic Models in Constrained Mixture Spaces", *Technometrics*, Vol. 17, pp. 149-159.

[5] St. John, R. C.(1984), "Experiments with Mixture, Ill-conditioning, and Ridge Regression", *Journal of Quality Technology*, Vol. 16, pp. 81-96.