

R을 이용한 회귀분석과 실험계획법 시스템 구축¹⁾

김성수²⁾, 박희진³⁾, 조영훈⁴⁾, 오진호⁵⁾

요 약

본 연구에서는 최근에 널리 사용되고 있는 R 프로그램을 이용하여 실험계획법 중 요인배치법과 반응표면분석을 구현하였다. 특히 반응표면분석에서 직교계획, 회전계획, 기울기 회전계획을 만족하는 실험계획을 제공함으로써 상업용 프로그램의 미진한 부분을 개선하여 실험선택의 폭을 넓게 하였다.

주요용어 : R, 반응표면분석, 직교계획, 회전계획, 기울기 회전계획

1. 서 론

실험계획법은 통계학을 이해하는 데 필수적인 학문분야인 동시에 산업 현장에서 널리 이용되고 있는 통계적 분석방법 중의 하나이다. 특히 최근 6 시즈마와 연계하여 생산성 향상을 위해 필수적인 통계적 분석방법으로 인식되고 있다. 따라서, 실험계획법의 분석방법은 SAS, Minitab 등의 범용소프트웨어에서 제공할 뿐만 아니라, Design-Expert 등의 실험계획법 전용프로그램도 개발되어 있다. 그러나 상업용 프로그램들은 많은 비용을 지불해야 사용가능하며, 실용적인 측면만을 강조하여 쉽게 사용할 수 있는 간단한 분석방법을 선호하기 때문에 다양한 분석기법과 최근에 연구된 기법들을 제공하지 못하는 한계를 가지고 있다. 이러한 상업용 통계프로그램의 한계는 실험계획법에 국한된 문제가 아니라 통계학의 여러 분야에서 나타나는 공통적인 현상이다. 이에 최근 생물통계와 데이터마이닝 분야를 중심으로 R 프로그램을 통해 다양한 분석방법을 제공함으로써 상업용 통계프로그램의 대안으로 자리 잡고 있다. R 프로그램을 선호하는 이유는 공개 프로그램으로 강력한 수치해석, 통계계산, 그래픽 기능을 특징으로 하고 있기 때문이다. 이러한 특징은 연구자에게는 편리한 과학계산을 가능하게 함으로써 다양한 방법과 최신의

1) 본 연구는 한국과학재단 목적기초연구 지원으로 부분적으로 수행되었음.(R01-2003-000-10220-0)

2) (110-791) 서울시 종로구 동숭동 169번지 한국방송통신대학 정보통계학과 교수

3) (151-747) 서울시 관악구 신림동 산 56-1, 서울대학교 통계학과 박사과정

4) (151-747) 서울시 관악구 신림동 산 56-1, 서울대학교 통계학과 석사과정

5) (151-747) 서울시 관악구 신림동 산 56-1, 서울대학교 통계학과 박사과정

R을 이용한 회귀분석과 실험계획법 시스템 구축

연구방법을 쉽게 프로그램화할 수 있게 하였고, 사용자에게는 공개 프로그램이므로 누구나 자유롭게 사용할 수 있게 하였다.

따라서, 본 연구에서는 실험계획법의 여러 기법을 R 프로그램을 이용하여 구현하였다. 특히, 실험계획법의 최근 연구방법과 현장에서 필요한 분석방법 중 상업용 프로그램에서 개발이 미진한 부분을 포함하였다.

2. 프로그램 구성 및 특징

본 연구에서 개발을 목표로 하는 실험계획법 분석방법은 크게 4가지로 아래와 같다.

- 1) 요인배치법(Factorial Design)
- 2) 반응표면분석(Response Surface Analysis)
- 3) 혼합물 실험(Mixture Design)
- 4) 다구치 파라미터 설계(Taguchi Parameter Design)

각 분석방법은 실험설계부분(Design Part)과 분석부분(Analysis Part)로 나누어지며, 현재 요인배치법과 반응표면분석은 개발이 완료되었으며, 혼합물 실험과 다구치 파라미터 설계는 이후 개발 계획에 있다. 현재 개발 완료된 요인배치법과 반응표면분석을 중심으로 간단히 살펴보도록 하겠다.

2.1 요인배치법(Factorial Design)

요인배치법의 실험설계부분과 분석부분을 살펴보자. 먼저 실험설계부분에서 아래와 같이 구성되어있다.

- 완전설시법(Full Factorial Designs)
- 일부설시법(Fractional Factorial Designs)
- 일반화 완전설시법(General Full Factorial Designs)
- Plackett-Burman 설계

완전설시법과 일부설시법은 2, 3수준 요인배치법을 가능하게 해준다. 특히 일부설시법에서는 사용자가 입력한 정의대비에 맞게 모든 별명관계를 출력하여 교락관계를 쉽게 파악할 수 있게 하였으며 레졸루션(resolution), 정의대비의 형태(word length pattern)등의 정보도 출력하게 하

였다. 일반화 완전실시법은 서로 다른 수준을 가지는 2~9개의 인자가 있을 경우 완전실시법으로 실험설계를 하며, Plackett-Burman 설계는 사용자가 지정한 인자수(2~47)와 실험횟수(12~48)에 맞는 실험설계를 해준다.

본 연구에서 개발한 요인배치법의 실험설계부분을 MINITAB의 요인배치법과 비교해보면, MINITAB에서는 2수준 완전실시법과 일부실시법만 가능한 반면, 본 연구에서 개발한 프로그램은 3수준 요인배치법도 가능하게 했으며, 일부실시법에서는 레졸루션(Resolution)과 정의대비의 형태(word length pattern)를 제공함으로써 최적실험설계를 선택하기 위한 정보를 제공하는 장점을 가지고 있다. 일반화 완전실시법과 Plackett-Burman 설계는 MINITAB과 동일하다.

분석부분은 다음과 같다

- 분산분석법(ANOVA)
- 절반확률도방법(Half Probability Plot Method)

분산분석법을 유의한 인자를 찾는 기본적인 방법으로 하고, 여기에 절반확률도를 이용한 그래픽 기법으로 사용자가 오차항에 풀링시키고자 하는 인자를 하나씩 선택하여 최종적으로 유의한 인자를 찾는 방법도 개발하였다.

2.2 반응표면분석(Response Surface Analysis)

반응표면분석도 요인배치법과 마찬가지로 실험설계부분과 분석부분으로 나누어져 있다. 먼저 실험설계부분은 다음과 같다.

- 중심합성계획(Central Composite Design)
- Box-Behnken 계획(Box-Behnken Design)
- Equiradial Design
- 12면체계획(Dodecahedron Design)
- 20면체계획(Icosahedron Design)

반응표면분석의 실험설계부분에서 중요한 특징은 실험계획에서 실험모수(design parameter)를 직교계획(orthogonal design), 회전계획(rotatable design), 기울기 회전계획(slope rotatable design)에 만족하는 값으로 결정할 수 있게 하였다. 각 실험계획을 맞게 실험모수를 결정하는 방법은 조재일(2004), Park(1987), Park과 Kim(1992)에서 살펴볼 수 있다.

분석부분은 아래와 같이 구성된다.

- 최소제곱법을 이용한 모형추정
- 정준분석(Canonical Analysis)
- 능형분석(Ridge Analysis)
- 등고선도(Contour Plot)
- Desirability 방법

반응표면분석의 기본모형은 2차 모형을 가정하여 최소제곱법을 모수추정 방법으로 사용하였다. 최적점을 찾는 방법으로 고유값과 고유벡터를 이용한 정준분석(Canonical Analysis)과 제한된 실험영역에서 최적점을 찾는 방법인 능형분석(Ridge analysis)도 가능하게 하였다. 또한, 두 개의 실험변수에서 반응표면을 살펴보기 위해서 등고선도 작성할 수 있게 하였으며, 다반응치의 동시최적화 문제를 풀기 위한 방법으로 Desirability 함수를 이용한 분석도 가능하게 하였다.

3. 프로그램 예

본 연구에서 개발한 실험계획법 프로그램의 절차를 요인배치법에서 일부실시법과 반응표면분석에서의 중심합성계획을 중심으로 살펴보도록 하겠다.

일부실시법은 2^{6-2} 실험으로 사용하는 정의대비는 $I = ABCD = BCDF$ 이다. <그림 3.1>은 일부실시법의 실험계획을 나타낸 것이다.

<그림 3.1> 일부실시법 실험계획

```

R Console
> Design(DFFD, level=2, n.factors=6, 1=cd("ABCD", "BCDF"))
Alias matrix:
[1,] [,1] [,2] [,3] [,4]
[2,] "a" "bcd" "abcde" "e"
[3,] "b" "acd" "cde" "abe"
[4,] "ab" "cd" "acde" "be"
[5,] "c" "abd" "bde" "ace"
[6,] "ac" "bd" "abde" "ce"
[7,] "bc" "ad" "de" "abce"
[8,] "abc" "d" "ade" "bce"
[9,] "g" "abcd" "bcdef" "aef"
[10,] "af" "bcd" "abdef" "ef"
[11,] "bf" "acd" "cdef" "abef"
[12,] "abf" "cdf" "acdef" "bef"
[13,] "cf" "abd" "bdef" "acef"
[14,] "bcf" "ad" "def" "abcef"
[15,] "abd" "cf" "aef" "bcdf"
[1,] [,1] [,2] [,3] [,4] [,5] [,6]
[1,] -1 -1 -1 -1 -1 -1
[2,] -1 -1 -1 -1 -1 1
[3,] -1 -1 1 1 -1 -1
[4,] -1 -1 1 1 -1 1
[5,] -1 1 -1 1 -1 -1
[6,] -1 1 -1 1 -1 1
[7,] -1 1 1 -1 -1 -1
[8,] -1 1 1 -1 -1 1
[9,] 1 -1 -1 1 1 -1
[10,] 1 -1 -1 1 1 1
[11,] 1 -1 1 -1 1 -1
[12,] 1 -1 1 -1 1 1
[13,] 1 1 -1 -1 1 -1
[14,] 1 1 -1 -1 1 1
[15,] 1 1 1 1 1 -1
[16,] 1 1 1 1 1 1
  
```

중심합성계획은 실험변수가 2개, 중심점은 6개인 경우 회전계획을 만족시키는 축점을 이용하여 실험계획을 하였으며, 반응변수의 값은 Myers와 Montgomery(2002)의 예제 7.6을 사용하였다. <그림 3.2>는 중심합성계획의 실험계획부분과 분석자료입력을 나타낸 것이고, <그림 3.3>은 분석부분에서 모형추정 및 최적점을 구하는 정준분석을 나타낸 것이며 <그림 3.4>은 등고선도를 나타낸 것이다.

4. 결 론

본 연구에서는 실험계획법 중 요인배치법과 반응표면분석을 R 프로그램을 이용하여 구현하였다. 특히 반응표면분석에서는 사용자의 요구에 따라 직교계획, 회전계획, 기울기 회전계획을 만족하는 실험계획을 제공함으로써 상업용 프로그램에 비해 실험선택의 폭을 넓게 할 수 있었다. 하지만 2차 모형에 가정한 분석과 최적화 방법만을 구현한 한계를 가지고 있다. 따라서 다양한 모형에서의 분석방법과 최적화 방법을 보완해야하며, 실험계획법을 위한 종합적인 프로그램이 되기 위해서는 혼합물 실험과 다구지 파라미터 설계의 분석방법이 추가되어야 할 것이다.

<그림 3.2> 중심합성계획법 실험계획 및 분석자료입력

The image shows two windows from an R environment. The left window is the R Console, and the right window is a data entry window titled '실험자료입력' (Experimental Data Input).

R Console Output:

```
> x<-Design("CCD",n.factor=2)
Number of center point : 6

1. Orthogonal design
2. Rotatable design
3. Slope rotatable design
4. User's value
Select : 2
> x
      [,1]      [,2]
[1,] -1.000000 -1.000000
[2,] -1.000000  1.000000
[3,]  1.000000 -1.000000
[4,]  1.000000  1.000000
[5,] -1.414214  0.000000
[6,]  1.414214  0.000000
[7,]  0.000000 -1.414214
[8,]  0.000000  1.414214
[9,]  0.000000  0.000000
[10,] 0.000000  0.000000
[11,] 0.000000  0.000000
[12,] 0.000000  0.000000
[13,] 0.000000  0.000000
[14,] 0.000000  0.000000
```

Data Entry Window (실험자료입력):

	col1	col2	var3	var4	var5	var6	var7
1	-1	-1	80.5				
2	-1	1	81.5				
3	1	-1	82				
4	1	1	83.5				
5	-1.414214	0	75.6				
6	1.414214	0	78.4				
7	0	-1.414214	77				
8	0	1.414214	78.5				
9	0	0	83.9				
10	0	0	84.3				
11	0	0	84				
12	0	0	79.7				
13	0	0	79.8				
14	0	0	79.5				
15							
16							
17							
18							
19							

<그림 3.3> 반응표면모형 모형추정 및 정준분석 결과

```

> rsm(data, n.res=1, type="max")

Call:
lm(formula = y ~ ., data = full)

Residuals:
    Min       1Q   Median       3Q      Max
-2.36667 -2.17947 -0.01667  2.21619  2.43333

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -81.8667    1.2052   67.929 2.46e-12 ***
col1          0.9325    1.0437    0.893  0.398
col2         -0.5777    1.0437   -0.553  0.595
sq           -1.3083    1.0863   -1.204  0.263
V2           -0.9333    1.0863   -0.859  0.415
cp            0.1250    1.4760    0.085  0.935
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.952 on 8 degrees of freedom
Multiple R-Squared: 0.2827,    Adjusted R-squared: -0.1656
F-statistic: 0.6306 on 5 and 8 DF,  p-value: 0.6827
-----

Analysis of Variance Table

Response: y
      Df Sum Sq Mean Sq F value Pr(>F)
col1   1  6.956   6.956   0.7982 0.3977
col2   1  2.670   2.670   0.3063 0.5951
sq     1 11.358  11.358   1.3034 0.2866
V2     1  6.433   6.433   0.7382 0.4153
cp     1  0.063   0.063   0.0072 0.9346
Residuals 8 69.718  8.715

-----

Eigenvalues
-0.923191 -1.318476

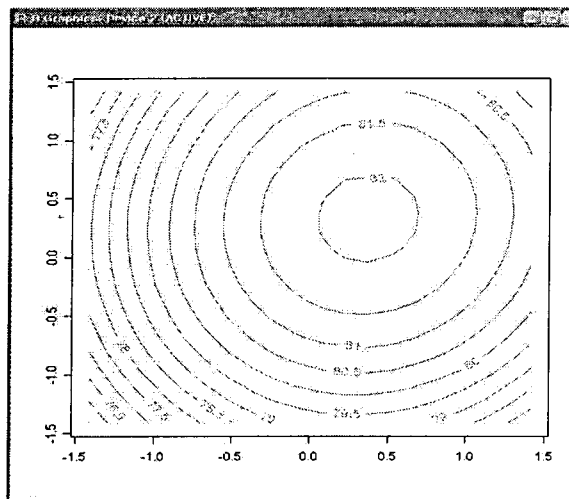
Eigenvectors
      [,1]      [,2]
[1,] 0.1601822  0.9870875
[2,] 0.9870875 -0.1601822

Canonical Point=( 0.3723341 0.3343965 )

Canonical point is maximum point

Predicted Response on the stationary point = 82.13685
    
```

<그림 3.4> 등고선도



참고문헌

- [1] 조재일 (2004), 이차반응표면 회귀모형 적합을 위한 최적 중심합성계획, 서울대학교 석사 논문
- [2] Faraway, J.J. (2005), Linear Models with R, Chapman & Hall/CRC, USA
- [3] Myers, R.H. and Montgomery, D.C. (2002), Response Surface Methodology, John Willy & Sons, Inc., USA
- [4] Park, S.H. (1987), A class of Multifactor designs for estimating the slope of response surface, Technometrics, 29, 449-453
- [5] Park, S.H and Kim, H.J. (1992), A measure of slope-rotatability for second order response surface experimental designs, Journal of Applied Statistics, 19, 391-404
- [6] Verzani, J. (2002), Using R for Introductory Statistics, Chapman & Hall/CRC, USA