Journal of the Korean Data & Information Science Society
2014, **25**(6), 1491–1498

http://dx.doi.org/10.7465/jkdi.2014.25.6.1491
한국데이터정보과학회지

다항회귀모형에서의 추가받힘점 선택

김영일[1] · 장대흥[2]

[1]중앙대학교 경영경제대학 · [2]부경대학교 통계학과

접수 2014년 10월 11일, 수정 2014년 10월 30일, 게재확정 2014년 11월 12일

요 약

최적실험의 제일 큰 약점은 실험기준이 지나치게 모형과 그에 수반되는 가정에 의존한다는 점이다. 이는 종종 모형의 모수의 개수와 받힘점의 개수가 일치를 하는 경우로 이루어지는데 이는 가정된 모형이 참이 아닌 경우를 대비한 실험이 될 수 없다. 이런 경우 문헌에서는 가정된 다항회귀모형의 차수보다 큰 차수를 가진 다항회귀모형을 가정하고 최적실험을 제안하나 이는 $D-$효율에 근거한 관행적인 방법일 뿐이다. 본 연구에서는 O'Brien (1995)이 제안한 가정된 모형의 일반적인 이탈을 염두에 둔 추가받힘점 생성에 관하여 알아보고 단순회귀모형과 2차 회귀모형에 대한 실험들을 $D-$효율로 카타로그화 하여 실험자로 하여금 선택을 할 수 있게 하였다. O'Brien은 비선형모형에 대해 추가받힘점의 선택 방법을 제시하였지만 방법을 구현하는 데 있어 명확치 않은 기준이 있어 모수에 의존하는 비선형모형에 대한 최적실험보다는 다항회귀모형을 중심으로 심층적으로 사용방법을 알아보았다.

주요용어: 다항회귀모형, 받힘점. $D-$최적실험, $D-$효율.

1. 머리말

반응변수 $n \times 1$벡터 $y(x)$는 식 (1.1)과 같이 x의 다항회귀 (polynomial regression) 함수로 연결되어 있다.

$$y(x) = \theta_0 + \sum_{i=1}^{m} \theta_i x^i + \epsilon = f_m^T(x)\theta + \epsilon, (x \in \Omega) \tag{1.1}$$

여기서 $f_m^T(x) = (1, x, \cdots, x^m)$이며 $\theta^T = (\theta_0, \theta_1, \cdots, \theta_m)$는 추정하여야 할 모수벡터이며 오차항 ϵ는 기댓값 0 그리고, 분산은 σ^2이다. 본 연구에서는 일반성의 손실 없이 $\sigma^2 = 1$로 설정한다. 받힘점 (support point)에 해당되는 x값이 위치하는 실험영역으로 통상 $x \in \Omega = [-1, 1]$으로 간주한다. 실험설계 ξ는 실험영역 Ω내의 유한개의 k개의 점 $x_i \in \Omega, i = 1, 2, \cdots, k$에 대한 확률질량함수 $\xi(x_i), i = 1, \cdots, k$로 기술되는데 전체 실험에 쓰이는 관측값의 개수 n이고 각 받힘점에서 반복되는 관측값의 개수가 $n_i, i = 1, \cdots, k$라 한다면 개개의 $\xi(x_i)$, $i = 1, 2, \cdots, k$는 n_i/n의 형태를 가지며 $\int_\Omega \xi d(x) = 1.0$의 제약조건을 가진다. 추가로 $n \times \xi(x_i)$이 정수라는 제약조건이 주어지면 정확실험설계 (exact design), 제약조건이 없는 경우를 근사실험설계 (approximate design)라 한다. 본 연구는 근사실험인 경우를 다룬다.

[1] 교신저자: (156-756) 서울특별시 동작구 흑석로 84, 중앙대학교 경영경제대학, 교수.
E-mail: yik01@cau.ac.kr
[2] (608-737) 부산광역시 남구 용소로 45, 부경대학교 자연과학대학 통계학과, 교수.

식 (1.1)의 모형 $f_m(x)$하에서 특정한 점 $x_0 \in \Omega$에서의 분산은 식 (1.2)와 같다.

$$d_m(\xi, x_o) = f_m^T(x_o) M_m^{-1}(\xi) f_m(x_0) \tag{1.2}$$

여기서 $M_m(\xi)$는 $\int_\Omega f_m(x) f_m^T(x) \xi d(x)$로 정보행렬 (information matrix)이다. 이 정보행렬은 비특이 (non-singular) 행렬로 가정한다. 최적실험설계분야에서는 이러한 $M_m(\xi)$의 행렬식을 최대화시키는 실험설계기준으로 D-최적실험설계기준이 있고 $x_0 \in \Omega$인 경우 해당하는 식 (1.2)의 예측분산 (prediction variance) $d_m(\xi, x_0), x_o \in \Omega$의 최대값을 최소화하는 G-최적실험설계기준이 존재한다. 이러한 D-와 G-최적 사이에는 동격성 (equivalence)이 존재한다. 그리고 최적성은 $d_m(\xi, x_o), x_o \in \Omega$의 최대값이 모형의 모수의 개수와 일치하는지 여부를 통해 확인할 수 있다. 물론 식 (1.1)에서의 다항모형의 경우 모수의 개수는 $m+1$이다.

다항회귀모형인 경우 $D-$최적의 해는 이미 1950년대에 구해진 바 있다. de la Garza (1954) 및 Guest (1958)에 의해 다항회귀모형 $f_m(x)$의 차수 m을 알고 있다는 가정에서 $D-$최적실험설계 해가 제공되었다. 그러나 이러한 해의 구조는 받힘점의 개수가 다항회귀모형의 모수의 개수 $m+1$와 정확히 일치하고 각 받힘점에 부여되는 질량은 $1/(m+1)$로 등질량인 구조를 가진다. 이러한 구조 때문에 $D-$최적실험은 가정된 모형에 불확실성이 존재하는 실제적인 실험환경 하에서는 채택되기 힘들다.

따라서 Kussmaul (1969)과 Cook과 Nacthsheim (1980) 등은 가정된 모형보다 더 높은 차수의 다항회귀모형을 가정하고 실험을 하는 것을 추천하곤 한다. 그러나 이러한 방법은 두 모형간의 효율성만을 가지고 추천된 방법이므로 다소 관행적이다. 기존 연구에서는 Dykstra (1971), Evans (1979), Kim 등 (2002) 등과 같이 기존 실험에 받힘점을 추가하는 추가실험의 논문들이 존재하나 이는 적합결여검정을 염두에 두고 받힘점을 찾는 문제와는 별도의 문제이다. 또한 D_s-최적, 즉 기본 모형에서 이탈되는 차수에 해당하는 추정계수의 분산을 최소화하는 기준이 있으나 이러한 최적의 받힘점의 개수가 확장된 모형의 모수의 개수와 일치하므로 유용성이 떨어진다. O'Brien (1995) 이외에는 모형의 일반적인 불확실성을 염두에 두고 추가받힘점을 찾는 연구는 기존문헌에 없는 편이다. 최적실험의 일반적인 참고문헌으로는 Fedorov (1972)나 Atkinson과 Donev (1992)등을 참조하기 바란다.

제 2절에서는 O'Brien (1995)의 방법을 소개하고 다항회귀모형으로 연계하여 추가받힘점에 대한 논의를 심도 있게 한다. 기존 방법과 달리 이 방법이 가지고 특징들을 분석하고 이에 대한 사용 용도를 제안하고자 한다. 제 3절에서는 결론과 토의를 한다.

2. 받힘점 추가

먼저 O'Brien의 방법을 설명하여 보자. 다항회귀모형 $f_m(x)$의 모수의 개수는 $m+1$이므로 $D-$최적의 받힘점의 개수는 $m+1$이다. 또한 ξ_x을 새로운 받힘점 x에 모든 질량을 부여하는 한 점 (one point) 실험이라 하면 다음과 같은 실험설계는

$$\xi_E = \frac{m+1}{m+2}\xi_D + \frac{1}{m+2}\xi_x \tag{2.1}$$

$D-$최적의 받힘점 $x_1, \cdots, x_{m+1}$ 및 x에 질량 $1/(m+2)$를 배정하는 실험이 된다.

이러한 실험 ξ_E와 D-최적실험 ξ_D의 차이를 나타내주는 방법으로서 ξ_E이 $D-$최적에 대해 가지는 $D-$효율 $e[\xi_E(\xi_D)]$을 이용한다.

$$e[\xi_E(\xi_D)] = \left(|M(\xi_E)| \ / \ |M(\xi_D)|\right)^{1/(m+1)}$$

이러한 $D-$효율은 ξ_E에 대해서는 다음과 같이 표현된다. 앞으로는 이런 $D-$효율을 de로 표기한다.

$$de = \frac{m+1}{m+2}(1 + d(x, \xi_D)/(m+1))^{1/m+1} \tag{2.2}$$

O'Brien (1995)은 이러한 효율을 일정 수준, 예를 들어 90%를 충족시키는 받힘점 x을 추가받힘점으로 설정하는 것을 제안하였다. 식 (2.2)를 $d(x, \xi_D)$에 대해 정리하면 O'Brien이 추천한 받힘점이 나온다.

$$d(x, \xi_D) = (m+1)\left[\left(\frac{m+2}{m+1}de\right)^{m+1} - 1\right] \tag{2.3}$$

즉 예측분산의 값이 식 (2.3)의 우변과 일치하는 모든 받힘점 x를 찾는 문제가 된다. 다만 O'Brien은 비선형모형을 이용하여 이런 방법을 제시하였지만 이에 대한 응용은 선형모형이 더 적합하다고 판단되어 본 연구에서는 다항회귀모형을 가정하여 추가받힘점을 카탈로그 (catalogue)화하려는 것이다. 또한 O'Brien이 간과한 de의 설정문제도 짚으려고 한다.

예제 2.1 O'Brien은 아무런 설명 없이 de값으로 0.9와 같은 임의의 숫자를 제시하였는데 그러나 de를 설정할 때는 고려할 사항이 존재한다. 왜냐하면 단순회귀모형인 경우 $D-$최적 하에서의 예측분산함수는 $d(x, \xi_D) = 1 + x^2$으로 표시되는데 이는 실험영역 $\Omega = [-1, 1]$에서 최대값 2와 최소값 1로 상한과 하한이 주어져 있기 때문이다. 식 (2.3)의 우변에 있는 식을 이 상하한선에 맞추어 계산된다면 0.81650=$\sqrt{2/3} \le de \le \sqrt{8/9}$=0.94281사이에서 de가 설정되어야 한다. 만약 de가 하한 $\sqrt{2/3}$으로 설정되면 중간 받힘점이 0으로 설정이 되고 상한 $\sqrt{8/9}$로 설정을 하면 기존 모형에 대한 효율을 100%로 요구하는 문제가 된다. 후자의 경우는 중간 받힘점의 필요성은 없어진다.

본 연구에서는 편의상 de를 $\sqrt{2/3}$과 O'Brien이 제안한 0.9만 다루고자 한다. de가 0.9로 설정되는 경우는 받힘점은 -1, 1 이외에 ±0.803에 설정된다. 최적실험에서의 양극점인 받힘점에 추가하여 실험영역내의 점 두 개가 생성된 것이다. Figure 2.1은 $d(x, \xi_D)$의 그림인데 식 (2.3)에서 우측 값에 해당되는 값을 수직으로 읽어 들여 x축에서 교점을 읽어 드리면 된다.

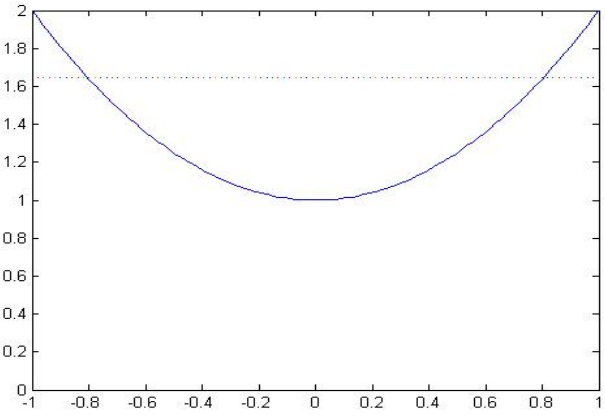

Figure 2.1 $d(x, \xi_D) = 1 + x^2$ with dotted line corresponding to de=0.9

기존 ξ_D의 받힘점과 O'Brien이 제안한 추가받힘점에 등질량을 부여하는 다음과 같은 실험의 효율성에 대해 알아보고자 한다.

$$\xi_O : \xi(\pm 1) = \xi(\pm x) = 1/4$$

참고로 본 연구에서는 받힘점에 대한 질량을 등질량 (equi-weight)으로 국한하여 설명을 전개할 것이나 반드시 그럴 필요는 없다. 기존의 $D-$최적실험을 ξ_D 그리고 추가받힘점만을 가정한 $D-$최적실험을

ξ_Δ라 한다면 다음과 같은 복합실험을 설정하여 볼 수 있다.

$$\xi_c = \alpha\xi_D + (1-\alpha)\xi_\Delta,\ 0 \le \alpha \le 1$$

그러나 본 연구의 특징이 O'Brien의 방법에 대한 다항회귀모형에 대한 적용성 문제를 검토하는 내용이므로 본 연구에서는 받힘점 개수만큼 해당하는 등질량의 실험만을 고려할 것이다. 왜냐하면 de의 상한값과 적절한 α를 이용하면 얼마든지 기본모형에 대한 D-효율을 올릴 수 있어 이러한 경우는 제외한다.

기본적으로 O'Brien이 제안한 실험은 가정된 모형에 대한 D-효율을 어느 정도 보장하는 조건을 만족하는 조건만 있을 뿐 다른 조건은 없다. 따라서 O'Brien (1995)의 방법은 기존모형에서 이탈되는 일반적인 모든 상황을 염두에 둔 실험이라 할 수 있고 그런 상황에서 사용을 추천하였다. 그러나 다항회귀모형인 경우는 참의 모형이 2차나 3차 다항회귀모형인 경우로 밝혀지는 경우가 있기 때문에 O'Brien이 제안한 실험의 $D-$효율을 이러한 관점에서 논의하는 것이 바람직하다.

$x = 0$인 경우 O'Brien의 실험은 2차 회귀모형을 가정한 $D-$최적실험, $\xi(\pm 1) = \xi(0) = 1/3$이므로 참의 모형이 2차 모형인 경우는 $e[\xi_O^2(\xi_D^2)]$ =1.0, 참의 모형이 가정한 모형 단순회귀모형인 경우는 $e[\xi_O(\xi_D^1)]$ =0.8165이 된다. ξ_O은 O'Brien이 제시한 실험이다. 그리고 ξ_D^1, ξ_D^2는 각각 1차 및 2차 다항회귀모형에 대한 $D-$최적이다.

그리고 $de = 0.9$으로 설정된 $x = \pm 0.803$인 경우는 $e[\xi_O(\xi_D^1)]$ =0.9069, $e[\xi_O^2(\xi_D^2)]$ =0.5593로 나타난다. 또한 $de = 0.9$인 경우는 ξ_O는 받힘점이 4개이므로 ξ_D^3과 비교하면 $e[\xi_O(\xi_D^3)]$=0.5947로 나온다. 결론적으로 O'Brien이 제안하였듯이 de값을 너무 높게 설정되고 참의 모형이 기존모형에서 이탈되면 그 효율은 급속히 떨어진다. 만약 de를 0.85로 낮추게 되면 $e[\xi_O(\xi_D^1)]$ =0.7909, $e[\xi_O^2(\xi_D^2)]$ =0.8396, $e[\xi_O^2(\xi_D^3)]$ =0.9909이므로 모든 이탈되는 모형에 대해 만족스러운 결과를 가져다준다.

참고로 3차 다항회귀모형에 대한 최적실험 $\xi_D^3 : \xi(\pm 1) = \xi(\pm\sqrt{1/5}) = 0.25$은 1, 2 차 모형에 대해서 가지는 효율은 다음과 같다.

$$e[\xi_D^3(\xi_D^3)] = 1.0,\ e[\xi_D^3(\xi_D^2)] = 0.865,\ e[\xi_D^3(\xi_D^1)] = 0.7736$$

결론적으로 실험자에게는 다음과 같은 3가지 실험 선택이 주어진다.

$$\xi_D^2,\ \xi_D^3,\ \xi_O^{de} : \sqrt{2/3} \le de \le \sqrt{8/9}$$

수치적으로 $de = \sqrt{2}/3$이면 $\xi_D^2 = \xi_O^{de}$가 되고 $de = 0.84327$이면 ξ_D^1에 받힘점 $\pm\sqrt{1/5}$이 추가되는 ξ_O실험이 된다. $de = 0.84327$를 기준으로 밑으로 de값을 잡으면 중간 받힘점은 $\pm\sqrt{1/5}$에서 이탈하여 0으로 향하는 반면 de값을 그 위로 잡으면 중간받힘점은 1로 향한다. de가 낮아지면 즉 기존 모형에 대한 효율을 낮게 잡더라도 다른 참의 모형이 될 가능성이 높은 2차나 3차에 대한 효율은 상대적으로 높아진다는 점을 생각하면 O'Brien이 제시한 $de = 0.9$의 경우는 다항회귀모형인 경우는 너무 높은 감이 없지 않다. 위에 언급한 3가지 실험 중 어느 실험을 선택할지는 실험자의 몫이 되겠지만 O'Brien이 제안한 방법을 선택할 시에는 조심스러운 de에 대한 설정이 필요하다. 또한 위의 세 가지 실험에 추가하여 다음과 같은 실험기준을 고려하여 보자.

$$\max_{\xi} \min \left\{ e[\xi(\xi_D^1)],\ e[\xi(\xi_D^2)],\ e[\xi(\xi_D^3)] \right\}$$

물론 이 경우는 O'Brien 의 방법과 달리 3차 항을 모형 집합에 포함시켜야 하는 조건이 필요하다. 비교 목적으로 받힘점에 등질량을 가정한다면 다음과 같은 실험이 나온다.

$$\xi_{mm} : \xi(\pm 1) = \xi(\pm 0.7739) = \xi(0) = 0.2$$

각 모형에 대한 $D-$효율의 각각 약 $e[\xi_{mm}(\xi_D^1)]$=0.7999, $e[\xi_{mm}(\xi_D^2)]$=0.8339, $e[\xi_{mm}(\xi_D^3)]$=0.7999로 나온다.

또한 O'Brien이 명기한 단순회귀모형에 대한 de를 0.8로 한 경우 다른 모형에 대한 효율에 어떤 변화가 일어나는지 다음과 같은 제약조건 실험을 실시하여 보았다. 이 방법은 Kim과 Lim (2007) 의 논문을 참조하기 바란다.

$$\max_{\xi} \min \left\{ e[\xi(\xi_D^2)],\ e[\xi(\xi_D^3)] \right\} \text{subject to } e[\xi(\xi_D^1)] \geq 0.80$$

이 경우도 비교 목적으로 받힘점에 등질량을 가정한다면 다음과 같은 실험이 나온다.

$$\xi_{mmc} : \xi(\pm 1) = \xi(\pm 0.7746) = \xi(0) = 0.2,$$
$$e[\xi_{mmc}(\xi_D^1)] = 0.80, e[\xi_{mmc}(\xi_D^3)] = 0.7969,\ e[\xi_{mmc}(\xi_D^2)] = 0.8343$$

이러한 경우 제약건식에 들어가는 하한 값은 O'Brien의 방법과 달리 실험자가 0과 1상에서 임의로 설정할 수 있다. 우연치 않게 ξ_{mm}과 ξ_{mmc}는 거의 유사한 결과를 보여 준다.

ξ_D^2, ξ_D^3, ξ_O^{de}: $\sqrt{2/3} \leq de \leq \sqrt{8/9}$에 추가하여 ξ_{mm}, ξ_{mmc}까지 실험의 선택의 폭은 넓어지는 것이다. 물론 이외에도 다양한 복합실험기준이 존재하나 여기서는 그 설명은 생략한다.

O'Brien의 방법이 de를 설정하는데 있어 다소 까다로운 점이 다항회귀모형인 경우는 존재하나 장점은 다른 방법과 달리 기본 모형에서 이탈 되는 구조식을 밝히지 않더라도 된다는 점이다. 기본 모형에 대한 효율만 너무 과다하게 설정하지 않는다면 기존의 문헌의 ξ_{mm}이나 ξ_{mmc}에 버금가는 실험을 손쉽게 설정할 수 있다.

또한 O'Brien의 방법은 기존의 받힘점에 추가하여 받힘점을 구하는 문제이므로 기존의 받힘점은 포기할 필요가 없다. 기존의 받힘점에 추가하여 추가받힘점이 생성된다는 점 역시 O'Brien방법의 매력이다.

예제 2.2 이차회귀모형에서 de가 정해지면 추가받힘점은 -1, 0, 1 이외에 $\pm x_1, \pm x_2$에 설정되는데 x_1, x_2값은 Table 2.1에 주어져 있다. 2차함수인 경우 $D-$최적의 예측분산함수는 $d(x, \xi_D) = 3 - 4.5x^2 + 4.5x^4$이고 실험영역 $\Omega = [-1, 1]$에서 받힘점 $-1, 0, 1$에서 분산의 최대값 3을 가지며 $\pm\sqrt{1/2}$=0.707에서 최저값, 1.875를 가진다. 따라서 예제 2.1과 마찬가지로 선택할 수 있는 de의 구간을 설정하여 보면 $0.88175 \leq de \leq 0.9450$이 나오는데 본 연구에서는 예제 2.1과 같은 맥락에서 de값을 임의로 0.88175, 0.9로 설정하였다. Figure 2.2를 참조하기 바란다.

Table 2.1 Extra support point for specific de

de	0.88175	0.90
추가받힘점	$\pm\sqrt{1/2}$= $\pm$0.7071	$\pm$0.9098, $\pm$0.4149

de가 0.88175인 경우 즉 $\xi(\pm 1) = \xi(0) = \xi(\pm\sqrt{1/2}) = 0.2$와 같은 실험인 경우 받힘점이 5개 까지 나오므로 차수가 1, 2, 3, 4인 다항 회귀모형에 대해 그 효율을 비교하여 볼 수 있을 것이다.

$$de = 0.88175 : e[\xi_O(\xi_D^1)] = 0.7746,\ e[\xi_O(\xi_D^2)] = 0.8277,\ e[\xi_O(\xi_D^3)] = 0.8599,\ e[\xi_O(\xi_D^4)] = 0.95979$$

참고로 ξ_D^4실험이 차수가 1, 2, 3, 4인 다항 회귀모형에 대해 가지는 효율은 다음과 같다.

$$e[\xi_D^4(\xi_D^1)] = 0.756,\ e[\xi_D^4(\xi_D^2)] = 0.7828,\ e[\xi_D^4(\xi_D^3)] = 0.895,\ e[\xi_D^4(\xi_D^4)] = 1.0$$

다음으로 de값을 0.9로 상향 조정하여 보자. 받힘점이 7개가 나온다. Figure 2.2를 참조하기 바란다. 기존 받힘점 3개에 4개의 추가받힘점이 생성된다.

$$\xi(\pm 1) = \xi(\pm 0.8730) = \xi(\pm 0.4878) = \xi(0) = 1/7$$

차수가 1, 2, 3, 4, 5, 6인 다항회귀모형 $D-$최적에 대해 가지는 효율은 다음과 같다.

$$de = 0.9 : e[\xi_O(\xi_D^1)] = 0.7559,\ e[\xi_O(\xi_D^2)] = 0.8168,\ e[\xi_O(\xi_D^3)] = 0.8494,$$
$$e[\xi_O(\xi_D^4)] = 0.8618,\ e[\xi_O(\xi_D^5)] = 0.8806,\ e[\xi_O(\xi_D^6)] = 0.9743$$

참고로 ξ_D^6 실험이 차수가 1, 2 3 4, 5, 6인 다항 회귀모형에 대해 가지는 효율은 다음과 같다.

$$e[\xi_D^6(\xi_D^1)] = 0.7385,\ e[\xi_D^6(\xi_D^2)] = 0.7975,\ e[\xi_D^6(\xi_D^3)] = 0.8464$$
$$e[\xi_D^6(\xi_D^4)] = 0.8676,\ e[\xi_D^6(\xi_D^5)] = 0.9026,\ e[\xi_D^6(\xi_D^6)] = 1.0$$

실험의 선택은 보호하고자 하는 다항회귀모형의 차수에 달려 있다. 기본 모형인 2차모형에서 이탈되는 정도가 심하지 않고 2차 모형에 대한 효율을 높게 유지한다면 de를 최소로 하는 O'Brien의 방법을, 그렇지 않고 참의 모형이 어떤 차수의 다항회귀모형인지 모를 경우에는 4차를 선택하는 것이 합리적이다. 혹은 이탈의 정도가 심하면 5, 6차 항의 다항회귀모형을 기준으로 하여 받힘점을 선택하는 것이 합리적이다. 그러나 차수가 높은 다항회귀모형을 염두에 둔 최적실험은 결과적으로 받힘점의 개수가 늘어난다. 기존 3개의 받힘점에서 2배 이상의 6, 7개의 받힘점의 개수는 다소 과다하다고 판단된다. 따라서 기존의 받힘점 3개에 신규 받힘점 두 개가 추가되는 최소 de를 가정한 O'Brien의 방법도 나쁜 선택은 아니라 보여 진다. 또한 이탈되는 모형의 구조식을 실험에 포함시키지 않아도 되는 장점이 있기 때문이다. ξ_O는 간단한 계산으로 결정되므로 ξ_{mm}, ξ_{mmc}와 같이 알고리즘에 의해 계산되는 실험보다 간편한 점이 이점이 될 수 있을 것이다. 물론 불확실한 모형들의 집합이 구조적으로 구성되어 있다면 Kim과 Lim (2007)의 방법이 추천된다.

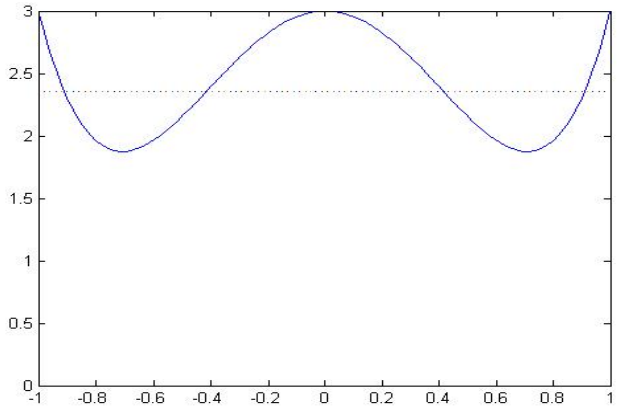

Figure 2.2 $d(x, \xi_D) = 3 - 4.5x^2 + 4.5x^4$ with dotted line corresponding to de=0.9

3. 결론

본 연구에서는 다항회귀모형을 기준으로 추가받힘점의 생성과정을 알아보았다. 최적실험이 가지고 있는 단점을 보완하면서 이론적인 근거에 의해 모델에 대한 불확실성을 감안한 실험을 실시할 수 있는

바탕이 마련된 것이다. O'Brien방법은 기존 모형에서 일반적인 이탈 (general departure)을 염두에 둔 방법이므로 기존문헌에서 추천되었던 확장 모형을 명시하지 않아도 되는 장점이 있다. O'Brien은 비선형모형을 염두에 두고 이러한 방법을 제안하였으나 비선형모형에 대한 실험은 모수 값에 의존하는 국지 $D-$최적이 결정되고 de를 설정하는데 있어 애매모호한 점이 있어 본 연구에서는 다항회귀모형을 염두에 둔 실질적인 실험을 구성하여 보았다.

References

Atkinson, A. C. and Donev, A. N. (1992). *Optimum experimental designs*, Oxford Press, New York.

Cook, R. D. and Nachtsheim, C. J. (1980). Model robust, linear-optimal designs. *Technometrics*, **24**, 49-54.

de la Garza, A. (1954). Spacing of information in polynomial regression. *Annals of Mathematical Statistics*, **25**, 123-130.

Dykstra, O. (1971). The augmentation of experimental data to maximize $|X^T X|$. *Technometrics*, **13**, 682-688.

Evans, J. W. (1979). Computer augmentation of experimental designs to maximize $|X^T X|$. *Technometrics*, **21**, 321-330.

Fedorov, V. V. (1972). *Theory of optimal experiments*, Academic Press, New York

Guest, P. G. (1958). The spacing of observations in polynomial regression. *Annals of Mathematical Statistics*, **29**, 294-299.

Kim, M. S., Hong, K. J. and Choi, D. H. (2002). Augmented D-optimal design for effective response surface modelling and optimization. *Journal of Mechanical Science and Technology*, **16**, 203-210.

Kim, Y. I. and Lim, Y. B. (2007). Hybrid approach when multiple objectives exist. *The Korean Communications in Statistics*, **14**, 531-540.

Kussmaul, K. (1969). Protection against assuming the wrong degree in polynomial regression. *Technometrics*, **11**, 677-682.

O'Brien, T. E. (1995). Optimal design and lack of fit for nonlinear regression models. In *Statistical Modelling*, edited by G. U. H. Seeber B. J Francis, R. Hatzinger and G. Steckel-Berger, Springer, New York, 201-206.

Journal of the Korean Data & Information Science Society
2014, **25**(6), 1491–1498

http://dx.doi.org/10.7465/jkdi.2014.25.6.1491
한국데이터정보과학회지

Selection of extra support points for polynomial regression

Young-Il Kim[1] · Dae-Heung Jang[2]

[1]School of Business and Economics, Chung-Ang University
[2]Department of Statistics, Pukyong National University

Received 11 October 2014, revised 30 October 2014, accepted 12 November 2014

Abstract

The major criticism of optimal experimental design is that it depends heavily on the model and its accompanying assumption that often leads the number of support points equal to the number of parameters in the model. Often in the past, a polynomial model of higher degree is assumed to handle the experimental design for the polynomial regression of lower degree. In this paper we searched the possible set of designs which are robust to the departure of the assumed model. The designs are categorized with respect to D-efficiency. The approach by O'Brien (1995) was discussed in univariate polynomial regression model setting.

Keywords: $D-$efficiency, D-optimal experimental design, polynomial regression, support points.

[1] Corresponding author: Professor, School of Business and Economics, Chung-Ang University, Seoul 156-756, Korea. E-mail: yik01@cau.ac.kr
[2] Professor, Department of Statistics, College of Natural Science, Pukyong National University, Pusan 608-737, Korea.