한국통계학회논문집 제2권 1호
1995년 4월 pp. 55-63

Iλ-최적실험계획의 특성에 대한 추가적인 연구[1)]

김영일[2)]

Abstract

The characteristics of Iλ-optimality are investigated with repsect to other experimental design's criteria, D-and G-optimality. The comparisons are based on D- and G-, and Iλ-efficiencies using the Beta(p,q) distribution as a weighting function for Iλ-optimality. Results indicate that serious consideration should be given to the Iλ-optimality criterion especially when the error variance function is not homogeneous.

1. 소개 및 표현방법

관측치, y(x)는 다음의 선형모형을 통하여 이루어진다고 가정한다.

$$y(x)=fT(x)\theta+\varepsilon \tag{1}$$

여기서 θ는 Rp 공간의 미지의 모수 벡터이며 x는 폐구간의 실험영역, χ = [-1,1]에서의 요소이며, f(x)는 가정된 반응함수의 형태에 의존되는 px1 벡터이며 χ에서 연속이다. 오차들 {ε}은 평균 0을 갖고 있으며 독립이다. 그리고 또한 개개의 오차의 분산인 vi는 xi의 함수로 가정을 하고, v= ω-1(x)로 표기한다. 앞으로는 ω(x)와 ω-1(x)는 각각 효율함수및 오차함수로 불리도록 한다. 실험계획문제는 실험영역, χ에서 벡터 x를 선택하는 문제이다. 그리고 어떤 주어진 설정된 의미에서 이러한 벡터들로서 형성된 실험계획이 최적이 될 수 있게 끔 하는 것이 최적실험계획이다. 이러한 아이디어는 쉽게 확률을 이용한 실험계획으로 연장할 수 있다. 즉 실험계획문제는 (f,χ ω)에 대한 확률적인 실험계획 $\xi \in E\chi$ 의 선택이라 할 수 있다.

식 (1)에 관련하여 다음과 같은 정보행렬을 정의할 수 있다.

$$M(\xi) = \int\chi\ \omega(x)f(x)fT(x)d\xi(x)$$

정의 1. 실험계획 ξD는 아래 조건이 만족할때 D-최적이라 한다.

$$\max_{\xi\in E_x} |M(\xi)| = |M(\xi_D)|.$$

1) 본 연구는 1995년도 중앙대학교 학술연구비 지원에 의해 수행되었음.
2) (456-756) 경기도 안성군 대덕면 내리 산 40-1 중앙대학교 산업정보학과

정의 2. 실험계획 ξG는 아래 조건이 만족할때 G-최적이라 한다.

$$\min_{\xi \in E_x} \max_{x \in \chi} d(x, \xi) = \max_{x \in \chi} d(x, \xi_G)$$

여기서 d(x,ξ)= fT(x)M-1(ξ)f(x)는 실험계획 ξ의 주어진 x값에서의 예측치의 분산의 함수이다.

정의 3. 실험계획 ξI 는 아래 조건이 만족할 때 Iλ-최적이라 한다.

$$\min_{\xi \in E_\chi} \int_\Omega d(x, \xi)\, d\lambda(x) = \int_\Omega d(x, \xi_I)\, d\lambda(x)$$

여기서 Ω는 실험계획자가 λ라는 가중치함수를 설정할 수 있는 영역을 말한다. 그러나 꼭 Ω가 실험계획영역인 χ와 일치할 필요는 없다.

최적실험계획에서는 D-및 G-기준이 많이 쓰이는데, 이의 한 이유로서 등분산 오차함수인(ω(x) = 1)인 경우 이 두 기준의 동치를 들 수 있다. 그러나 이분산오차함수인 경우에는 동치정리가 성립되지 않기 때문에 이러한 장점은 소멸한다. 최근에 Wong과 Cook(1992)은 위에서 제시한 G-최적성을 위시한 일반적인 Mini-Max 접근방법을 이용한 실험계획을 연구하였다. 그러나 일반적으로 이분산오차함수인 경우의 G-최적은 최적의 확인절차가 복잡한 반면 Iλ-최적은 간편하다.

정의 4. D-최적실험계획 ξD를 기준으로 실험계획 ξ의 D-효율성은 다음과 같다.

$$D(\xi) = \{det\ M(\xi_d)^{-1}\ det\ M(\xi)\}^{\frac{1}{p}}.$$

정의 5. G-최적실험계획 ξG을 기준으로 실험계획 ξ의 G-효율성은 다음과 같다.

$$G(\xi) = \max_{x \in \chi} d(x, \xi_G) / \max_{\xi \in \chi} d(x, \xi).$$

정의 6. Iλ-최적실험계획 ξI를 기준으로 실험계획 ξ의 Iλ-효율성은 다음과 같다.

$$I_\lambda(\xi) = \int_\Omega d(x, \xi_I)\, d\lambda(x) / \int_\Omega d(x, \xi)\, d\lambda(x)$$

실험계획자가 예측치의 분산함수, d(x,ξ)의 가중치를 반영하기 위한 함수로 다양한 함수형태를 제시하는 Beta(p,q)함수를 생각할 수 있음은 이미 본인이 밝힌 바 있다(1993).

$$\lambda'_{p,q}(x) = 1/(b-a)^{p+q-1}\Gamma(p+q)/\Gamma(p)\,\Gamma(q)\,(x-a)^{p-1}(b-x)^{q-1}:$$
$$a<x<b,\quad p>0,\quad q>0$$

여기서 λ'p,q(x)는 dλ'p,q(x)/dx를 의미한다.

최적실험기준들간의 비교는 어렵기도 하거니와 비교 자체이외의 다른 생산적인 의미를 갖지 않을 수 있다(1975, Kiefer). 특수한 경우를 제외하면 질적인 평가도 제한적인 범위내에서 제시되는 예의 범위에서만 받을 수 있다. 그럼에도 불구하고 이러한 비교는 사용자로 하여금 최적기준의 특성들을 직접적으로 비교하여 주는 장점을 갖고 있기 때문에 때때로 시도되고 있다. 2절및 3절에서는 실험계획자들이 흔히 쓰는 다항회귀모형을 기준으로 Iλ-최적기준의 특징을 파악하고자 한다. (참고로 Beta(p,q)의 특성상, p > q (p < q)면 영역 오른쪽 (왼쪽)에 더 많은 가중치를 두는 형태가 발생하고 p=q=1 인 경우는 일양함수, p=q<1 이면 양극에 가중치를 두는 함수 p=q >1이면 영역 중심에 가중치를 두는 형태가 발생한다.)

2. Beta(p,q)의 p와 q가 같고 오차가 등분산인 경우

2.1 다항회귀모형에서의 Iλ-최적의 D-(G-)효율성: p=q=1인 경우

가중치 함수인 Beta함수가 Ω = [-1,1]에서 p=q=1인 일양함수일때 fT(x)= (1,x,..,xn) 이면 표 1에서와 같이 각각의 질량과 효율성을 계산할 수 있는데 먼저 n차 다항함수인 경우 D-최적의 Iλ-효율성은 n이 2에서 3으로 갈때는 떨어지지만 4에서 5로 갈때는 약간 올라감을 알 수 있다. (n이 1인경우는 D-최적이 바로 Iλ-최적이 되기 때문에 제외시켰음). Iλ-최적의 D-효율성은 90%이상을 상회하고 있고 G-효율성은 n이 커짐에 따라 급격히 감소함을 알 수 있다. 이는 후에도 지적이 되지만 p=q=1인 경우는 Iλ-최적과 D-최적의 연관성이 전반적으로 G-최적에 비해 높다는 암시가 될 수도 있다. Atwood(1969)에 의하면 주어진 실험계획의 D-효율성은 G-효율성에 비해 항상 크거나 같다고 하였기 때문에 만약 n이 불확실한 경우에는 G-효율성을 높이는데 실험계획의 촛점을 맞추는 모델 로버스트적인 실험계획이 필요하다고 생각된다. 이에는 Cook과 Nachtsheim(1982)의 연구를 예로 참조할 수 있다.

<표 1> Iλ-최적의 질량및 효율성
(Iλ-최적의 질량: p=q=1)

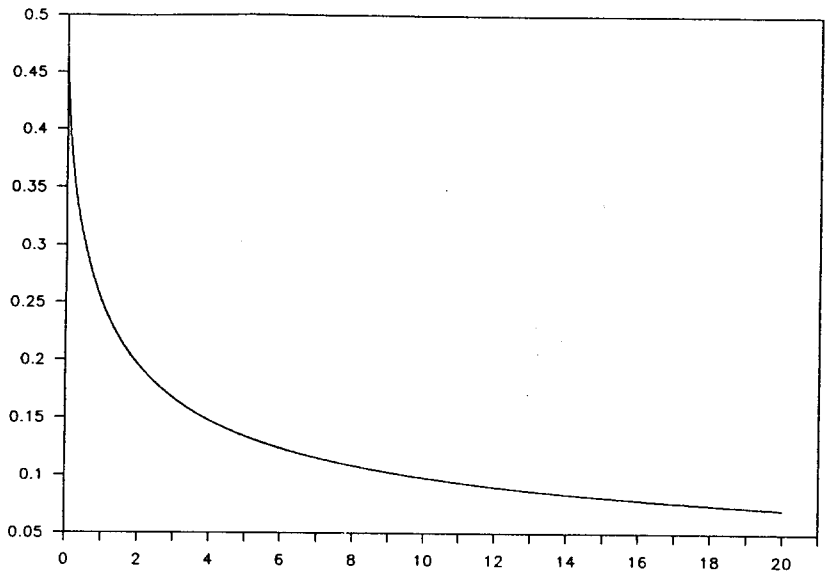
n=2(Quadratic) ξ(±1) = 0.25 ξ(0)= 0.5
n=3(Cubic) ξ(±1) = .154 ξ(±.4467) =.346
n=4(Quartic) ξ(±1) = .107 ξ(±.6543) =.250 ξ(0) = .286
n=5(Quintic) ξ(±1) = .0797 ξ(±.765) = .1903 ξ(±.2852)= .230

	$D(\xi_\lambda)$	$G(\xi_\lambda)$	$I(\xi_D)$
n=2(Quadratic)	94.49%	75.00%	88.88%
n=3(Cubic)	92.36	61.80	87.21
n=4(Quartic)	91.43	53.60	87.02
n=5(Quintic)	90.99	47.90	87.20

2.2 이차형식의 회귀모형에서의 Iλ-최적의 D-(G-)효율성: p=q인 경우

2.2.1 질량의 변화

Iλ-최적은 이차형식의 회귀모형인 경우 (p=q=1) $\xi(\pm1) = 1/4$, $\xi(0) = 1/2$ 을 배치한다(표 1참조). p=q가 1보다 커지면 점점 중앙에 더 많은 질량을 배치하고 p=q가 1보다 작아지면 $\xi(0) = 1/2$에서 작아짐을 알 수가 있다. 이에 착안하여 D-최적에 가장 가까운 p=q을 수치해석적인 방법(분석적인 방법으로는 해가 불가능)을 통하여 p=q의 값은 약 .33임이 됨을 알 수 있는데 여기에서는 p=q가 0에서 멀어짐과 동시에 $\xi(\pm1)$의 값이 어떻게 변하는지 그림을 통하여 알아 보았다.



<그림 1> p=q가 변함에 따른 x=-1에서의 질량 : 이차형식 회귀모형 : 등분산오차함수

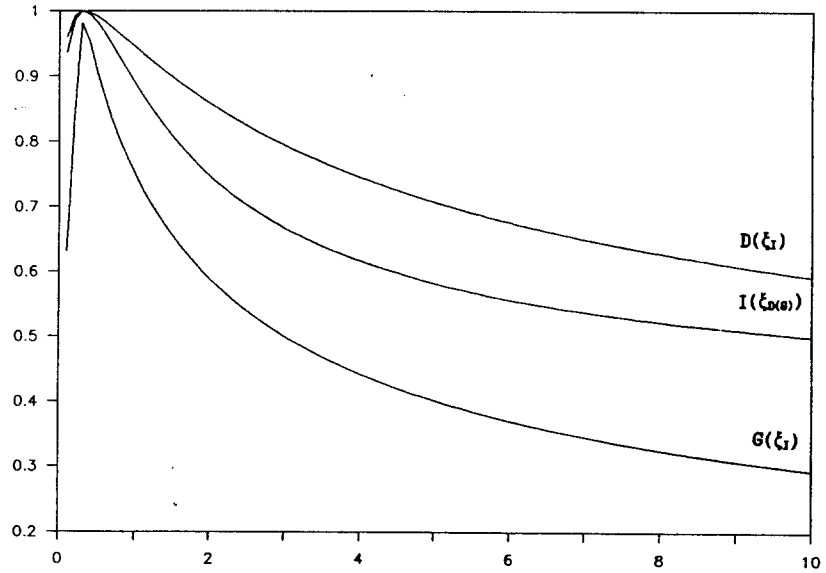
p=q의 값이 .33을 기준으로 $\xi(\pm1)$의 값은 1/3을 갖고 있음을 확인 할 수 있다. 즉 D-최적은 Iλ-최적의 관점으로 보았을 때는 양극점 -1과 1의 주위에 가중치를 부여하는 실험기준임을 알 수 있다. 그리고 특이할 만한 사항은 p=q가 커짐과 동시에 $\xi(\pm1)$의 값은 급격하게 질량의 감소를 가져다 주지만 p=q가 어느 정도 커지면 $\xi(\pm1)$의 값은 안정적인 최소한의 값을 유지하고자 한다. 물론 p가 무한대로 가까워지면 $\xi(\pm1)$의 값은 0에 근접하겠지만 p가 20주위의 값이라도 대체로 0.05 이상의 값을 유지하려고 한다. 이의 의미는 실질적으로 Iλ-최적의 실험기준을 쓰고 실험영역중심에 집중적으로 가중치를 행사한다 하더라도 양극점에서의 질량은 p=q=10을 기준으로 총20%는 남겨두어야 함이 바람직하다고 볼 수 있다.

이러한 분석은 3차다항식의 회귀모형이상에서는 수학적인 계산과정의 복잡성때문에 어려움이 많다. 그러나 p=.33을 이용한 Iλ-최적의 질량을 3차다항식의 회귀모형에 대해 계산하여 보아도 역시 이차형식의 회귀모형의 경우와 마찬가지로 D-최적과 ($\xi(\pm1) = \xi(\pm1/\sqrt{5}) = 1/4$) 거의 일치하는 것을 알 수 있다. 이를 바탕으로 추정하건데 다항식의 회귀모형을 모델로 Iλ-최적을 설정할고 그리고 대칭인 가중치 함수를 염두에 두면 p=q=.33을 기준으로 D-최적과 비교하면 될 것이고

p=q 가 불확실한 경우에는 일양함수인 경우에 D-효율성이 90%이상임을 기억하면 p=q=1도 좋은 기준이 될것이다.

2.2.2 p=q가 변하는 경우 이차형식의 회귀모형에서의 Iλ-최적의 D-(G-) 효율성

그림 2는 p=q가 0에서 커짐과 동시에 Iλ-최적의 D-효율성과 G-효율성이며 D-(G)최적의 Iλ-효율성이다. 후자인 경우는 오차가 등분산이므로 D-와 G-의 구분이 필요치 않는다. 결과는 2.1에서 논의되었던 내용과 유사하다. Iλ-최적의 D-효율성과 Iλ-최적의 G-효율성사이에 D-(G)최적의 Iλ-효율성이 위치하고 있다. 그리고 Iλ-최적의 D-효율성은 p=q가 약 6이상을 넘어서면 70%이하로 떨어짐을 알 수 있고 Iλ-최적의 G-효율성은 40%선까지 떨어진다. 이로보아 이차형식의 회귀모형에서 p=q가 불확실하게 설정되는 상황에서는 효율성면에서 문제가 있다고 보여진다. 이러한 경우에는 등분산의 D-와 G-최적의 동치를 감안하여 D-최적을 선택하는 것도 바람직 할 수 있다.



<그림 2> p=q가 변함에 따른 효율성 : 이차형식 회귀모형 : 등분산오차함수

3. 오차가 이분산인 경우($\omega(x)\neq 1$)

3.1 이분산오차함수와 Beta(p,q)에서의 p및q의 관계

p=q가 같은 경우는 단순회귀모형이나 이차형식의 회귀모형에서 이미 그 특성을 알아 보았으나 p와 q가 같지 않은 경우는 복잡한 양상을 띈다. 이차형식의 회귀모형에서는 중간받힘점이 더 이상 0이 되지 않음으로 분석적인 방법으로서는 이분산 오차함수와의 관계를 따질 수 없게 되어 여기서는 단순선형회귀식만을 갖고 분석하고자 한다.

오차항에 대한 분산의 지식은 다음과 같다고 가정을 하자. 제일 작은 분산의 값과 제일 큰 값의 비는 고정된 값 $\gamma > 1$이면 실험영역 χ에서는 분산은 선형으로 증가하고 $0 < \gamma < 1$이면 실험

영역 χ = [-1,1]에서는 분산은 감소한다. 이를 수학적으로 규정하면 ω-1(x) $\propto$ (γ-1)x + (γ+1)이다. 등분산(γ=1)일 경우 문제의 단순성을 위하여 ω-1(x) = [(γ-1)x + (γ+1)]/2로 한다.

정리 1. fT(x)= (1,x), Beta(p,q)이면 x=-1에서의 Iλ-최적의 질량, ξ(-1)과 p와 q는 다음의 관계가 성립한다.

$$(p2+p)/(q2+q) = (\xi(-1)-1)2/\xi(-1)2$$

증명. Fedorov(1972)의 동치정리를 이용하면 된다.

정리 1에서 당연히 p=q이면 단순회귀식에서의 Iλ-최적은 D-최적과 같아 지지만 p>q이면 가중치가 실험영역 왼쪽에 쏠림으로 x =1 에 .5보다큰 질량이 부여되고 p<q이면 그 반대이다. 이러한 맥락에서 위에서 정의한 이분산의 경우의 G-최적의 x=-1과 1에서의 질량은 각각 1/(γ+1), γ/(γ+1)임을 비교하면 p와 q의 변화에 따른 이분산 G-최적의 상관성을 따질 수 있을 것이다. 예를 들어 p=5와 q=2인 경우는 계산에 의해 γ = 2.236에 해당이 된다고 할 수 있다.

3.2 단순선형회귀식에서 Iλ-최적의 특성 (p=q인 경우)

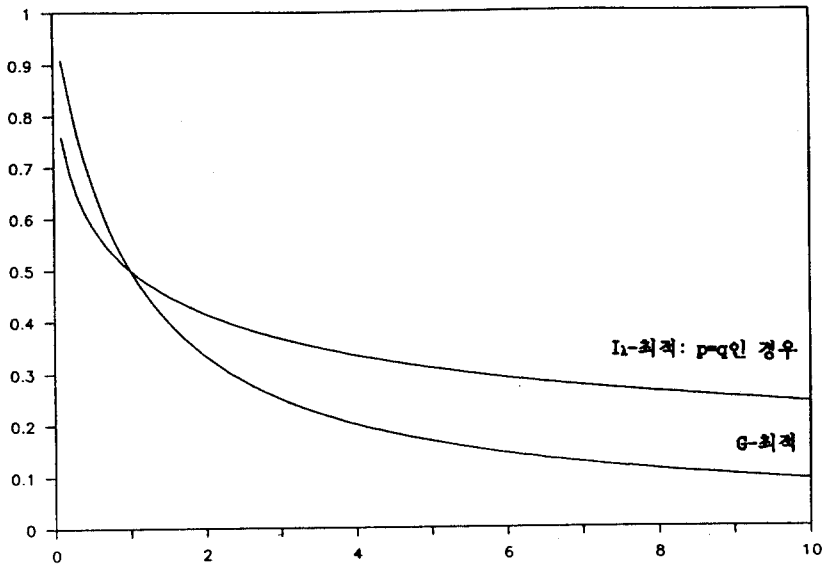
오차가 이분산인 경우는 위의 p와 q가 다른 경우와 마찬가지로 이차형식의 회귀모형에서는 중간받힘점이 0이 아니므로 분석적인 방법으로서는 Iλ-최적의 해를 일반화시키기 어렵기때문에 단순 선형회귀식에 국한하여 특징을 살피기로 하자. 그러나 설사 단순 선형회귀식이라 할지라도 이분산과 p와 q가 다른 경우가 혼재되어 있는 경우는 실용성의 문제상 제외키로 한다.

정리 2. 단순선형식에서의 오차함수, ω-1(x) = [(γ-1)x + (γ+1)]/2이면 γ에 따른 Iλ-최적의 x = ±1에서의 질량, ξ(±1)은 p와 q가 같은 경우는 다음과 같다.

$$\xi(-1) = (\sqrt{\gamma}-1)/(\gamma-1) \text{및} \quad \xi(1) = (\gamma-\sqrt{\gamma})/(\gamma-1)$$

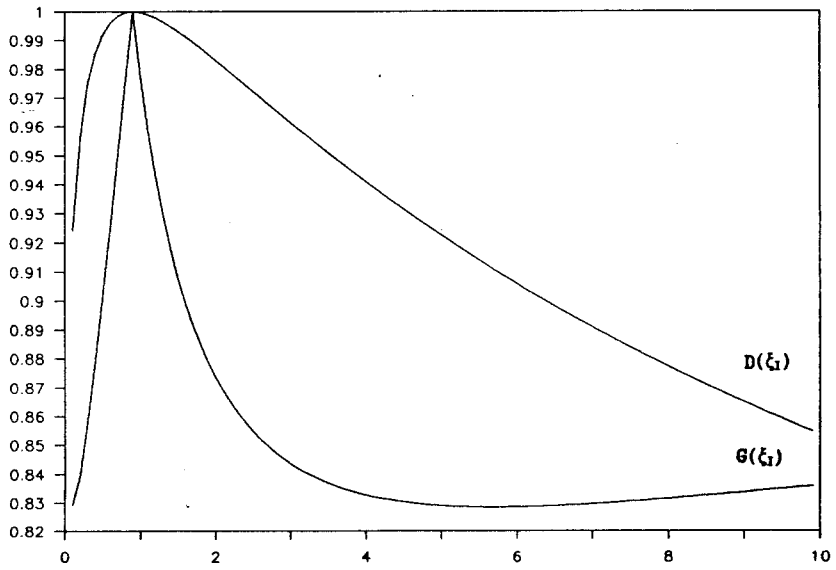
예를 들어 γ가 4이면 x=-1에서의 질량은 1/3이다. 이는 이분산 G-최적실험기준일때 γ=2인 경우에 해당한다.

정리 2를 이용하여 그림 3에서는 x=-1에서의 질량을 G-최적의 ξ(-1), 1/(γ+1)과 비교하여 보았다. Iλ-최적의 특징으로서 이분산하에서는 p와 q가 같은 경우 p=q < 1인 구간에서는 Iλ-최적의 x=-1에서의 질량이 G-최적의 질량보다는 상대적으로 작고 전구간에서의 질량의 감소율은 완만함을 알 수 있다.



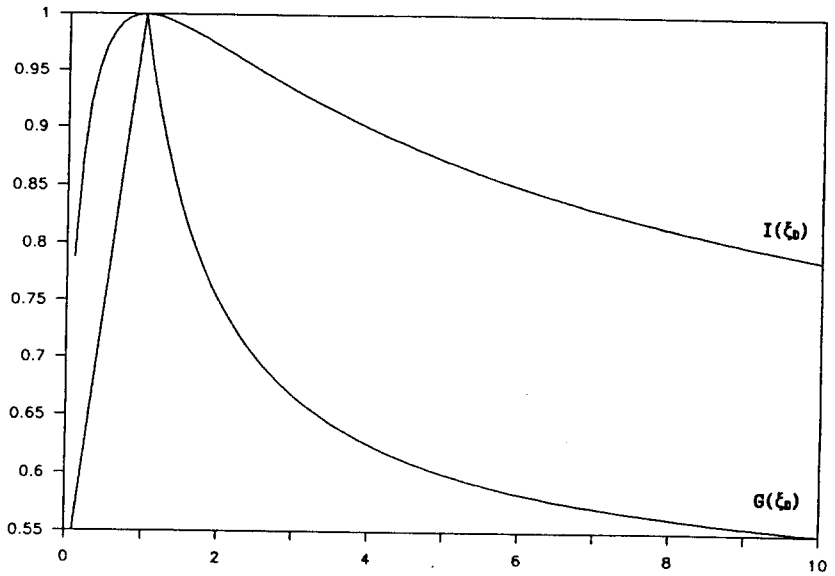
<그림 3> γ가 변함에 따른 x=-1에서의 질량 : 단순선형 회귀모형

그림 4에서는 Iλ-최적의 D-효율성및 G-효율성을 계산하였다. 80%이상의 효율성을 유지하며 γ가 커지면 G-효율성은 상승함을 알 수있다. γ가 커지면 이분산 G-최적이나 Iλ-최적이나 다같이 x=1에 질량을 치우치기 때문이다.

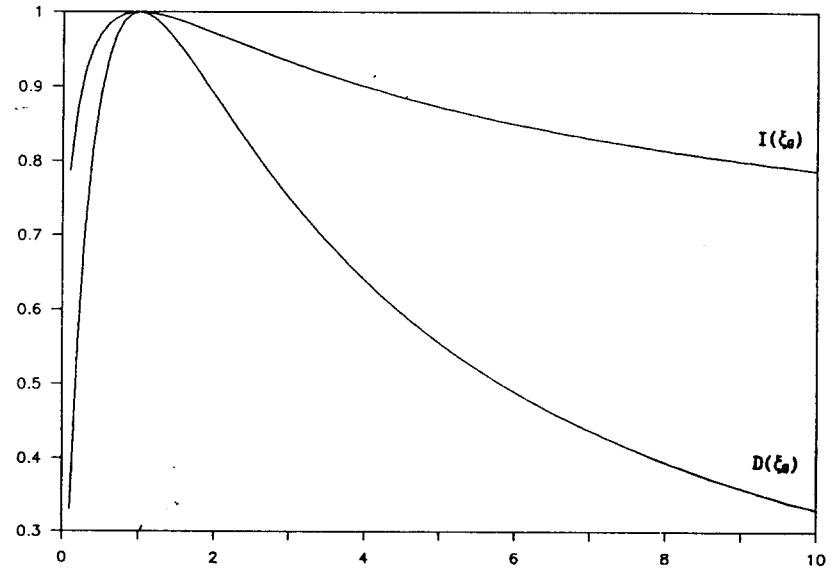


<그림 4> γ가 변함에 따른 I_λ-최적의 효율성 : 단순선형 회귀모형 : p=q

그림 5에서는 D-최적의 Iλ-효율성및 G-효율성, 그리고 그림 6에서는 이분산 G-최적의 D-효율성및 Iλ-효율성을 γ이 변화함에 따라 그려보았다.



<그림 5> γ가 변함에 따른 D-최적의 효율성 : 단순선형 회귀모형 : p=q



<그림 6> γ가 변함에 따른 G-최적의 효율성 : 단순선형 회귀모형 : p=q

그림 4의 80%효율성에 비해 그림 5에서는 γ가 커짐에 따라 G-효율성이 55%및 그림 6에서는 D-효율성이 30%수준으로 떨어짐을 알 수 있다. 이분산 오차함수인 경우는 p와 q가 같고 단순선형회귀 모형인 경우 Iλ-최적이 효율성면에서 D-최적이나 G-최적보다 나을 수도 있다는 반증이다. 이상의 결과는 단순선형회귀모형에서 얻은 제한적이기 하지만 이차형식의 선형회귀모형에서도 통용될 수 있는지 여부는 추후 수치적인 확인작업이 이루어져야 하겠다.

4. 결론

이상 단순선형및 이차형식의 선형회귀모형을 갖고 지금까지 그 실용성이 잘 알려져 있지 않은 Iλ-최적실험기준의 특징을 본인의 1993년 논문과는 다른 각도에서 알아보았다. Wong과 Cook(1992)이 제시한 이분산 G-최적 실험기준과 비교하여 Iλ-최적의 특성을 Beta분포를 가정한 가중치함수를 이용하여 설명하였으며 그리고 각각의 D- G- 그리고 Iλ-효율성을 간단한 예제를 통한 제한적인 결과이지만 첨부하여 사용자로 하여금 특징을 이해케 하였다. Iλ-최적기준은 이분산 오차함수인 경우는 그 효율성이 뛰어남을 알 수 있었다. 그리고 G-최적성에 비해서 최적성의 확인절차가 간편할 뿐 아니라 가중치함수의 성질을 감안하면 오히려 등분산 G-최적과 동치인 D-최적이나 이분산 G-최적보다는 이분산오차함수인 경우에는 실험기준이 더 유용하다고 볼 수 있다.

참고문헌

[1] 김영일 (1993). 단순선형회귀와 이차형식을 중심으로 D-와 G-최적에 비교한 I_λ-최적실험기준의 특성, 『한국품질관리학회지』, 21권 2호, 140-155.

[2] Atwood, C.L. (1969). Optimal and Efficient Design of Experiments, *Annals of Mathematical Statistics,* Vol. 40, 1570-1602.

[3] Cook, R.D. and Nachtsheim, C.J. (1982). Model Robust, Linear-Optimal Designs, *Technometrics,* Vol. 24, 49-54.

[4] Fedorov, V.V. (1972). *Theory of Optimal Experiments,* New York, Academic Press.

[5] Kiefer, J. (1975). Optimal Design: Variation in Structure and Performance under Change of Criterion, *Biometrika,* Vol. 62, 277-288.

[6] Wong, W.K. and Cook, R.D.(1992). *Heteroscedastic G-optimal Designs,* Technical Manuscript, Dept. of Bio-Statistics, University of California, Los Angeles.