한국시뮬레이션학회 논문지
Vol. 19, No. 1, pp. 133-141 (2010. 3)

층화추출에 의한 통제변수의 시뮬레이션 성과분석

권치명[1] · 김성연[1] · 황성원[1†]

Simulation Analysis of Control Variates Method Using Stratified sampling

Chi-Myung Kwon · Seong-Yeon Kim · Sung-Won Hwang

ABSTRACT

This research suggests a unified scheme for using stratified sampling and control variates method to improve the efficiency of estimation for parameters in simulation experiments. We utilize standardized concomitant variables defined during the course of simulation runs. We first use these concomitant variables to counteract the unknown error of response by the method of control variates, then use a concomitant variable not used in the controlled response and stratify the response into appropriate strata to reduce the variation of controlled response additionally. In case that the covariance between the response and a set of control variates is known, we identify the simulation efficiency of suggested method using control variates and stratified sampling. We conjecture the simulation efficiency of this method is better than that achieved by separated application of either control variates or stratified sampling in a simulation experiments. We investigate such an efficiency gain through simulation on a selected model.

Key words : Simulation Efficiency, Stratified Sampling, Control Variates Method, Variance Reduction Method

요 약

본 연구는 층화추출법과 통제변수기법을 시뮬레이션 실험설계에서 동시에 사용하여 파라미터의 추정 효율성을 개선하는 방법을 제안하였다. 시뮬레이션을 수행하는 도중에 수집한 표준화 부가변수를 통제변수기법에 활용하여 파라미터 추정을 위한 반응변수의 오차를 줄이도록 하였으며 통제변수 구성에 사용되지 않은 부가변수를 층화변수로 활용하는 층화추출기법을 적용하여 통제반응변수의 변이성을 추가로 감소하였다. 반응변수와 통제변수 사이의 공분산 관계가 알려진 경우에 두 방법을 동시에 활용하여 파라미터를 추정하는 기법의 시뮬레이션 효율성을 이론적으로 도출하였다. 반응변수와 통제변수 사이의 상관관계 구조가 알려지지 않은 경우 선택된 모형의 시뮬레이션 결과는 제안된 층화 통제 추정기법이 층화추출법이나 통제변수기법을 개별적으로 적용하여 파라미터를 추정하는 것보다 우수한 것으로 나타나고 있다.

주요어 : 시뮬레이션 효율성, 층화추출, 통제변수기법, 분산감소기법

1. 서 론

복잡한 대기행렬 시스템의 연구에서 이론적인 측면이나 실제 응용 측면에서 시뮬레이션 분석이 자주 활용된다. 시뮬레이션에 의한 분석은 시뮬레이션 실험과 관련된 비용이 많이 소요될 수 있다. 시뮬레이션을 통한 시스템 성과(반응변수) 파라미터의 추정에서 추정치의 정확성을 사용자가 수용할 수 있는 수준이 되게 하려면 방대한 량의 시뮬레이션 런이 요구될 수도 있다. 따라서 같은 량의 시뮬레이션 런을 수행하여 시스템 반응변수에 대한 추정치의 정확도를 높일 수 있는 방법에 대한 연구는 의미가 있으며 이러한 목적으로 시뮬레이션 실험 설계에서 성과 추정치의 효율성을 개선하는 분산감소기법이 제안되었다.

분산감소기법 중에서 부가변수를 활용하여 시뮬레이션의 효율성을 재고하는 방법으로는 통제변수기법(control variates method)과 층화추출법(stratified sampling method)이 사용되고 있다[3]. 통제변수기법은 반응변수와 부가변

* 이 논문은 2008학년도 동아대학교 학술연구비(단기연구 파견과제)에 의하여 연구되었음.
2009년 12월 1일 접수, 2010년 3월 11일 채택
[1] 동아대학교 경영대학 경영정보학과
주 저 자 : 권치명
교신저자 : 황성원
E-mail; sense64@gmail.com

수(concomitant variable) 사이의 선형관련성을 이용하여 반응변수의 추정치의 분산을 감소시키며 층화추출법은 반응변수와 부가변수 사이에 내재한 복잡한 비선형 관계를 이용하여 반응변수를 층화함으로써 추정치의 분산을 감소시키는 확률표본추출법을 사용한다[4].

본 연구에서는 시뮬레이션 과정에서 반응변수와 함께 적은 비용으로 다수의 부가변수를 수집하고 이를 활용하여 반응변수의 변이성을 감소시키고자 한다. 수집 부가변수가 반응변수와 선형상관성이 높을 경우에 통제변수기법은 반응변수를 추정하는데 매우 효과적인 기법으로 알려져 있다. 층화변수기법은 층화변수가 반응변수를 효과적으로 층화할 수 있는 경우 반응변수의 정도(precision)를 효율적으로 개선시킬 수 있다. 층화변수 기법은 부가변수가 반응변수와 선형상관성이 높을 경우 효과적일 수 있지만 비선형인 경우에도 사용될 수 있다. 따라서 수집 부가변수 중에서 반응변수와 선형관련성이 적으나 반응변수를 효과적으로 층화할 수 있는 1개의 변수를 층화변수로 선정할 수 있으면 수집 부가변수를 통제변수와 층화변수로 각각 사용하여 층화추출법과 통제변수기법을 동시에 시뮬레이션 실험에 적용하여 반응변수 추정에 활용할 수 있을 것으로 기대한다.

본 연구는 선형상관성은 적으나 반응변수를 효과적으로 층화하는 부가변수를 선정하고 이를 이용하여 반응변수를 층화하고 층화된 내부에서는 층화에 사용되지 않은 부가변수를 이용하여 반응변수의 변이성을 감소시키는 방안을 연구하고자 한다. 이러한 방법은 시뮬레이션 실험 설계에서 부가변수를 활용하여 반응변수 추정치의 분산을 감소시키는 통제변수기법과 층화추출법을 동시에 적용하는 것으로 만일 두 기법의 장점이 반응변수의 추정에 함께 반영될 수 있다면 반응변수의 추정 정도는 개별 기법을 독립적으로 적용하여 추정하는 경우보다 개선될 수 있을 것으로 기대할 수 있다. 반응변수와 관련성을 가지는 부가변수를 층화변수와 통제변수로 활용하는 두 기법의 결합 방법에서 반응변수 파라미터 추정치의 특성과 그 효율성을 조사하고자 한다. 이를 위해 본 연구에서는 결합 방법에 의한 반응변수의 추정치를 불편성과 유효성의 측면에서 이론적으로 고찰하고 선택된 모형에 대한 시뮬레이션을 통하여 그 타당성을 조사하고자 한다.

2. 관련 연구

시스템 반응변수 y의 기대치 $E(y)=\mu_y$을 추정하기 위해 시간 $(0,\ t)$ 동안 시뮬레이션 런을 독립적으로 n회 수행하는 반복분석에서 μ_y의 단순추정량으로 n회 독립적인 시뮬레이션 런으로부터 얻은 반응변수 $y_i(i=1,\ ...,\ n)$의 표본평균, $\bar{y}$가 사용된다. 분산감소기법은 원래의 단순 확률표본 추출 과정을 새로운 표본추출 과정으로 대체하여 얻은 표본으로부터 μ_y에 대한 새로운 추정치를 제공한다.

새로운 추정치 $\hat{\mu}_y$가 μ_y의 불편추정치이며 만일 그 분산이 단순 추정치의 분산보다 적게 되면 새로운 추정치는 단순 추정치에 비해 효율적으로 μ_y을 추정한다고 볼 수 있다. 통제변수기법과 층화추출법은 시뮬레이션 도중에 반응변수와 관련이 있는 변수, 즉 부가변수를 수집하고 그 정보를 활용하여 이러한 목적을 달성한다. 층화추출법에서는 반응변수를 층화하는 기준으로 부가변수를 층화변수로 사용하고 통제변수기법에서는 부가변수를 반응변수의 변이성을 축소하는 통제변수로 사용한다.

본 장에서는 관심 반응변수 기대치의 추정을 효율적으로 하기 위해서 시뮬레이션 도중에 반응변수와 함께 수집하는 부가변수의 설정과 그 분포 그리고 이를 활용하는 통제변수기법과 층화추출법의 추정 효율성에 대해 살펴보고자 한다.

2.1 부가변수의 표준화와 분포함수

시뮬레이션 실험에서 시스템 요소의 확률적인 재현은 알려진 확률분포나 경험적인 확률분포로부터 확률변수를 추출하는 과정이다. 이러한 확률변수 중에는 시스템 반응변수 y와 어느 정도 선형상관성이 있는 확률변수가 있을 수 있는데 보통 이러한 변수가 부가변수로 선택된다. 예를 들어 대기행렬 네트워크의 k번째 대기행렬에서 평균이 μ_k이고 표준편차가 σ_k인 확률분포로부터 재현되는 서비스 시간은 시스템 체재시간이라는 반응변수와 어느 정도 선형상관성을 가질 수 있다. 여기서 $s_j(k)$을 시뮬레이션 과정에서 난수의 부여로 재현되는 k번째 대기행렬의 j번째 서비스 시간이라고 하면 $\{s_j(k):j\geq 1\}$는 평균이 μ_k이고 표준편차가 σ_k인 확률분포로부터 추출된 확률표본으로 생각할 수 있다.

시뮬레이션 런 시간 $(0,\ t)$ 동안 k번째 대기행렬에서 추출된 서비스 시간의 표본 개수를 $n(k,t)$라 할 때 표준화 부가변수(standardized concomitant variable)는 다음과 같이 정의할 수 있다.

$$c_k(t)=n(k,t)^{-1/2}\sum_{j=1}^{n(k,t)}[s_j(k)-\mu_k]/\sigma_k \qquad (1)$$

이 식에서 만일 시뮬레이션 런 시간 t가 충분히 커지게 될 때 $c_k(t)$를 c_k로 표시하면 표준화 부가변수 c_k의 분포는 평균이 0이고 분산이 1인 정규분포에 수렴한다[10].

$$c_k \rightarrow N(0, 1) \quad as \ \ t \rightarrow \infty \tag{2}$$

식 (1)에서 수집 부가변수의 수가 1개인 경우는 부가변수의 수가 q인 일반적인 경우로 확장될 수 있다. 시뮬레이션 과정에서 수집된 q개의 부가변수, 즉 $c_k(t)$ $(k=1, 2, .., q)$는 시뮬레이션 모형의 확률적 요소에 서로 독립적인 난수를 부여하여 실현되는 확률변수이므로 q개의 부가변수는 서로 독립적이다. 이러한 경우 시뮬레이션 런 시간 t가 충분히 커지게 되면 q-변량 부가변수 벡터 $C=(c_1, c_2, \cdots, c_q)$는 q-변량 정규분포(q-variates normal distribution)로 수렴한다[10].

$$C \rightarrow N_q(0_q, I_q) \quad as \ \ t \rightarrow \infty \tag{3}$$

단 여기서 0_q와 I_q는 각각 q 차원의 0 벡터와 단위행렬이다.

2.2 층화추출법

층화추출법은 표본조사에서 반응변수 y의 기대치 μ_y에 대한 추정치의 정도를 높이기 위해 자주 사용되는 표본추출 방법이다. 층화추출법으로 반응변수의 기대치를 추정하려면 반응변수와 이를 효과적으로 층화할 수 있는 층화변수를 시뮬레이션 도중에 부가적으로 수집하여야 한다. 시간 $(0, t)$동안 시뮬레이션 런을 독립적으로 n회 반복수행하는 경우, i번째 런으로부터 얻은 반응변수 y_i와 층화변수 s_i의 쌍$\{(y_i, s_i) : 1 \le i \le n\}$에 대해 층화변수 s_i가 소속하는 구간 $(\omega_{h-1}, \omega_h]$ $(h=1, 2, \cdots, L)$에 따라 반응변수$\{y_i : 1 \le i \le n\}$는 L개의 층으로 층화된다. 독립적인 n회 시뮬레이션 결과로부터 L개의 층으로 층화된 반응변수의 수를 각각 $n_h (1 \le h \le L)$라 하고 층 h에서 반응변수의 평균을 $\bar{y}_h$라 하면 μ_y의 층화 추정량 $\bar{y}_{st}$는 다음과 같다[2].

$$\bar{y}_{st} = \sum_{h=1}^{L} n_h \bar{y}_h / n \tag{4}$$

각 층에서 반응변수의 평균은 반응변수가 단순 확률추출된 경우이므로 $E[\bar{y}_{st}] = \mu_y$ 이다. 시뮬레이션 런 시간 t가 충분히 클 때 표준화 부가변수를 층화변수를 s_i로 정의하면 s_i의 확률분포는 표준정규분포 $N(0, 1)$에 수렴한다. 층화변수가 h번째 층에 속하는 확률을 $\pi_h = \Pr[\omega_{h-1} < s_i \le \omega_h]$로 정의하면 h층 반응변수 $\{y_{hi} : 1 \le i \le n_h\}$의 기대치와 분산은 다음과 정의할 수 있다.

$$\mu_{yh} = E[y \mid \omega_{h-1} < s_i \le \omega_h] \tag{5}$$

$$\sigma_{yh}^2 = E[(y-\mu_{yh})^2 \mid \omega_{h-1} < s_i \le \omega_h] \tag{6}$$

시뮬레이션에서 반응변수 y_i는 일반적으로 정규분포 $N(\mu_y, \sigma_y^2)$를 따르는 표본으로 가정하는데 이와 유사하게 층화된 각 층에서의 반응변수도 정규분포 $N(\mu_{yh}, \sigma_{yh}^2)$를 따르는 것으로 가정할 수 있다. 만일 관찰된 표본의 수 $N=n$에서 독립적으로 수행한 시뮬레이션 런의 수 n가 적지 않고 층의 수 L이 크지 않는 경우 h층에 속하는 반응변수가 하나도 없을 확률 $\Pr(N_h = 0)$은 거의 무시할 수 있을 정도로 적다$(h=1, 2, \cdots, L)$. 이러한 경우 μ_y의 층화 추정량 $\bar{y}_{st}$은 다음과 같이 표현할 수 있다[2].

$$\bar{y}_{st} = \sum_{h=1}^{L} \pi_h \bar{y}_h \tag{7}$$

위 식에서 층화 추정량은 $E[\bar{y}_{st}] = \mu_y$이며 그 분산은 다음과 같다.

$$Var(\bar{y}_{st}) = \sum_{h=1}^{L} \pi_h^2 n_h^{-1} \sigma_{yh}^2 \tag{8}$$

표준화 부가변수를 층화변수로 사용하려면 먼저 층의 수 L과 층화변수의 층 구간 $(\omega_0 \equiv -\infty, \omega_1, \cdots, \omega_L \equiv \infty)$을 지정하여야 한다. 층화변수가 표준정규분포를 따르는 경우 층화 추정량의 분산을 최소로 하는 Sethi[8]의 최적 층화 전략에 의하면 적절한 층의 수 L는 $2 \le L \le 6$으로 h층에 속하는 표본의 수는 $n\pi_h \ge 10$ $(1 \le h \le L)$ 조건을 만족해야 한다[2]. 부가변수 s_i의 확률분포가 표준정규분포 $N(0, 1)$에 수렴한다는 조건하에서 Wilson과 Pritsker[10]는 추정치의 분산을 최소화하는 최적 층화 전략으로 층의 수와 이에 따른 각 층의 구간 상한을 다음 표와 같이 제시하였다.

표 1. 층의 수에 따른 각 층의 상한(ω_h)

층수 (L)	층					
	1	2	3	4	5	6
2	0.000	∞				
3	-0.612	0.612	∞			
4	-0.982	0.000	0.982	∞		
5	-1.244	-0.382	0.382	1.244	∞	
6	-1.447	-0.659	0.000	0.659	1.447	∞

층의 수 L이 지정되면 층 h에 속하는 층화변수의 확률 π_h은 표준정규분포로부터 쉽게 계산된다. 시뮬레이션 결과로부터 h층에 속하는 반응변수의 평균과 분산을 각각 $\bar{y}_h$와 S_h^2이라하면 $\{(\bar{y}_h, S_h^2) : 1 \le h \le L\}$는 서로 독립적이다. 그리고 μ_y의 층화 추정량 $\bar{y}_{st}$는 μ_y의 불편 추정량이 되며 $\bar{y}_{st}$의 분산 추정량 $\hat{V}ar(\bar{y}_{st})$은 다음 식과 같으며

$$\hat{V}ar(\bar{y}_{st}) = \sum_{h=1}^{L} \pi_h^2 n_h^{-1} S_h^2 \tag{9}$$

이 분산추정량은 식 (8)의 $Var(\bar{y}_{st})$에 대한 불편 추정량이 된다[10].

2.3 통제변수기법

통제변수기법은 반응변수 y의 기대치 μ_y을 추정하기 위해 시뮬레이션 도중에 반응변수 y와 선형상관관계를 가지는 변수들, 즉 통제변수 벡터 $C = (c_1, c_2, ..., c_q)$를 부가적으로 수집하고 반응변수와 통제변수 벡터 사이의 상관관계를 이용하여 μ_y의 추정량에 대한 변이성을 감소시키는 기법이다. 시간 $(0, t)$동안 시뮬레이션 런을 독립적으로 n회 반복수행하는 시뮬레이션 실험에서 표준화 통제변수 벡터 C를 사용하는 μ_y의 통제추정량 $\bar{y}(C)$는 다음과 같이 정의할 수 있다[3].

$$\bar{y}(C) = \bar{y} - \overline{C}\alpha \tag{10}$$

여기서 $\bar{y}$와 $\overline{C}$는 각각 반응변수 y_i와 통제변수 C_i의 평균과 평균 벡터이며 α는 통제변수 벡터의 q차원 계수 벡터이다. 식 (10)에서 통제변수 벡터가 표준화 통제변수로 구성되었다고 가정하면 각 통제변수의 기대치는 0임으로 $E[\overline{C}] = 0_q$이다. 따라서 통제추정량 $\bar{y}(C)$는 μ_y의 불편추정량이 된다. 또한 통제변수의 공분산행렬을 Σ_C 그리고 반응변수와 통제변수와의 공분산 벡터를 σ_{yC} 로 표시하면 $Var(\bar{y}(C))$는 $\alpha = \Sigma_C^{-1}\sigma_{yC}$ 일 때 최소가 되며 그 값은 다음과 같다[5].

$$\begin{aligned} Var(\bar{y}(C)) &= n^{-1}(\sigma_y^2 - \sigma_{yC}^{'}\Sigma_C^{-1}\sigma_{yc}) \\ &= n^{-1}(1 - R_{yC}^2)\sigma_y^2 \end{aligned} \tag{11}$$

이 식에서 $R_{yC}^2 = \sigma_y^2\sigma_{yC}^{'}\Sigma_C^{-1}\sigma_{yC}$는 y와 C 사이의 다중상관계수의 제곱이며 다중상관계수가 커질수록 통제추정량은 반응변수의 기대치를 추정하는데 효율적이다. 일반적으로 식 (10)에서 계수 벡터 α의 최적 값은 알 수 없으므로 시뮬레이션 런을 독립적으로 n회 반복수행하여 얻은 반응변수와 통제변수의 결과로부터 추정해야한다. S_C와 S_{yC}를 각각 표본으로부터 구한 Σ_C와 σ_{yC}의 표본추정치라고 하면 α의 표본추정치는 $\hat{\alpha} = S_C^{-1}S_{yC}$ 이며 μ_y의 통제추정량 $\bar{y}(C)$의 분산은 다음과 같다[5].

$$Var[\bar{y}(C)] = n^{-1}[(n-2)/(n-q-2)](1 - R_{yc}^2)\sigma_y^2 \tag{12}$$

위 식에서 통제변수의 효율성은 반응변수와 통제변수 사이의 다중상관계수 R_{yC}^2에 의하여 결정되지만 이 식의 두 번째 항 $(n-2)/(n-q-2)$의 값은 통제변수의 수 (q)가 증가하면 1보다 상당히 큰 값이 될 수 있다. 여기서 $(n-2)/(n-q-2)$를 손실인자(loss factor)라 부르는데 반응변수와 상관성이 적은 통제변수가 통제추정량에 포함되면 손실인자의 값은 증가하지만 다중상관계수는 별 증가가 없을 수 있어 전체적으로는 통제변수 추정량의 분산이 증가할 수 있다.

3. 층화통제변수기법

통제변수기법은 통제변수와 반응변수 사이에 내재한 선형관계를 이용하여 반응변수에 대한 통제추정량의 분산을 감소시키며 층화추출법은 반응변수와 층화변수 사이에 내재한 비선형 관계를 이용하여 추정치의 정도를 개선하고자 한다. 시뮬레이션 실험설계에서 반응변수와 통제변수 사이의 관계를 활용하여 통제반응변수를 구성한 다음 이들에 대하여 층화추출법을 적용하여 반응변수 기대치의 추정 정도를 높이고자 한다. 이러한 방법은 부가변수를 활용하여 추정치의 분산을 감소시키는 통제변수 기법과 층화추출기법을 결합하는 것으로 시뮬레이션 실험설계에 동시에 적용할 경우 두 기법을 개별적으로 적용하는 것보다 추정 효율성이 개선될 것으로 기대한다. 본 장에서는 통제변수 및 층화변수를 활용하기 위한 부가변수의 선정과 통제추정치의 층화 전략, 층화통제 추정량의 제안 및 추정치의 효율성에 대해 알아보고자 한다.

3.1 층화변수와 통제추정량의 분포

반응변수와 관련성을 가지는 부가변수는 시뮬레이션 모형의 구조와 반응변수의 특성에 따라 달라진다. 서비스 시스템의 시뮬레이션에서 관심 반응변수가 개체의 시스템 체제시간인 경우에 개체가 서비스를 받는 빈도가 높은

서버의 서비스 시간이 보통 부가변수로 선택된다[10]. 시뮬레이션 과정에서 수집된 부가변수는 반응변수와 선형적이거나 또는 비선형적인 관계를 가질 수 있으며 또한 거의 이러한 관계를 보이지 않을 수도 있다. 수집된 부가변수 군에서 어떤 변수를 통제변수로 사용할지 아니면 층화변수로 사용할지에 대한 명확한 기준은 없으나 반응변수와 깊은 선형관계를 나타내는 변수들은 일반적으로 좋은 통제변수 후보가 될 수 있다. 그리고 부가변수의 값에 따라 층화된 반응변수에서 층 내부의 반응변수는 변이성이 적으며 층간 반응변수는 변이성이 크게 되는 경우 이 부가변수는 좋은 층화변수의 조건을 갖추게 된다.

시뮬레이션 과정에서 시스템 요소의 확률적 재현은 독립적인 난수로부터 생성된 확률 표본으로 생각할 수 있다. 반응변수와 관련성을 가지는 시스템 요소의 확률표본으로부터 식 (1)과 같은 표준화 부가변수를 $(q+1)$개 정의하고 그 중 1개의 변수 s는 층화변수로 나머지 q차원의 부가변수는 통제변수 벡터 $C=(c_1, c_2, ..., c_q)$로 사용하고자 한다. 서로 독립적인 난수의 부여로 실현된 부가변수들은 독립적이므로 시뮬레이션의 런 시간 t가 충분히 클 때 독립적인 n회 반복 시뮬레이션 런으로부터 얻은 반응변수 y와 $(q+1)$차원의 부가변수 벡터 $(s, c_1, \cdots, c_q) = (s, C')$, 즉$\{(y_i, s_i, C'_i) : 1 \le i \le n\}$는 다음과 같은 $(q+2)$-변량 정규분포의 확률표본으로 생각할 수 있다[6].

$$\begin{pmatrix} y \\ s \\ C \end{pmatrix} \sim N_{q+2}\left[\begin{pmatrix} \mu_y \\ 0 \\ 0_q \end{pmatrix}, \begin{pmatrix} \sigma_y^2 & \sigma_{ys} & \sigma'_{yC} \\ \sigma_{ys} & 1 & 0'_q \\ \sigma_{yC} & 0_q & I_{q\times q} \end{pmatrix}\right] \tag{13}$$

여기서 σ_{ys}는 반응변수 y와 층화변수 s 사이의 공분산이며 σ_{yC}는 y와 통제변수 벡터 C 사이의 공분산 벡터이다. q차원의 통제변수 벡터 C를 관찰하였을 때 반응변수와 층화변수의 조건부 분포는 다음과 같은 2변량 정규분포를 따른다[1].

$$\begin{pmatrix} y|C \\ s|C \end{pmatrix} \sim N_{q+2}\left[\begin{pmatrix} \mu_y + \sigma'_{yC} C \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_y^2 - \sum_{j=1}^{q} \sigma_{yj}^2, & \sigma_{ys} \\ \sigma_{ys}, & 1 \end{pmatrix}\right] \tag{14}$$

단 여기서 σ_{yj}^2은 반응변수 y와 통제변수 c_j의 $Cov(y, c_j)^2$이다. 위의 식으로부터 통제변수 벡터 C가 관찰되었을 때 반응변수의 기대치 μ_y를 추정하기 위한 통제추정량을

$$y(C) = y - \sigma'_{yc} C \tag{15}$$

로 정의하면 $E(y-\sigma'_{yC}C)|C] = \mu_y - \sigma'_{yC}C$ 이며 C에 대한 비조건부 기대치 $E[\mu_y - E[\sigma'_{yC}C]]$에서 $E[C] = 0_q$이므로 $y(C) = (y-\sigma'_{yC}C)$는 μ_y의 불편 추정치이다. 통제추정치 $y(C)$는 분산이 $(\sigma_y^2 - \sum_{j=1}^{q}\sigma_{yj}^2)$으로 단순 추정치의 분산 σ_y^2 보다 적어 단순추정치보다 μ_y의 효율적인 추정치가 된다. 층화변수 s는 통제변수 벡터 C와는 독립적으로 관찰됨으로 분포함수 (14)로부터 통제반응변수와 층화변수,$(y(C),\ s)$의 C에 대한 조건부 결합 확률분포는

$$\begin{pmatrix} y(C)|C \\ s \end{pmatrix} \sim N_{q+2}\left[\begin{pmatrix} \mu_y \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_y^2 - \sum_{j=1}^{q}\sigma_{yj}^2, & \sigma_{ys} \\ \sigma_{ys}, & 1 \end{pmatrix}\right] \tag{16}$$

이다. 이 분포를 기반으로 하여 층화변수 s에 따른 층화 통제추정량을 도출하고자 한다.

3.2 층화 통제추정치

시뮬레이션 런을 통해 수집되는 층화변수는 통제변수와는 독립적으로 관찰됨으로써 층화변수 s에 의하여 반응변수 y를 층화하는 방법은 통제반응변수 $y(C)$의 층화추출 과정에 그대로 적용될 수 있다. 표준정규분포를 따르는 부가변수 s를 층화변수로 사용하여 통제반응변수 $y(C)$를 층화할 때 Sethi의 최적 층화 전략에 의하면 적절한 층의 수는 2 내지 6개이다[10]. 만일 시뮬레이션의 런 수 n이 적지 않으면 시뮬레이션 결과 각 층에 속하는 표본의 수는 10개 이상 나타날 것으로 기대할 수 있다. 이러한 조건이 만족되지 않는다면 층의 수 L을 축소하여 층의 구간 확률 $\pi_h = \Pr[\omega_{h-1} < s_i \le \omega_h]$를 증가시킴으로써 $n\pi_h \ge 10\ (1 \le h \le L)$ 조건을 만족시킬 수 있다.

독립적인 i번째 런으로부터 얻은 반응변수와 층화변수의 쌍 $\{(y_i, s_i) : 1 \le i \le n\ \}$에서 층화변수 s_i가 소속되는 구간 $(\omega_{h-1}, \omega_h]$에 따라 반응변수 $\{y_i : 1 \le i \le n\}$는 h개의 층으로 나누어지며 h층의 하한(ω_{h-1})과 상한(ω_h)은 층의 수 L에 따라 표 1로부터 결정되며 구간 확률 π_h도 결정된다. h층에 속하는 반응변수 y_{hj} $(j=1, \cdots, n_h)$를 이에 대응하여 관찰된 통제변수 $C_{hj}(j=1, \cdots, n_h)$를 사용하여 조정할 때 만일 반응변수와 통제변수 사이의 상관관계 σ_{yC}가 알려져 있다면 통제반응변수는 다음과 같이 정의된다.

$$y_{hj}(C) = y_{hj} - \sigma'_{yC} C_{hj} \tag{17}$$

h층의 반응변수의 $\{y_{hj}(C) : 1 \le j \le n_h\}$의 평균을 $\bar{y}_h(C)$ 이라고 하면 μ_y의 층화 통제추정치 $\bar{y}_{st}(C)$는 다음과 같다[2].

$$\bar{y}_{st}(C) = \sum_{h=1}^{L} \pi_h \bar{y}_h(C) \tag{18}$$

만일 파라미터 σ_{yC}가 알려진 경우가 아니라면 반응변수와 통제변수로부터 σ_{yC}를 추정해야 한다. 식 (17)에서 σ_{yC}의 표본 대응 추정치 S_{yC}를 사용하는 통제반응변수 $y_{hj}(C: S_{yC})$는

$$y_{hj}(C: S_{yC}) = y_{hj} - S'_{yC} C_{hj} \tag{19}$$

이며 $\{y_{hj}(C: S_{yC}) : 1 \le j \le n_h\}$의 평균을 $\bar{y}_h(C: S_{yC})$으로 표시하면 μ_y의 층화통제추정치 $\bar{y}_{st}(C: S_{yC})$는 다음 식과 같다.

$$\bar{y}_{st}(C: S_{yC}) = \sum_{h=1}^{L} \pi_h \bar{y}_h(C: S_{yC}) \tag{20}$$

3.3 층화 통제추정치의 효율성

미지의 파라미터에 대한 추정량은 그 분산이 적을수록 효율적이다. 시뮬레이션에서는 보통 단순 추정량의 분산에 비하여 분산감소기법에 의해 달성한 추정량의 분산감소비율에 의하여 추정량의 효율성을 평가한다. 표준화 통제변수 $c_j (j = 1, 2, \cdots, q)$는 $N(0,1)$을 따르므로 q차원의 통제변수 벡터 C의 기대치는 $E[C] = 0_q$이다. 또한 c_j는 층화변수 s_j와는 독립적으로 실현되는 확률변수 값이므로 층 h와는 상관없이 $E[C_h] = 0_q$이다. 따라서 통제변수의 계수 벡터 σ_{yC}가 알려진 경우 층화 통제변수 추정량의 기대치는

$$E[\bar{y}_{st}(C)] = \sum_{h=1}^{L} \pi_h E[\bar{y}_h - \sigma'_{yC} \bar{C}_h]$$

$$= \sum_{h=1}^{L} \pi_h [E[\bar{y}_h] - \sigma'_{yC} E[\bar{C}_h]] = \sum_{h=1}^{L} \pi_h \mu_{yh} = \mu_y \tag{21}$$

이며 층화 통제추정치는 μ_y에 대한 불편 추정량이다. 그리고 층화 통제추정치 $\bar{y}_{st}(C)$의 분산은 다음과 같다.

$$Var[\bar{y}_{st}(C)] = Var[\sum_{h=1}^{L} \pi_h (\bar{y}_h - \sigma'_{yC} \bar{C}_h)]$$

$$= \sum_{h=1}^{L} \pi_h^2 [Var[\bar{y}_h] - 2\sigma'_{yC} Cov(\bar{y}_h, \bar{C}_h)$$

$$+ \sigma'_{yC} Cov(\bar{C}_h) \sigma_{yC}] = \sum_{h=1}^{L} \pi_h^2 n_h^{-1} [\sigma_{yh}^2 - \sigma'_{yC} \sigma_{yC}]$$

$$= \sum_{h=1}^{L} \pi_h^2 n_h^{-1} [\sigma_{yh}^2 - \sigma'_{yC} \sigma_{yC}]$$

$$= \sum_{h=1}^{L} \pi_h^2 n_h^{-1} [\sigma_{yh}^2 - \sum_{j=1}^{q} \sigma_{yj}^2] \tag{22}$$

식 (11)의 통제추정량의 분산에서 $\sigma'_{yC} \Sigma_C^{-1} \sigma_{yc} = \sum_{j=1}^{q} \sigma_{yj}^2$ 이고 $\Sigma_C^{-1} = I_{q \times q}$이므로 식 (11)을 다시 쓰면

$$Var(\bar{y}(C)) = n^{-2} (\sum_{h=1}^{L} n_h (\sigma_y^2 - n_h \sum_{j=1}^{q} \sigma_{yj}^2)$$

$$= \sum_{h=1}^{L} \pi_h^2 n_h^{-1} (\sigma_y^2 - \sum_{j=1}^{q} \sigma_{yj}^2) \tag{23}$$

이다. 식 (22)의 층화 통제추정치 $\bar{y}_{st}(C)$의 분산을 식 (8)에서의 층화추정량의 분산과 식 (23)에서의 통제추정량의 분산과 각각 비교하면

$$Var(\bar{y}_{st}) - Var[\bar{y}_{st}(C)] = \sum_{h=1}^{L} [\pi_h^2 n_h^{-1} \sum_{j=1}^{q} \sigma_{yj}^2] \tag{24}$$

$$Var(\bar{y}(C)) - Var[\bar{y}_{st}(C)] = \sum_{h=1}^{L} \pi_h^2 n_h^{-1} (\sigma_y^2 - \sum_{h=1}^{L} \sigma_{yh}^2) \tag{25}$$

이다. 위의 결과로부터 통제반응변수의 계수 벡터 σ_{yC}가 알려진 경우에는 층화 통제추정치는 층화추정치에 비하여 효율적이다. 통제추정치와 비교하면 식 (25)에서 $\sum_{h=1}^{L} \sigma_{yh}^2$와 σ_y^2의 크기에 의해 층화통제추정량의 효율성이 결정된다. 반응변수의 전체분산 σ_y^2은 층내분산과 층간분산의 합으로 $\sigma_y^2 \ge \sum_{h=1}^{L} \sigma_{yh}^2$(층내분산) 이므로 식 (25)에서 층화통제추정량의 분산이 통제추정량의 분산에 비해 층간분산 만큼 감소됨을 알 수 있다.

통제반응변수의 계수 벡터 σ_{yC}가 알려져 있지 않는 경우, 식 (19)의 층화통제추정치의 기대치에서 S_{yC}와 $\bar{C}_h$는 서로 독립이므로 $E[S_{yC} \bar{C}_h] = E[S_{yC}] E[\bar{C}_h]$이며 $E[\bar{C}_h] = 0$ 이므로 층화통제추정치는 μ_y의 불편추정량이다. 즉

$$E[\bar{y}_{st}(C: S_{yC})] = E[\sum_{h=1}^{L} \pi_h (\bar{y}_h - S_{yC} \bar{C}_h)]$$

$$= E[\sum_{h=1}^{L}\pi_h\bar{y}_h] - \sum_{h=1}^{L}\pi_h E[S_{yC}\bar{C}_h] = \mu_y \quad (26)$$

σ_{yC}을 모르는 경우 식 (19)에서의 층화통제추정치의 분산은 매우 복잡한 형태를 취하고 있어 식 (22)에서와 같이 직접적으로 그 분산을 유도하는 과정은 쉽지 않다. 이론적으로 추정량의 분산을 비교 하는 대신 본 연구에서는 층화추정량, 통제추정량 및 층화 통제추정량의 표본분산을 서로 비교하여 제안된 기법의 효율성을 평가하고자 한다.

4. 시뮬레이션 실험

4.1 시뮬레이션 모형

그림 1은 집중치료, 심장치료 및 회복치료 시설을 보유하고 있는 특수병원을 이용하는 환자의 시설 이용경로를 보여주고 있는데 Schruben과 Margolin[9]은 이 모델의 시뮬레이션을 통하여 3가지 시설용량이 도착환자의 병원 수용비율에 미치는 영향을 조사하였다. 본 연구에서는 현재의 시설용량에서 수용환자의 병원 입원 평균시간을 추정하고자 한다.

그림 1에서 환자의 병원 도착 간 시간은 평균이 3인 Poisson 분포를 따르며 환자의 75%는 집중치료, 25%는 심장치료를 받고자 한다. 집중치료를 받은 후 27%는 퇴원하고 나머지 73%는 중간 치료를 추가로 받고 퇴원한다. 그리고 심장치료를 받은 환자 중 20%는 퇴원하고 나머지 80%는 중간 치료를 추가로 받고 퇴원한다. 치료 시설이 부족한 경우 도착환자는 병원에 수용될 수 없으며 심장치료나 집중치료를 받고 회복치료가 필요한 환자는 회복치료 시설이 여유가 있을 때까지 심장치료실 또는 회복치료실에 대기한다. 환자의 치료시설 체류시간은 모두 log-normal 분포를 따르며 그 평균과 표준편차는 표 2와 같다.

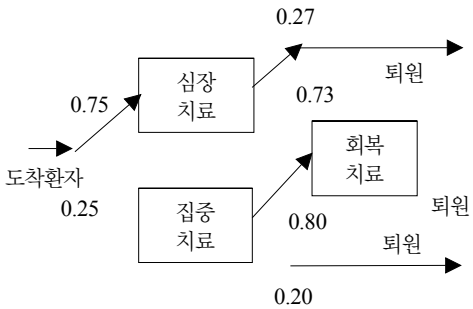

그림 1. 환자의 이동경로

4.2 시뮬레이션 실험결과

그림 1의 시뮬레이션 모형을 AweSim[7]으로 구현한 후 1500 단위시간(일) 동안 시뮬레이션 런을 독립적으로 70회 수행하였다. 각 시뮬레이션 런에서 처음 300 단위시간 동안은 warm-up 기간으로 시뮬레이션 결과를 통계량 수집에서 제외하였다. 시뮬레이션 도중에 반응변수(환자의 입원시간)와 함께 5개의 부가변수를 수집하였다.

3개의 치료시설에서 치료시간은 환자의 병원 전체 입원시간(y)과 상관성이 있으므로 집중 및 심장 치료시설의 진료시간으로부터 2개의 부가변수 (c_1, c_2)을 수집하고 집중치료 및 심장치료 후의 회복시설에서의 치료시간과 관련하여 2개의 부가변수 (c_3, c_4)을 수집하였다. 마지막으로 환자의 병원 도착 간 시간을 부가변수 (s)로 설정하였다. 부가변수는 모두 표준화변수로 정의하였으며 3개의 치료시설의 진료시간은 통제변수로 환자의 도착 간 시간은 층화변수로 사용하였다. 통제변수 (c_3, c_4)는 반응변수 y와 선형상관계수 값은 각각 (-0.128, 0.055)로 (c_1, c_2)와 y의 선형상관계수 값 (0.204, 0.261)보다 적으며 통제추정량은 통제변수로 (c_1, c_2)을 사용하는 경우가 최적이었다.

층화변수 s에 따라 반응변수는 편의상 층의 수(L)를 2와 3으로 하여 층화추정을 실시하였다. 표본 자료에서 모든 층에 속하는 자료의 수가 가장 적은 경우는 층의 수가 3일 때 층 1에 속하는 경우로써 이 때 이 층에 속하는 자료의 수는 15이며 $n\pi_h \geq 10$ $(1 \leq h \leq 3)$ 조건을 만족하였다.

분산감소기법의 효율성은 기법의 분산추정량이 단순추정 방법에 의한 추정량분산에 비해 분산이 감소한 비율로 측정한다[10]. 표 3은 시뮬레이션 결과로부터 얻은 반응변

표 2. 치료시설의 체류시간 및 치료실 수

치료 시설	평균(일)	표준편차(일)	치료실 수(개)
집중	3.4	3.5	15
심장	3.8	1.6	6
회복(집중치료 후)	15.0	7.0	17
회복(심장치료 후)	17.0	3.0	

표 3. 분산감소기법의 효율성

추정방법	$\hat{E}(y)$	추정량 분산	분산감소 비율
단순	4.136	0.280	-
통제	4.151	0.228	19.6%
층화추출($L=2$)	4.126	0.253	9.7%
층화추출($L=3$)	4.130	0.260	7.3%
층화통제($L=2$)	4.132	0.213	24.0%
층화통제($L=3$)	4.130	0.215	23.2%

수 기대치의 단순추정치, 통제추정치, 층화추정치와 층화 통제추정치와 추정치의 표본분산 및 추정기법의 분산의 감소비율을 나타내고 있다.

시뮬레이션 결과에 의하면 층화 통제추정법이 통제변수법이나 층화추출법보다 효율적으로 반응변수를 추정하는 것으로 나타나고 있다. 통제변수기법은 층화추출법에 비해 효율적으로 반응변수를 추정하고 있는데 이는 2개의 통제변수와 반응변수 사이의 표본 상관계수가 20%-26% 정도로 나타나 반응변수와의 관련성이 적지 않은 편이며 층화변수는 반응변수의 층화에 상대적으로 기여한 부분이 적은 것이 그 이유로 판단된다. 층화추출에서 층의 수가 2인 경우가 3인 경우보다 추정 효율성에서 2.4%정도 효율적인 것으로 나타나고 있다.

5. 결 론

본 연구는 시뮬레이션 과정에서 관심 반응변수와 함께 적은 비용으로 수집한 부가정보를 활용하여 반응변수의 변이성을 감소하는 통제변수기법과 층화추출법을 시뮬레이션 실험 설계에 동시에 응용하는 방법을 분석하였다. 두 기법의 적용과정에서 반응변수와 관련성을 가지는 부가변수를 통제변수로 사용하여 반응변수의 변이성을 감소시키고 통제변수와 독립적으로 수집하는 부가변수를 층화변수로 사용하여 반응변수의 변이성을 추가로 감소시키고자 하였다.

반응변수와 통제변수 사이의 상관관계 구조가 알려진 경우에 반응변수 기대치에 대한 층화 통제변수 추정치는 층화 추정치에 비해 통제변수와 반응변수 사이의 상관계수의 제곱에 비례하여 분산이 감소되며 통제추정치와 비교해서는 층화된 층 사이의 분산 즉 층간분산 만큼 제안된 추정량의 분산이 감소되는 것으로 나타났다. 일반적으로 반응변수와 통제변수 사이의 상관관계 구조는 알 수 없으므로 식 (22)에서와 같이 층내 반응변수의 분산으로 구성된 층화통제 추정량의 기대치를 구하여 층화 통제변수 추정치의 효율성을 이론적으로 규명하는 것이 쉽지 않다. 대신 본 연구에서는 시뮬레이션 실험을 통하여 제안된 기법의 효율성을 평가하였다.

표본 반응변수와 통제변수의 표본 공분산을 사용한 시뮬레이션 결과에서는 층화 통제추정에 의한 추정 효율성이 통제변수기법이나 층화추출법에 의한 추정보다 우수한 것으로 나타나고 있다. 간단한 시뮬레이션 모형에 적용한 결과이지만 이러한 결과는 대형 모형의 시뮬레이션에도 적용될 수 있을 것으로 기대된다. 시뮬레이션 도중에 수집한 부가변수 중에서 어떤 변수를 통제변수로 또는 층화변수로 사용하는 것이 효과적인가에 대한 연구는 의미가 있을 것으로 사료되며 또한 반응변수와 통제변수 사이의 상관관계를 추정할 경우에 층화 통제추정량의 이론적인 추정 효율성을 규명하는 부분도 의미가 있을 것으로 판단된다.

참 고 문 헌

1. Anderson, T. W., An Introduction to Multivariates Statistical Analysis, John Wiley & Sons, New York, 1984.
2. Cochran, W.C., Sampling Techniques, John Wiley & Sons, New York, 1977.
3. Law, A.M. and Kelton, W.D., Simulation Modeling and Analysis, 2nd Edition, MacGraw-Hill, New York, 2000.
4. Kleijnen J.P.C., Statistical Techniques in Simulation, Part I. Marcel Decker, Inc., New York, 1974.
5. Kwon, C. and Tew, J.D., "Strategies for Combining Antithetic Variates and Control Varuates in Designed Simulation Experiments," Management Science, vol. 40, no. 8, pp. 1021-1034, 1994.
6. 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. pp. 49-75, 1994.
7. Pritsker, A.A.B. and O'Reilly, J., Simulation with Visual SLAM and AWeSim, John Wiley & Sons, New York, 1999.
8. Sethi, V.K., "A note on Optimum Stratification of populations for estimating population means," Australian Journal of Statistics, vol. 5. pp. 20-33, 1963.
9. Schruben, L.W. and Margolin, B.H., "Pseudo-random Number Assignment in Statistically Designed Simulation and Distribution Sampling Experiments," JASA, vol. 73, pp. 504-525, 1978.
10. Wilson, J.R. and Pritsker, A.A.B., "Experimental Evaluation of Variance Reduction Techniques for Queueing Simulation Using Generalized Concomitant Variables," Management Science, vol. 30, pp. 1459-1472, 1984.

권 치 명 (cmkwon@dau.ac.kr)

1978 서울대학교 산업공학과 졸업
1983 서울대학교 대학원 산업공학과 졸업
1991 VPI & SU 산업공학과 박사
현재 동아대학교 경영정보학과 교수

관심분야 : Simulation Modeling & Output Analysis, Simulation Optimization, FMS

김 성 연 (sykim@dau.ac.kr)

1981 서울대학교 계산통계학과 학사
1983 서울대학교 대학원 통계학 전공(석사)
1997 North Carolina State University(통계학 박사)
현재 동아대학교 경영정보학과 교수

관심분야 : 선형모형, 비선형모형

황 성 원 (sense64@gmail.com)

1987 동아대학교 응용통계학과 학사
1993 경성대학교 대학원 산업정보학과 석사
2006 동아대학교 경영정보학과 박사
현재 동아대학교 경영정보학과 겸임교수

관심분야 : 정보시스템 평가, CRM, 정보보안