분산감소기법을 이용한 파라미터 추정의 효율성

황성원*, 권치명*, 김성연*

Efficiency of Estimation for Parameters by Use of Variance Reduction Techniques

Sung-won Whang, Chi-myung Kwon and Sung-yeon Kim

Abstract

본 연구는 시뮬레이션 반응변수가 입력 인자의 선형 1차식으로 표현된 경우에 인자의 파라미터를 효과적으로 추정하기위해 사용될 수 있는 분산감소기법을 제안하였다. 이 기법은 하나의 실험설계에 공통난수와 대조난수를 동시에 사용하는 Schruben과 Margolin의 방법과 시뮬레이션하는 도중에 얻어지는 통제변수를 활용하는 기법을 결합하는 방법으로 시뮬레이션의 효율성을 개선하고자 하였다. 시뮬레이션 결과 제안된 기법은 주어진 모형의 평균 반응치를 추정한 데는 S-M 기법보다 효과적이었으며 인자의 다른 파라미터를 추정하는 데는 S-M 기법과 비슷한 성과를 보이고 있다. 만일 시뮬레이션 과정에서 반응변수와 상관성이 높은 통제변수들을 선택할 수 있는 경우에는 제안된 기법이 S-M 기법보다 보다 파라미터 추정에 효과적일 것으로 판단된다.

Key Words: Control Variates, Common Random Numbers, Antithetic Variates, Variance Reduction Techniques, Simulation Efficiency

* 동아대학교 경영정보과학부

1. 서론

분산감소기법(variance reduction technique)은 시뮬레이션 실험에서 시스템 반응변수의 값을 추정하는데 효과적으로 사용되고 있다. 종종 반응변수는 입력 인자의 1차 선형모형으로 가정되며 이러한 경우 입력 인자의 수준에서 모형의 파라미터를 정확히 추정하는 문제는 시뮬레이션의 효율성과 관련된다. 인자의 수준조합(표본점)에서 반응변수의 값과 모형의 파라미터를 효과적으로 추정하기위해 사용되는 대표적인 분산감소기법으로는 공통난수기법(common random numbers:CRN), 대조난수기법(anrithetic variates:AV), 통제변수기법(control variates method:CV)을 들 수 있겠다. 분산감소기법들에 대한 많은 연구 중에서 두 기법을 결합하여 하나의 시뮬레이션 실험설계 동시에 사용할 수 있는 방법에 대한 연구로 대표적인 것은 우선 Schruben과 Margolin[7]의 연구를 들 수 있는데 이들은 CRN과 AV 기법을 사용하여 모형의 파라미터를 추정하는 방법을 제안하였다. 통제변수기법은 보통 반응변수의 평균을 추정하는데 효율적으로 사용되는데 Nozari, Arnold와 Pegden[5]은 이 기법을 다중 표본점 모형에 사용하여 파라미터 추정의 효률성을 평가하였다. Kwon과 Tew[2]는 CV기법과 AV기법을 결합하여 대형 시뮬레이션 모형의 파라미터를 추정하는 방법을 제안하였다.

Schruben과 Margolin의 기법(S-M 기법)은 CRN 기법보다 우수한 것으로 알려졌으며 CV 기법은 표본점(입력인자의 조합)에서 반응변수의 평균을 추정하는데 매우 효과적인 기법으로 사용되고 있다[4, 6]. 이 방법은 시뮬레이션 도중에 반응변수와 상관성이 높은 시뮬레이션 모형인자를 통제변수로 선택하여 반응변수와 통제변수들 사이에 다중상관성을 이용하여 반응변수의 변이성을 감소시키는 방법이라고 볼 수 있다. S-M 기법은 실험설계에서 설계행렬(design matrix)의 표본점을 두 개의 직교 블럭(block)으로 나누고 같은 블럭의 표본점에는 CRN을 사용하고 다른 블럭에 속하는 표본점에는 AV 기법을 적용한다. 이 기법의 효율성은 같은 블록의 두 표본점 사이에서 유도되는 상관성과 관계가 깊다. 반면 CV 기법의 효율성은 반응변수와 통제변수 사이의 다중 상관성에 따라 결정된다. 보통 통제변수는 입력인자공간에서 독립적으로 관찰되며 주로 단일 점에서 반응변수를 정확히 추정하는 데 이용되며, 반응변수 사이에 관찰되는 상관성을 이용하여 파라미터의 추정치의 신뢰도를 개선시키는 S-M 기법과는 다른 면에서 시뮬레이션의 효율성을 개선시킨다고 볼 수 있다.

이러한 측점에서 S-M 기법을 적용하는 과정에서 얻어지는 통제변수를 효과적으로 사용하여 반응변수의 변이성을 감소시킬 수 있고 아울러 S-M 기법이 통제 반응변수(controlled response)의 두 블록에 속하는 표본점에도 통제변수를 사용하지 않는 경우의 표본점과 비슷한 상관성을 보인다면 두 방법의 장점을 살려 파라미터의 추정을 보다 효과적으로 달성할 수 있을 것으로 판단된다.

본 연구에서는 CV 기법과 S-M 기법을 결합하는 방법을 제안하고 결합된 방법이 S-M 기법보다 어떠한 조건에서 파라미터 추정에 효율적인가를 조사하고자 한다. 아울러 선정된 시뮬레이션 모형을 대상으로 시뮬레이션을 통한 효율성 문제를 다루고자 한다.

2. S-M 기법

시뮬레이션 시스템에서 입력 인자의 수가 p 개이고 인자의 각 조합수준(표본점) i에서 평균 반응변수 y_i가 인자들의 1차 회귀모형으로 근사될 수 있다면 y_i는 다음과 같이 표현된다.

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i, i = 1,2,...,m \quad (1)$$

여기서 x_{ij}는 표본점 i에서 인자 j의 수준이며 β_0는 미지의 상수이고 β_j는 1차식의 계수이며 $\epsilon_i \sim N(0, \sigma^2)$는 모형의 오차를 각각 나타낸다. 여기서 반응벡터를 $y = (y_1, y_2, ..., y_m)'$, 모형의 파라미터 벡터를 β, X을 계획행렬, 오

차 벡터를 $\epsilon = (\epsilon_1, ..., \epsilon_m)'$로 각각 정의하면 위의 식은 다음과 같은 행렬식으로 표현된다.

$$y = X\beta + \epsilon, \qquad (2)$$

여기서 $\epsilon \sim N(0, \sigma^2 I)$이며 설계행렬 X는 인자 수준의 re-parameterization 으로 직교한다고 가정하자.

설계행렬이 두 개의 직교하는 블록로 나눌 수 있을 때 Schruben과 Margolin은 표본점의 시뮬레이션에서 난수를 할당하는 방법으로 CRN과 AV를 제안하였다. 만일 CRN을 두 표본점에 사용하면 두 표본점에서의 반응치는 양의 상관관계(ρ_1)를 가지고 두 표본점에서 AV를 사용하면 두 표본점에서의 반응치는 음의 상관관계(ρ_2)를 가지는 가지는 것으로 알려져 있다. S-M 기법은 (a) 같은 불럭에 속하는 모든 표본점에는 CRN을 할당하여 시뮬레이션을 수행하고 (b) 다른 불럭에 있는 표본점에는 CRN과 대조되는 AV를 할당하여 시뮬레이션을 수행하는 난수할당법을 사용한다. 독립적인 난수(independent random numbers)를 모든 표본점에 할당하여 시뮬레이션을 수행할 때 모형 (2)의 파라미터 β의 최소자승 추정량 (ordinary least squares estimator: OLS)과 그 분산은 각각 다음과 같다.

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'y;$$

$$Cov(\hat{\beta}_{OLS}) = \sigma_y^2(X'X)^{-1}] \qquad (4)$$

설계행렬을 두개의 직교 블록으로 나눌 수 있는 경우 S-M 기법에 의한 난수할당법으로 얻은 파라미터 β의 최소자승 추정량은 (4)와 같으며 그 분산은 다음과 같다 [7].

$$Cov(\hat{\beta}_{OLS}) = \sigma_y^2[(\rho_1 + \rho_2)G/2 + (1-\rho_1)(X'X)^{-1}] \qquad (5)$$

여기서 G는 (mxm)행렬로 첫 번째 행과 열의 값이 1이고 나머지 원소의 값은 모두 0인 행렬이다.

파라미터 추정량의 분산 측면에서 독립적인 난수할당법과 S-M 방법의 난수할당법을 비교하면 식 (4)-(5)으로부터 $\beta_j (j = 1, 2, ..., p)$를 추정하는데는 S-M 방법이 독립적인 난수할당법보다 우수하며 β_0를 추정하는데는 독립적인 난수할당법이 S-M 방법보다 우수하다는 사실을 알 수 있다. 식 (5)에서 보는바와 같이 S-M 기법의 효율성은 CRN에 의해 같은 블록의 두 반응치 사이에 유도되는 상관계수의 값(ρ_1) 따라 결정된다. 따라서 S-M 방법과 결합하여 CV 기법을 효율적으로 적용하려면 (a) CV에 의하여 같은 블록의 조정된 두 반응치사이에 나타나는 양의 상관관계의 크기와 (b) 각 표본점에서 반응치의 분산을 감소시키는 데 CV의 효과가 어느 정도인가에 달려있다고 할 수 있다.

3. 통제변수 기법의 응용

보통 통제변수는 인자의 각 수준에서 독립적으로 관찰되며 각 통제변수의 평균은 알려져 있다. 표본점 i에서 얻은 s개의 통제변수의 평균 벡타를 c_i라하자. 만일 각 표본점에서 시뮬레이션 런 시간이 충분히 크다면 평균 반응치와 통제변수의 평균 벡타는 다음과 같은 다변량 정규분포를 따르는 것으로 생각할 수 있다 [3].

$$\begin{pmatrix} y_i \\ c_i \end{pmatrix} \sim N_{s+1}\left[\begin{pmatrix} \mu_i \\ \mu_c \end{pmatrix}, \begin{pmatrix} \sigma_y^2 & \sigma_{yc}' \\ \sigma_{yc} & \Sigma_c \end{pmatrix}\right]. \qquad (6)$$

여기서 $\mu_i = E(y_i)$, $\mu_c = E(\mu_c)$, $\sigma_{yc} = Cov(y_i, c_i)$이며 $\Sigma_c = Cov(c_i)$이다. 위의 가정이 성립하는 경우, m개의 표본점에서 통제 반응치(controlled response)는 m-변량 정규분포를 따르며 또한 통제 반응치

$$y_i(\alpha_i) = y_i - \alpha_i' c_i \qquad (7)$$

는 평균 반응치의 불편추정량으로 알려져 있다 [4]. 따라서 선형모형 (2)에서와 비슷하게 통제 반응치는 다음과 같은 선형모형으로 나타낼 수 있다.

$$y(A) = X\beta + \epsilon^*, \qquad (8)$$

단, 여기서 $y(A) = (y_1(\alpha_1), ..., y_m(\alpha_m))'$이며 ϵ^*는 에러 벡타이다.

위의 모형 (8)에 S-M 기법의 난수할당법을 적용하여 β의 추정량과 그 분산을 각각 구하면 다음과 같은 식으로 주어진다.

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'y(A);$$

$$Cov(\hat{\beta}_{OLS}) = \sigma_y^2[(\rho_1 + \rho_2 - R_{yc}^2 - R_{yc}^*)G/2 + (1-\rho_1)(X'X)^{-1}] \qquad (9)$$

여기서 R_{yc}^2는 반응변수와 통제변수 벡타 사이의 다중 상관계수의 제곱이며, R_{yc}^*는 두 블록에 속하는 반응변수와 통제변수 벡타 사이의 복잡한 관계식으로 기술된다.

식 (5)와 (9)의 해석적인 비교를 통하여 모형 추정량의 분산 측도 측면에서 두 기법의 우월성을 평가하는 것이 바람직하나 이는 쉬운 일이 아니므로 대신 시뮬레이션을 통하여 두 기법의 효율성을 비교 평가하고자 한다.

4. 시뮬레이션 실험 및 결과분석

4.1 시뮬레이션 모형

위에서 언급한 두 기법의 효율성을 비교하기 위해 S-M 기법이 매우 효과적으로 적용된 병원 자원배분 문제를 대상으로 시뮬레이션 실험을 실시하였다 [7]. 병원에 도착하는 환자는 도착률이 3.3명/일인 Poisson 분포를 따르며 도착 환자의 75%는 Itensive Care(IC)를 받고 나머지 25%는 Coronary Care(CC)를 요한다. IC 서비스 시간은 평균과 분산이 각각 3.4와 3.5인 lognormal 분포를 따르며 IC 치료 환자의 27%는 병원을 떠나며 73%는 Intermidiate Care(InC)로 이동하며 InC에서의 서비스 시간은 평균과 분산이 각각 15.0, 7.0인 lognormal 분포를 한다. CC의 서비스 시간은 평균과 분산이 각각 3.8과 1.6인 lognormal 분포를 따르고 CC 치료 환자의 20%는 병원을 떠나며 80%는 InC로 이동하며 이러한 환자의 경우, InC에서의 서비스 시간은 평균과 분산이 각각 17.0과 3.0인 lognormal 분포를 한다. 도착 환자는 병원 시설이 부족하면 시스템을 이탈하며(balking), 3가지 종류의 병원 시설을 어떤 수준으로 결정하는가에 따라 병원에 수용될 수 있는 환자에 대한 시스템 이탈 수준이 달라지게 된다. Schruben과 Margolin은 병원 이탈 비율(반응치)에 영향을 미치는 인자로 병원의 3가지 치료시설의 크기를 가정하였으며 세가지 인자의 주요인(main effect)과 두 인자 간의 3가지 상호작용(pairwise interaction)을 모형의 독립변수로 도입하여 2^3요인 시뮬레이션 실험을 수행하였다(표 1 참조).

< 표 1> 2^3 요인 실험의 표본점

표본점		IC	CC	InC
블록 1	1	13(-1)	4(-1)	15(-1)
	2	13(-1)	6(+1)	17(+1)
	3	15(+1)	4(-1)	17(+1)
	4	15(+1)	6(+1)	15(-1)
블록 2	5	13(-1)	4(-1)	17(+1)
	6	13(-1)	6(+1)	15(-1)
	7	15(+1)	4(-1)	15(-1)
	8	15(+1)	6(+1)	17(+1)

실험 결과 2요인에 의한 상호작용 효과는 거의 무시할 수 있어 전체 평균(β_0)과 주 요인 효과(β_1, β_2, β_3)만을 모형에 포함시켰다. <표 1>의 표본점 i 행에서의 수는 IC, CC, InC의 수준을 의미하며 괄호 안의 수는 reparamete -rization을 통하여 얻은 인자의 수준이다. 블록 1과 2의 표본점에 대응하는 설계행렬은 직교행렬임을 알 수 있으며 통제변수는 환자의 도착간 시간을 표준화하여 사용하였다. 이 모형을 AweSim으로 모델링하고 각 표본점의 시뮬레이션 시간은 1,500일로 하였다.

4.2 시뮬레이션 결과 분석

8개 표본점에서 얻은 반응치의 분산은 S-M 기법에 의한 것이 대략 1.88-2.19 정도이고 제안된 기법으로 얻은 것이 0.43-0.67 정도이었다. 같은 블록에 속한 두 반응치 사이의 상관계수는 S-M 기법에 의한 것이 대략 0.98-0.99 이며

제안된 기법은 0.91-0.97정도 상관계수를 나타내었다. 다른 두 블록에 위치한 두 표본점에서의 반응치 사이의 상관계수는 S-M 기법에 의한 것이 대략 -0.54와 -0.51 사이의 값으로 나타났으며 제안된 기법은 대략 -0.23에서 -0.19 정도의 값으로 나타났다. <표 2>는 병원 자원 배분 문제에 적용한 두 기법으로부터 얻은 파라미터의 추정치와 그 분산을 각각 요약하였다. β_0를 추정하는데는 제안된 기법이 우수하였으며 주요인을 추정하는데는 두 기법이 비슷한 수준의 효율성을 보이고 있다.

< 표 2> 파라미터 추정치와 분산

파라미터	S-M 방법		제안된 기법	
	추정치	분산	추정치	분산
β_0	45.722	0.472	45.670	0.200
β_1	-0.291	0.003	-0.290	0.003
β_2	-0.378	0.005	-0.378	0.005
β_3	-1.805	0.001	-1.805	0.001

5. 결론

같은 블록에 속하는 두 표본점 사이의 상관계수의 값이 클 경우 S-M 기법은 파라미터를 추정하는데 매유 효율적이라고 볼 있는데 병원 자원 배분 모형에서는 그 값이 098-0.99 정도로 나타나 S-M 기법은 추정량의 분산을 감소시키는데 매우 효과적이었다. 비록 제한된 시뮬레이션 결과이지만 이러한 모형에서도 제안된 기법은 S-M 기법보다 파라미터를 추정하는데 우수한 결과를 보이고 있다.

만일 반응변수와 깊은 상관성을 보이는 통제변수의 집합을 선택할 수 있고 모형의 특성상 두 표본점에서 공통난수의 할당에 따라 나타나는 syncronization 효과가 적을 경우 S-M 기법에 통제변수 기법을 결합하는 방법이 S-M 방법만을 사용하는 것보다 우수할 것으로 판단된다.

참고문헌

[1] Anderson T.W.: An Introduction to Multivariate Statistical Analysis. John Wiley & Sons, New York (1984).

[2] Kwon, C. and Tew, J.D.: Combined Correlation Methods for Meta-model Estimation in Multi-population Simulation Experiments. J. Statistical Computation and Simulation, Vol. 49. (1994) 49-75.

[3] Kwon, C and Tew, J.D.: Combining Antithetic and Control Variates in Designed Simulation Experiments. Management Science, Vol. 40. (1994) 1021-1034.

[4] Lavenberg, S.S., Moeller, T.L., Welch, P.D.: Statistical Results on Control Variates with Application to Queuing Simulation. Operations Research, Vol. 27. (1982). 182-202.

[5] Nozari, A., Arnold, S.F., Pegden, C.D.: Control Variates for Multi-population Simulation Experiments. IIE Trans., Vol. 16. (1984) 159-169.

[6] Rubinstein, R.Y. and Marcus, R.: Efficiency of the Multi-variates Control Variates in Monte Carlo Simulation. Operations Research, Vol. 33. (1985) 661-677.

[7] Schruben, L.W., Margolin, B.H.:Pseudorandom Number Assignment in Statistically Designed Simulation and Distribution Sampling Experiments. JASA, Vol. 73. (1978) 504-525.

[8] Pritsker, A.A.B., O'Reilly, J.J.: Simulation with Visual SLAM and AWeSim. John Wiley & Sons, New York (1999).