The Korean Journal of Applied Statistics (2015) DOI: http://dx.doi.org/10.5351/KJAS.2015.28.3.561
28(3), 561–571

Robust Extrapolation Design Criteria under the Uncertainty of Model and Error Structure

Dae-Heung Jang[a] · Youngil Kim[b,1]

[a]Department of Statistics, Pukyong National University
[b]School of Business and Economics, ChungAng University

(Received April 22, 2015; Revised May 24, 2015; Accepted May 25, 2015)

Abstract

When we consider an optimal design to predict the response corresponding to the point outside the design region, we are extremely careful about choosing the design criteria for selecting the support points. The assumed model and its accompanying error structure should be assumed to extend beyond the design region for the selected design criteria to be valid. Thus, we modify the existing design criteria such as extrapolation-optimality to be suited to those situations. We propose some maximin approaches in this paper. Simple and quadratic regression models are tested to find the basic characteristics of such maximin approaches. Some main findings are discussed in the conclusion.

Keywords: extrapolation, extrapolation-optimality, maximin approach

1. 서론

R^k 공간의 어떤 조밀한 (compact)한 부분집합 Ω에 정의된 m개의 독립적인 연속회귀함수(continuous regression function)로 구성된 $f^T(x) = (f_1(x), f_2(x), \ldots, f_m(x))$가 있다고 가정하고 Ω 안에 있는 모든 x에 대하여 단변량(univariate) 반응변수 $y(x)$는 식 (1.1)과 같은 통계적인 선형모형에 의거 관측이 된다고 가정하자.

$$y(x) = f^T(x)\theta + \frac{e}{\sqrt{\lambda(x)}} = \sum_i^m f_i(x)\theta_i + \frac{e}{\sqrt{\lambda(x)}}, \tag{1.1}$$

여기서 $\theta^T = (\theta_1, \theta_2, \ldots, \theta_m)$은 추정하여야 할 모수벡터이고 e는 서로 비상관인 실수 값을 가지는, 기댓값이 0이고 분산이 σ^2인 오차확률변수이다. 앞으로는 일반성의 손실 없이 σ^2은 1로 가정한다. 여기서 $\lambda(x)$는 Ω에 정의된 형태가 알려진 양의 실수 값을 가지는 유계(bounded) 함수라 가정한다. 이러한 $\lambda(x)$는 효율함수(efficiency function)라 한다. 만약 $\lambda(x)$가 상수함수라면 오차들은 등분산(constant variance) 구조를 가진다고 하고, 그렇지 않으면 이분산(non-constant) 구조를 가진다고 한다.

실험계획문제는 Ω 안에 있는 s개의 점, $x_i = 1, 2, \ldots, s$로 구성된 유한집합에 질량을 부여하는 확률 질량함수 ξ로 기술된다. 앞으로는 점 x_i에 배정되는 질량은 $\xi(x_i)$로 표기한다.

[1]Corresponding author: School of Business and Economics, ChungAng University, 84 Heukseok-ro, Dongjak-gu, Seoul 156-756, Korea. E-mail: yik01@cau.ac.kr

전체 실험에 쓰이는 관측 값의 개수, n을 곱한 $n\xi(x_i)$가 정수라는 제약조건이 주어지면 실험계획은 정확실험(exact design), 제약조건이 없는 경우는 근사실험(approximate design)이라 한다. 본 연구에서는 후자를 택하는데 이는 이론적인 최적실험분야의 연구에서는 이 방법을 통하여 실험설계의 근본적인 성질을 연구할 수 있기 때문이다.

식 (1.1) 하에서는 실험설계 ξ의 정보행렬 $M(\xi)$은 식 (1.2)과 같이 나타난다.

$$M(\xi) = \int_{\Omega} \lambda(x) f(x) f^T(x) d\xi(x). \tag{1.2}$$

$M(\xi)$로 실험설계 ξ의 정보의 양을 측정하는데 비특이(non-singular) 행렬로 가정한다. 그리고 이를 이용하여 특정한 점 z에서의 예측분산(prediction variance), $d(z,\xi)$은 식 (1.3)과 같다. 여기서 z는 실험영역내의 점일 필요는 없다.

$$d(z,\xi) = f^T(z) M^{-1}(\xi) f(z). \tag{1.3}$$

최적실험설계 분야에서는 많은 실험기준이 존재하는데 대표적으로 $M(\xi)$의 행렬식을 최대화시키는 D-최적 실험기준과, 예측분산, $d(z,\xi), z \in \Omega$의 최대값을 최소화하는 G-최적 실험기준 등을 들 수 있다. $\lambda(x)$가 상수함수인 경우 Kiefer와 Wolfowitz (1960)에 의하면 이러한 D-최적과 G-최적 사이에는 동격성(equivalence)이 존재한다. 그리고 최적성(optimality)은 $d(z,\xi), z \in \Omega$의 최대값이 모형의 모수의 개수인 m과 일치하는지 여부를 통하여 확인할 수 있다.

본 연구에서 다루는 주된 실험기준은 실험영역을 벗어나는 점, $z_0 \notin \Omega$에 해당되는 식 (1.4)의 예측분산을 최소화하는 외삽(extrapolation)-최적이다. 앞으로는 이러한 실험기준을 편의상 ext-최적으로도 병기한다.

$$d(z_0,\xi) = f^T(z_0) M^{-1}(\xi) f(z_0), \quad z_0 \notin \Omega. \tag{1.4}$$

z_0가 Ω 내의 한 점이고 식 (1.3)의 분산을 최소화 하는 실험 즉, $f^T(z_0)\theta$를 가능하면 정확하게 추정하기 위한 실험은 z_0에 모든 질량을 부여하는 실험기준이 최적이 되기 때문에 본 연구에서는 이를 제외하고 실험영역 바깥에 있는 점만을 대상으로 한다. z_0가 실험영역 바깥에 있는 경우는 z_0에 질량을 부여하지 못하고 실험영역 내의 받힘점(support points)에 질량을 부여하여야 하는 실험계획이 된다. 앞으로는 $z_0 \notin \Omega$와 같은 점을 외삽점으로 부른다. 흥미롭게도 많은 경우 받힘점의 위치는 외삽점의 위치에 상관없이 정해지며 외삽최적의 문제는 받힘점에 배정되는 질량에 의해서만 결정된다.

그러나 외삽최적의 해는 Berger와 Wong (2009, 242p)이 지적하였듯이 실험영역을 벗어나도 가정된 모형이 여전히 타당하여야 한다는 전제조건이 있어야 한다. 그러나 $z_0 \notin \Omega$가 실험영역을 많이 벗어날수록 이러한 가정은 불안전하다. 이러한 점에 유의하여 Dette와 Wong (1996)이 Läuter (1974)의 실험기준을 이용하여 외삽점을 대상으로 모형의 불확실성이 존재할 경우 최적실험을 제안하였다. 그러나 이들의 논문들은 모형의 불확실성에 따르는 $z_0 \notin \Omega$에서의 예측분산의 선형결합만을 고려한 실험기준을 고려하였을 뿐 모형추정을 위한 실험기준은 고려하지 않아, Kim과 Jang (2012)은 모형추정문제까지 포함한 복합실험기준을 다루었다.

외삽문제에서는 이처럼 모형의 타당성도 문제가 되지만 한편으로는 실험영역을 벗어나게 되면 오차의 구조 역시 달라질 가능성이 많기 때문에 이에 대한 논의가 필요하다. 기존문헌에서는 주로 등분산을 가정한 모형을 다루었는데 이는 식 (1.1) 양변에 $\sqrt{\lambda(x)}$를 곱하면 모형이 $\sqrt{\lambda(x)}y(x) = \sqrt{\lambda(x)}f^T(x) + e$로 바꾸어지기 때문에 등분산의 오차구조 하에서의 D-최적의 해를 구할 수 있기 때문이다. 그러나 더 이상 오차항의 등분산 구조 하에서의 D-최적과 G-최적의 동격성은 존재하지 않는다. 따라서 $z_0 \in \Omega$인

경우 이분산하에서의 식 (1.3)을 최소화하는 논의가 상당부분 이루어져 있다. 대표적인 참고문헌으로 Wong과 Cook (1993)의 효시적인 논문을 들 수 있다. 그러나 일반적인 G-최적과 달리 이분산하에서의 외삽최적인 경우는 이분산의 구조가 특수한 경우가 아니면 해석적인 해를 구하기 어렵다. Chen 등 (2008)의 논문 등에서 일부 해석적인 해가 존재한다.

따라서 본 연구에서는 오차항의 구조가 이분산인 경우 외삽최적의 특징을 알아보고 더 나아가 오차항의 구조식이 불확실한 경우 다항회귀모형을 중심으로 maximin 실험기준을 알아본다. 또한 분산구조의 불확실한 문제 뿐 아니라 모형이 불확실한 경우도 포함시켜 외삽최적에 대한 문제를 알아보고자 한다.

본 연구에서 정의될 maximin 실험설계를 논하기 전에 임의의 실험설계 ξ가 최적실험에 대하여 가지는 효율을 먼저 정의할 필요가 있다.

D-**효율**: *D*-최적에 대한 실험설계 ξ의 *D*-효율 $\phi_\xi(\xi_D^*)$는 다음과 같이 정의된다.

$$\phi_\xi(\xi_D^*) = \left\{\frac{\det(M(\xi))}{\det(M(\xi_D^*))}\right\}^{\frac{1}{m}},$$

여기서 ξ_D^*는 D-최적 실험설계이며 det는 행렬식을 의미한다.

외삽효율: ext-최적에 대한 실험설계 ξ의 외삽효율 $\phi_\xi(\xi_{ext}^*)$는 다음과 같이 정의된다.

$$\phi_\xi(\xi_{ext}^*) = \frac{f^T(z_0)M^{-1}(\xi_{ext}^*)f(z_0)}{f^T(z_0)M^{-1}(\xi)f(z_0)}, \quad z_0 \notin \Omega,$$

여기서 ξ_{ext}^*는 ext-최적 실험설계를 의미하고 z_0는 실험영역 Ω 바깥에 위치하는 외삽점이다.

참고로 위와 같은 효율은 어느 실험기준, ψ이든지 정의가 될 수 있다. 간단히 표현하면 실험설계 ξ의 ψ-효율이란 실험기준 ψ에 대하여 실험설계 ξ가 최적 실험설계 ξ_ψ^*에 대하여 가지는 효율 $\phi_\xi(\xi_\psi^*)$이라 정의가 된다.

제 2절에서는 1절에서 정의된 개념을 이용하여 오차항이 이분산의 구조를 가지고 있는 경우 그 특성을 알아보고 이분산의 구조의 불확실한 상황에 대처한 maximin 방법을 제안코자 한다. 더 나아가 모형의 불확실한 경우도 고려한다. 제 3절에서는 제 2절에서 제안된 실험기준을 이용하여 다항회귀모형을 중심으로 그 예제를 살펴볼 것이다. 그리고 제 4절에서는 결론과 토의가 뒤따른다.

2. 이분산 구조하에서의 외삽문제

오차항의 이분산 구조를 위하여 하나의 편의적인 이분산 구조를 설정하여 본다. 분산은 받힘점의 값이 커질수록 Ω 안에서 선형으로 증가하고 예측분산의 최소값과 최대값의 비(ratio)는 상수 γ로 알려져 있다고 가정한다. 이는 수학적으로 $\lambda^{-1}(x) \propto (\gamma-1)x+(\gamma+1)$로 표현이 가능한데 $\gamma = 1$일 때 $\lambda(x) = 1$인 표준화된 상수함수가 되기 위해서 편의상 $\lambda^{-1}(x) = 1/2 \times [(\gamma-1)x + (\gamma+1)]$로 정한다.

그러나 실험자는 실험 전에는 γ의 값을 확신하지 못한다. 그렇다 하더라도 실험자는 일반적으로 γ의 값이 포함된 구간은 비교적 쉽게 설정할 수 있다. γ의 값의 범위를 가정할 수 있다면 즉, $\gamma \in \Gamma = [1, \gamma_0], \gamma_0 > 1$으로 범위를 설정하고 γ값이 실험 후에 추정된다면 다음과 같은 실험설계기준을 정의하여 볼 수 있다.

정의 2.1 실험계획 ξ^*는 아래 조건이 만족할 때 이분산 오차-강건(HETEROCEDASTIC ERROR-ROBUST) 외삽실험이라 한다.

$$\max_{\xi} \min_{\gamma\in\Gamma} \phi_\xi(\xi_{ext}^*) = \min_{\gamma\in\Gamma} \phi_{\xi^*}(\xi_{ext}^*). \tag{2.1}$$

정의 2.1은 γ의 값이 실험 후 어떻게 추정이 되더라도 실험설계 ξ가 Γ 내의 γ에 대하여 가지는 외삽-효율 중 최소값을 최대화시키는 실험설계이다.

이와 마찬가지로 γ가 주어진 상황이라면 가정한 모형 $f(x)$가 불확실한 경우도 maximin 실험이 가능하다. 즉, 모형에 대한 확신이 없는 경우라도 모형 $f(x)$을 k개의 모형으로 구성된 모형공간의 한 요소, 즉 $f(x) \in F = [f_1(x), f_2(x), \ldots, f_k(x)]$와 같이 정의하고 $f(x)$가 실험 후에 F의 한 모형으로 추정된다면 정의 2.1와 유사하게 정의 2.2가 가능하다.

정의 2.2 실험계획 ξ^*는 아래 조건이 만족할 때 이분산 모형-강건(HETEROCEDASTIC MODEL-ROBUST) 외삽실험이라 한다.

$$\max_{\xi} \min_{f \in F} \phi_{\xi}(\xi^*_{ext}) = \min_{f \in F} \phi_{\xi^*}(\xi^*_{ext}). \tag{2.2}$$

정의 2.1과 2.2를 융합하여 정의 2.3을 정의할 수 있다. 오차항의 분산의 구조뿐 아니라 모형에 대한 불확실성이 존재할 때 이를 동시에 다룰 수 있는 기준이 된다. 그러나 향후 예제를 통하여 정의 2.3에 의한 실험설계의 특징을 설명하겠지만 실제로는 이와 같이 복잡한 기준을 설정할 필요는 없어 보인다.

정의 2.3 실험계획 ξ^*는 아래 조건이 만족할 때 이분산 오차-모형 강건(HETEROSCEDASTIC ERROR-MODEL ROBUST) 외삽실험이라 한다.

$$\max_{\xi} \min_{\gamma \in \Gamma} \min_{f \in F} \phi_{\xi}(\xi^*_{ext}) = \min_{\gamma \in \Gamma} \min_{f \in F} \phi_{\xi^*}(\xi^*_{ext}). \tag{2.3}$$

이와 같이 정의된 실험기준에 따른 실험설계는 다음과 같은 알고리즘으로 실현가능하다. 알고리즘은 Fedorov (1972)의 전통적인 교환알고리즘을 변형하여 만든 것이다. 교환알고리즘은 매 단계별로 기존 받힘점과 질량을 목적함수가 수렴될 때까지 갱신하는 과정에서 생기는 새로운 받힘점과 질량을 교환하는 반복적인 알고리즘이다. 아래 주어지는 알고리즘 2.1은 정의 2.1을 기준으로 한다. 다른 정의도 유사하게 정의할 수 있다.

알고리즘 2.1 :

단계 2.1 정보행렬이 정칙 (non-singular)인 실험계획인 ξ_0을 설정한다 ($i = 0$).

단계 2.2 다음이 성립되는 c_i를 구한다.

$$\min_{\gamma \in \Gamma} \phi_{\xi}^{-1}(\xi^*_{ext}) = c_i.$$

단계 2.3 동시에 단계 2.2에서 나오는 최대값 c_i를 가져다주는 $x_i \in \Omega$를 찾는다.

단계 2.4 $\alpha_i = 1/(i+s)$, $s \geq 0$로 정한다.

단계 2.5 $\xi_{i+1} = (1-\alpha_i)\xi_i + \alpha_i \xi_{x_i}$로 새로운 실험을 구성한다. 여기서 ξ_{x_i}는 x_i에 질량전체를 부여하는 실험이다.

단계 2.6 두 개의 연속된 c_i의 차이가 충분히 작으면 알고리즘을 정지하고 그렇지 않으면 단계 2.2로 돌아간다.

참고로 단계 2.4에서의 s값은 알고리즘의 수렴(convergence) 속도를 조절하는 역할을 하는데 s의 크기를 정하는 명확한 규칙이 있는 것은 아니다. 그리고 수열(sequence) α_i는 c_i의 단조 감소(monotonically decreasing)를 유발하지는 않는다. 알고리즘의 핵심은 단계 2.3에서 $x_i \in \Omega$를 찾는 문제이다.

Table 3.1. Extrapolation design for each combination of z_0 and γ for simple linear regression

z_0	$\gamma = 1$	$\gamma = 2$	$\gamma = 3$
1.1	0.9545	0.9674	0.9732
2.0	0.7500	0.8093	0.8386
4.0	0.6250	0.7021	0.7427

3. 예제

2절에서 정의된 실험설계 기준을 다항회귀 모형을 중심으로 예제들을 통하여 그 의미를 알아본다. 모든 예제에서 $\Omega = [-1, 1]$을 가정한다.

예제 3.1: $f_1^T(x) = (1, x)$인 선형회귀모형과 2차형식의 회귀모형 $f_2^T(x) = (1, x, x^2)$을 중심으로 $\gamma = 1, 2, 3$인 경우 외삽-최적 실험설계의 구조를 먼저 살펴본 다음 2절에서 정의된 3가지 실험기준의 형태를 파악할 필요가 있다. Table 3.1이 $f_1(x)$를 대상으로 한 외삽-최적 실험설계이고 Table 3.2가 $f_2(x)$를 대상으로 한 외삽-최적 실험설계이다. Table 3.1에서 주어진 숫자는 $\xi(1)$의 값이다.

외삽점 z_0의 위치는 Dette와 Wong (1996)의 논문을 참조하여 편의적으로 세 값을 선정하였다. 외삽점 z_0의 위치가 1.1이란 의미는 실험영역 인근에 위치하였다는 의미이며 $z_0 = 4.0$의 의미는 외삽점의 위치가 실험영역을 아주 많이 벗어났다는 의미가 될 것이다. 먼저 $\gamma = 1$인 등분산의 경우는 Hoel과 Levine (1964)에 의하여 다항회귀모형 $f^T(x) = (1, x, x^2, \ldots, x^{m-1})$을 알고 있다는 가정 하에서 외삽-최적실험 $\xi^*_{z_0}, z_0 \notin \Omega$에 대한 해가 제공되어 있다. $z_0 = 1.1$인 경우 받힘점 1에 많은 질량 0.9545가 배정됨을 알 수 있는데 제 1절에서 언급되었지만 이는 z_0가 실험영역 안에 있는 내삽-최적의 문제인 경우는 해당되는 점 $z_0 \in \Omega$에 모든 질량이 배정된다는 사실을 아는 이상, z_0가 1에 가까우면 그리 짐작키 어려운 사항은 아니다. 그러나 z_0가 1에서 멀어질수록 받힘점 1에 배정되는 질량은 줄어든다. 그만큼 불확실성이 증대되어 오히려 받힘점 1에 배정되는 질량이 감소된다는 의미이다. 그리고 $\gamma = 2, \gamma = 3$인 경우 $\gamma = 1$에 비하여 $\xi(1)$의 값이 상대적으로 크다는 사실을 알 수 있다. 또한 $z_0 = 1.1$인 경우 $\gamma = 1$에 비하여 $\gamma = 3$인 경우 $\xi(1)$의 질량은 2% 늘어나는 반면 $z_0 = 4.0$인 경우는 $\xi(1)$의 질량이 18.8%로 증가한다. 그러나 $\xi(1)$의 값은 기본적으로 외삽의 문제로 질량이 늘어나는 것이지 γ의 값에 따라 추가적으로 아주 크게는 영향을 받지는 않는 것으로 보인다. 한편으로는 $\gamma = 3$을 기준으로 이분산 G-최적인 경우가 $\xi(1) = 0.75$임을 감안하면 외삽점의 위치에 따른 질량변화와 이분산에 의한 질량변화는 중복되어 나타나는 현상으로 이해된다.

Table 3.1에서 나타난 γ의 값이 크고 외삽점의 위치가 1에서 가까울수록 1에 배정된 질량 $\xi(1)$의 값이 커지는 현상은 Table 3.2에서 비슷하게 확인된다. 2차 회귀모형은 받힘점이 3개로 나타나는데 가운데 받힘점의 질량 $\xi(0)$과 1에서의 질량 $\xi(1)$의 값의 패턴을 보게 되면 이해가 될 것이다.

예제 3.2부터 예제 3.4에서는 정의 2.1부터 정의 2.3에서 정의된 실험설계의 형태를 살펴볼 것이다. 또한 $z_0 = 1.1$인 경우 $\gamma = 1$에 비하여 $\gamma = 3$인 경우 $\xi(1)$의 질량은 4.6% 늘어나는 반면 $z_0 = 4.0$인 경우는 $\xi(1)$의 질량이 20.6%로 증가한다. 상대적으로 $\xi(1)$의 증가폭이 단순회귀모형보다는 높으나 패턴은 비슷하다.

예제 3.2: Table 3.3은 $f^T(x) = (1, x)$를 염두에 두고 정의 2.1에 정의된 실험기준을 $\gamma \in \Gamma = [1, 3]$으로 설정하여 세 외삽점, $z_0 = 1.1, z_0 = 2, z_0 = 4$ 각각의 경우 구한 실험설계이다. 오차-강건 외삽실험 역시 외삽점이 1에서 멀어질수록 받힘점 1에서의 질량 $\xi(1)$은 상대적으로 작아진다.

Table 3.2. Extrapolation design for each combination of z_0 and γ for quadratic linear regression

z_0	$\gamma = 1$	$\gamma = 2$	$\gamma = 3$
1.1	$\xi(-1) = 0.0387$	0.0283	0.0234
	$\xi(0) = 0.1479$	0.1322	0.1262
	$\xi(1) = 0.8134$	0.8395	0.8504
2.0	$\xi(-1) = 0.1429$	0.1121	0.0958
	$\xi(0) = 0.4286$	0.4121	0.4064
	$\xi(1) = 0.4286$	0.4758	0.4977
4.0	$\xi(-1) = 0.1935$	0.1558	0.1347
	$\xi(0) = 0.4839$	0.4770	0.4763
	$\xi(1) = 0.3226$	0.3672	0.3889

Table 3.3. Error-robust extrapolation design for simple linear regression

z_0	$\xi(-1)$	$\xi(1)$
1.1	0.0361	0.9639
2.0	0.2057	0.7943
4.0	0.3162	0.6839

Table 3.4. Extrapolation-efficiency of error-robust extrapolation design for each combination of z_0 and γ for simple linear regression

z_0	$\gamma = 1$	$\gamma = 2$	$\gamma = 3$
1.1	0.9975	0.9996	0.9975
2.0	0.9881	0.9986	0.9881
4.0	0.9842	0.9985	0.9842

Table 3.5. Error-robust extrapolation design for quadratic linear regression

z_0	$\xi(-1)$	$\xi(0)$	$\xi(1)$
1.1	0.0313	0.1365	0.8322
2.0	0.1211	0.4129	0.4661
4.0	0.1649	0.4780	0.3571

Table 3.4에 표기된 각 γ에 대한 외삽효율을 살펴보면 오차-강건 외삽실험은 매우 높은 효율을 유지함을 알 수 있다. 특히, $z_0 = 1.1$인 경우 어떤 γ가 추후 밝혀지더라도 모두 99% 이상의 효율을 가지고 있다. 그리고 모든 z_0에 대해서 $\gamma = 1, \gamma = 3$에서의 효율이 같다는 사실은 γ의 값을 최소값과 최대값만 지정하면 된다는 사실을 의미한다. 이를 이용하여 해석적인 해를 구할 수 있는 여지가 나오는데 Dette와 Biedermann (2003) 등이 비선형 모형인 Michaelis-Menten 모형에서 모수의 불확실성을 염두에 둔 maximin 방법을 연구할 때 밝힌 바 있다. Kim과 Jang (2014)의 최근 논문에도 이를 확장하여 밝힌 바 있다.

이분산의 구조가 다양한 형태로 표시되는 상황 하에서는 굳이 특정 이분산 구조에 대한 해석적인 해를 구할 필요는 없다고 판단되나 단순회귀모형인 경우 그 해석적인 해가 비교적 쉽게 밝혀져 부록 A에 명시하였다.

Table 3.5는 $f^T(x) = (1, x, x^2)$를 염두에 둔 정의 2.1에 정의된 실험기준을 $\gamma \in \Gamma = [1, 3]$으로 설정하여 3가지 외삽점 각각의 경우 구한 실험설계이다.

Table 3.6에 표기된 각 γ에 대한 외삽효율을 살펴보면 오차-강건 외삽실험은 Table 3.4와 마찬가지로 매우 높은 효율을 유지하며 양극단에서의 효율은 일치함을 보인다.

Table 3.6. Extrapolation-efficiency of error-robust extrapolation design for each combination of z_0 and γ for quadratic linear regression

z_0	$\gamma = 1$	$\gamma = 2$	$\gamma = 3$
1.1	0.9969	0.9995	0.9969
2.0	0.9925	0.9991	0.9925
4.0	0.9917	0.9992	0.9917

Table 3.7. Heterocedastic model-robust extrapolation design and its extrapolation efficiency for each model in model space

γ	z_0	$\xi(-1)$	$\xi(0)$	$\xi(1)$	efficiency
	1.1	0.0410	0.0739	0.8851	0.9260
1	2.0	0.1968	0.2143	0.5889	0.7857
	4.0	0.2709	0.2418	0.4873	0.7575
	1.1	0.0336	0.0667	0.8998	0.9352
2	2.0	0.1586	0.2058	0.6356	0.7934
	4.0	0.2040	0.2382	0.5578	0.7600
	1.1	0.0296	0.0644	0.9060	0.9399
3	2.0	0.1173	0.2052	0.6775	0.8008
	4.0	0.2100	0.2435	0.5465	0.7722

Table 3.8. Heterocedastic error-model robust extrapolation design for each extrapolation point, $z_0 = 1.1, 2, 4$

z_0	$\xi(-1)$	$\xi(0)$	$\xi(1)$
1.1	0.0418	0.0739	0.8843
2.0	0.1995	0.2143	0.5862
4.0	0.2799	0.2419	0.4782

예제 3.3: 정의 2.2에서 정의한 실험기준으로 모형 공간 F에 들어가는 f를 $f_1^T(x) = (1, x)$, $f_2^T(x) = (1, x, x^2)$로 설정하여 즉, $f \in F = [f_1(x), f_2(x)]$로 하고 $\gamma = 1, 2, 3$에 대하여 강건실험을 구성하여 보면 각각 Table 3.7과 같다. Table 3.7의 마지막 열은 efficiency로서 이는 모형 f_1 혹은 f_2가 참의 모형일 때 발생하는 외삽효율인데 모형공간의 원소가 2개이므로 이는 Imhof와 Wong (2000)의 전개에 의하면 두 외삽효율은 같다. 주어진 γ 에서의 외삽효율의 패턴은 어떤 γ라 하더라도 비슷하게 나온다. 이는 기본적으로 γ가 주어져 있다면 외삽점의 위치에 상관없이 이분산 모형-강건 실험계획은 안전성을 유지한다는 의미이다. 그러나 외삽점이 1에서 멀어 질수록 외삽효율은 떨어지는 현상이 나타난다. Table 3.1과 Table 3.2와 비교하여 보면 된다. f_1과 f_2의 혼합모형에 대한 실험기준으로 해석이 가능하다.

예제 3.4: 여기서는 정의 2.3에서 정의된 실험기준을 살펴보도록 하자. 이는 $\gamma \in \Gamma = [1, 3]$과 $m \in M = [f_1, f_2]$를 동시에 고려한 실험이다. Table 3.8이 정의 2.3에 의한 실험설계이다. 참고로 예제 3.3과 마찬가지로 $f_1^T(x) = (1, x)$, $f_2^T(x) = (1, x, x^2)$이다.

Table 3.8의 첫 번째 행의 실험은 Table 3.7의 $\gamma = 1$에 해당하는 $z_0 = 1.1$에 관련된 실험과 유사함을 알 수 있다. 다른 행도 정확하게 일치하지는 않지만 마찬가지의 유형을 보이는데 이는 $\gamma = 1$을 대상으로 한 모형-강건(heterocedastic model-robust) 외삽실험이 정의 2.3에서 정의한 실험이 된다는 의미이다. 즉, 모형과 이분산을 동시에 고려할 때는 오차의 이분산의 크기를 나타내는 γ의 값은 그리 신경을 쓰지 않고 모형의 불확실성에 더 주안점을 두어야 한다는 의미이다. 다시 말하면 외삽과 이분산은 동시에 받힘점 1에 더 많은 질량을 부여하기 때문에 모형까지 고려하는 경우에는 γ의 크기는 고려하지 않아도 무

Table 3.9. Extrapolation-efficiency for each combination of γ and model for each for each extrapolation point, $z_0 = 1.1, 2, 4$

z_0	γ	f_1	f_2
	1	0.9260	0.9260
1.1	2	0.9274	0.9501
	3	0.9290	0.9557
	1	0.7857	0.7857
2.0	2	0.8054	0.7860
	3	0.8067	0.7865
	1	0.7580	0.7580
4.0	2	0.7638	0.7592
	3	0.7581	0.7593

방하다는 결론이 나온다.

4. 결론과 토의

본 연구에서는 전통적인 외삽최적이 가지고 있는 문제점을 두 가지로 요약하여 보았다. 외삽점 $z \notin \Omega$가 실험영역인 Ω를 벗어나게 되면 가정된 모형과 오차항의 구조에 대한 불확실성이 증대되기 때문이다. 먼저 오차항의 분산에 관한 불확실성이 증대하는 경우 실험의 문제점을 알아보았다. 이를 해결하기 위하여 정의 2.1에서 정의한 실험기준을 설정하였다. 정의 2.2는 모형이 불확실한 경우이고 정의 2.3은 정의 2.1과 정의 2.3을 복합하여 maximin 방법을 적용하였다. 이분산과 외삽은 공통적인 특징을 가지고 있다. 외삽점은 실험영역에서 이탈하면 실험영역의 극단에 있는 받힘점에 배정되는 질량은 증가하고 실험영역을 따라서 분산의 크기가 커지는 이산구조라면 이분산이 존재하는 경우 역시 마찬가지로 실험영역의 극단에 있는 받힘점에 배정되는 질량은 증가하기 때문이다. 정의 2.3의 결과는 이를 반영하여 오히려 $\gamma = 1$에 국한하여 정의 2.2에서 정의한 모형의 불확실성이 있는 경우 이분산 외삽실험을 추천한다. 본 연구에서는 단순회귀모형과 다항회귀모형을 벤치마킹삼아 예제를 구성하였으나 이런 현상은 모든 모형에 적용할 수 있을 것이다. 다항회귀모형과 이분산의 복합적인 설정을 정확(exact) 실험설계로 코딩한다면 이는 외삽을 염두에 둔 많은 실험자에게 도움을 줄 것이다.

부록 A:

아래 모든 정리는 $f^T(x) = (1, x)$을 가정한다.

실험계획은 $\xi(1) = p, \xi(-1) = 1 - p$의 구조로 $\gamma \geq 1$이 주어졌을 때 $d(z_0, \xi)$의 식은 다음과 같다

$$d(z_0, \xi) = \frac{-p(z_0-1)^2 + \gamma(p-1)(z_0+1)^2}{4(p-1)p}, \quad z_0 > 1. \tag{A.1}$$

따라서 식 (A.1)을 최소화하는 p가 유도된다.

정리 A.1 외삽-최적의 해는 $\gamma \geq 1$와 $z_0 > 1$이 주어졌을 때 받힘점 1에 주어지는 질량 p는 다음 같이 주어진다.

$$p = \frac{\gamma(z_0+1)^2 - \sqrt{\gamma(z_0^2-1)^2}}{\gamma(z_0+1)^2 - (z_0-1)^2}. \tag{A.2}$$

따라서 이를 $d(z_0,\xi) = f^T(z_0)M^{-1}(\xi)f(z_0), z_0 \notin \Omega$에 대입하면 $d(z_0,\xi)$의 최저값을 구할 수 있는데 다음과 같이 단순화된다.

$$d(z_0,\xi^*) = \frac{\sqrt{\gamma(z_0^2-1)^2}\left((z_0^2-1)^2-\gamma(z_0+1)^2\right)^2}{4\left((z_0-1)^2\sqrt{\gamma(z_0^2-1)^2}+\gamma(z_0+1)^2\left(-2+4z_0-2z_0^2+\sqrt{\gamma(z_0^2-1)^2}\right)\right)}, \tag{A.3}$$

여기서 ξ^*는 외삽-최적 실험으로 $\xi(0) = 1-\xi(1)$, $\xi(1) = \{\gamma(z_0+1)^2 - \sqrt{\gamma(z_0^2-1)^2}\}/\{\gamma(z_0+1)^2 - (z_0-1)^2\}$이다.

참고로 $\gamma = 1$이면 $p = (1+z_0)/(2z_0)$, $d(z_0,\xi) = z_0^2$으로 단순화 되는데 예를 들어 $z_0 = 2$이면 $p = 0.75$가 되어 Table 3.1에서 확인된 $\xi(1)$의 질량이 나온다. 다른 γ와 z_0인 조합도 마찬가지로 대입하여 구하면 된다.

정리 A.2 $\gamma \geq 1$이 주어졌을 때 $\lim_{z_0\to\infty}\xi(1) = \sqrt{\gamma}/(1+\sqrt{\gamma})$.

예를 들면 z_0가 무한대로 가면 $\gamma = 1$인 경우 외삽-최적은 $\xi(\pm1) = 0.5$인 D-최적과 같아지고 $\gamma = 3$인 경우 $\sqrt{3}/(1+\sqrt{3}) = 0.6340$으로 $\xi(1)$은 수렴된다. 또한 z_0가 무한대로 가고 γ가 커질수록 $\xi(1)$의 질량이 1에 가까워진다. 이러한 사실에 입각하여 Table 3.1의 내용을 확인하여 보면 된다.

정의 2.1에 정의된 오차-강건 외삽 실험계획은 $\gamma \in [1,\gamma_0], \gamma_0 > 1$일 때 $\gamma = 1$과 $\gamma_0 > 1$에서 발생하는 외삽-효율이 같아지도록 질량 p을 부여하는 작업이 된다.

즉, $\gamma = 1$일 때는 외삽효율이 z_0^2이므로 실험계획 $\xi(1) = p, \xi(-1) = 1-p$에서의 외삽효율은 다음과 같다.

$$\frac{4(p-1)pz_0^2}{-p(z_0-1)^2+(p-1)(z_0+1)^2}. \tag{A.4}$$

$\gamma \in [1,\gamma_0], \gamma_0 > 1$인 경우 γ_0에서의 외삽효율은 다음과 같이 주어진다.

$$\frac{\dfrac{\sqrt{\gamma(z_0-1)^2}\left((z_0-1)^2-\gamma(z_0+1)^2\right)^2}{4\left((z_0-1)^2\sqrt{\gamma(z_0^2-1)^2}+\gamma(z_0+1)^2\left(-2+4z_0-2z_0^2+\sqrt{\gamma\left(z_0^2-1\right)^2}\right)\right)}}{\dfrac{-p(z_0-1)^2+\gamma_0(p-1)(z_0+1)^2}{4(p-1)p}}. \tag{A.5}$$

식 (A.4)와 식 (A.5)가 같아지는 p가 정의 2.1에서 정의한 오차-강건 외삽최적에 해당하는 p가 되므로 정리 A.3을 구할 수 있다.

정리 A.3 $\gamma \in [1,\gamma_0], \gamma_0 > 1$가 주어져 있을 때 정의 2.1에서 정의한 오차-강건 외삽-최적인 p의 구조는 다음과 같다.

$$p = \frac{(z_0+1)\left((z_0-1)^3\sqrt{\gamma_0(z_0^2-1)^2}-2\gamma_0\sqrt{\gamma_0(z_0^2-1)^2}(-1-z_0-z_0^2+3z_0^3)-\gamma_0^2(z_0+1)^2\left(8z_0^2-8z_0^3+\sqrt{\gamma_0(z_0^2-1)^2}+3z_0\sqrt{\gamma_0(z_0^2-1)^2}\right)\right)}{4z_0\left(-(z_0-1)^2+\gamma_0(z_0+1)^2\right)\left((z_0-1)\sqrt{\gamma_0(z_0^2-1)^2}+\gamma_0(z_0+1)\left(2z_0-2z_0^2+\sqrt{\gamma_0(z_0^2-1)^2}\right)\right)}.$$

다소 복잡해 보이는 해의 구조이지만 알고리즘으로 구한 해와 일치함을 밝힌다. 또한 정의 2.1에 의한 실험계획의 각 γ에 대하여 효율을 구할 수 있다. 이러한 p를 이용하여 $\gamma \in [1,\gamma_0], \gamma_0 > 1$에서의 최소 외삽-효율의 최대값의 공식은 복잡하고 실험자에게는 무의미한 공식이므로 여기서는 생략한다.

오차항의 이분산 구조와 모형이 달라질 때마다 해석적인 해를 구하는 작업은 그리 간단치 않기 때문에 본 연구에서는 단순회귀모형에 국한해서 해를 제공하였다. 응용적인 알고리즘이 더 중요한 측면이 있다.

References

Berger, M. P. F. and Wong, W. K. (2009). *An Introduction to Optimal Designs for Social and Biomedical Research*, Wiley, New York.

Chen, R. B., Wong, W. K. and Li, K. Y. (2008). Optimal minimax designs over a prespecified interval in a heteroscedastic polynomial model, *Statistics and Probability Letter*, **78**, 1914–1921.

Dette, H. and Biedermann, S. (2003). Robust and efficient designs for the Michaelis-Menten model, *Journal of the American Statistical Association*, **98**, 679–686.

Dette, H. and Wong, W. K. (1996). Robust optimal extrapolation designs, *Biometrika*, **83**, 667–680.

Fedorov, V. V. (1972). *Optimal Experimental Design*, Academic Press, New York.

Hoel, P. G. and Levine, A. (1964). Optimal spacing and weighting in polynomial prediction, *Annals of Statistics*, **35**, 1553–1560.

Imhof, L. and Wong, W. K. (2000). A graphical method for finding maximin designs, *Biometrics*, **56**, 113–117.

Kiefer, J. and Wolfowitz, J. (1960). The equivalence of two extremum problems, *Canadian Journal of Mathematics*, **12**, 363–366.

Kim, Y. I. and Jang, D. H. (2012). Hybrid constrained extrapolation experimental design, *Communications for Statistical Applications and Methods*, **19**, 65–75.

Kim, Y. I. and Jang, D. H. (2014). The maximin robust design for the uncertainty of parameters of Michaelis-Menten Model, *The Korean Journal of Applied Statistics*, **27**, 1269–1278.

Läuter, E. (1974). Experimental planning in a class of models, *Mathematishe Operationsforshung und Statistik*, **5**, 673–708.

Wong, W. K. and Cook, R. D. (1993). Heteroscedastic G-optimal designs, *Journal of the Royal Statistical Society, Series B*, **55**, 971–980.

모형과 오차구조의 불확실성하에서의 강건 외삽 실험설계

장대흥[a] · 김영일[b,1]

[a]부경대학교 통계학과, [b]중앙대학교 경영학부

(2015년 04월 22일 접수, 2015년 05월 24일 수정, 2015년 05월 25일 채택)

요 약

실험영역을 벗어나는 점에 해당하는 반응값 예측을 위한 최적실험을 고려할 때 실험에 필요한 받힘점을 위한 실험기준을 선택하는 경우 매우 신중하여야 한다. 왜냐하면 가정한 모형과 오차구도가 실험영역을 벗어나도 타당하다는 가정을 하여야 되기 때문이다. 따라서 기존문헌의 외삽최적의 실험기준을 이러한 상황에 맞게 설계될 수 있도록 수정하였다. 본 연구에서는 maximin방법을 적용하여 새로운 실험기준의 특징 및 강건성을 단순회귀모형과 이차회귀모형을 기준으로 검정하였다.

주요용어: 외삽, 외삽-최적, 최대최소 실험

[1]교신저자: (156-756) 서울특별시 동작구 흑석동 84, 중앙대학교 경영경제대학 경영학부.
E-mail: yik01@cau.ac.kr