부스트래핑(bootstrapping) 기법을 활용한 회귀분석

Bootstrapping Regression Models: A Statistical Simulation

저자 (Authors)

심준섭
Shim Jun Seop

출처 (Source)

정책분석평가학회보 14(2), 2004.6, 167-184(18 pages)
Korean Journal of Policy Analysis and Evaluation 14(2), 2004.6, 167-184(18 pages)

발행처 (Publisher)

한국정책분석평가학회
The Korean Association For Policy Analysis And Evaluation

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE07516554

APA Style

심준섭 (2004). 부스트래핑(bootstrapping) 기법을 활용한 회귀분석. 정책분석평가학회보, 14(2), 167-184

부스트래핑(bootstrapping) 기법을 활용한 회귀분석

심 준 섭(중앙대)

jsshim@cau.ac.kr

본 연구는 표본의 크기가 제한된 상황 하에서 전통적인 OLS 회귀분석에 의한 모수추정치들이 갖는 불안정성의 문제를 극복하기 위한 대안적 통계기법으로 회귀모형 부스트래핑(bootstrapping regression models)의 적용가능성을 Monte Carlo 시뮬레이션을 통하여 평가하였다. 통계적 시뮬레이션을 위해 정해진 모수들을 투입조건으로 가상의 데이터를 구성하고, 회귀모형에 쌍부스트래핑(pairs bootstrapping) 기법을 적용한 후 얻어진 회귀계수들이 어느 정도나 모수의 근사치를 제공하는가를 분석하였다. 연구결과에 따르면 부스트래핑을 활용한 회귀분석은 기존의 단일표본에 기초한 OLS 회귀분석에 비해 R^2와 조정-R^2의 계산에 있어서 상대적으로 우수한 기법으로 확인되었다.

■ 주제어 : 부스트래핑(bootstrapping), 회귀분석(regression analysis), 반복표본추출(resampling), 표준오차(standard error), Monte Carlo 시뮬레이션 simulation)

Ⅰ. 서 론

모수통계(parametric statistics)기법은 모집단의 확률적 분포에 대한 가정을 기반으로 한다. 연구자는 표본의 분포가 모집단의 분포와 동일하다고 가정하고 표본에 대해 통계분석을 하며, 더 나아가 모집단의 분포를 추론한다. 그러나 연구자가 사전(priori)에 모집단의 확률적 분포를 알지 못하는 경우가 많으며, 따라서 이러한 상황에서 연구자는 단지 제한된 표본자료만으로 모수(parameter)들을 추정할 수밖에 없게 된다. 물론 제한된 자료의 불확실성은 경험적인 관찰로부터의 추론에 커다란 제약이 된다. 통계학자들이(Tabachnic & Fidell, 1996; Pedhazur & Schmelkin, 1991) 지적하고 있는 것처럼 대부분의 통계기

법들의 활용도는 표본의 크기가 추정치의 정확성에 미치는 영향에 따라 달라진다. 물론 표본의 크기를 증대시킬수록 안정적인 모수 추정치(parameter estimate)들을 획득할 수 있으나, 시간과 비용 등 여러 가지 제약들로 인해 표본의 크기를 증대시키는 일이 쉽지는 않다. 아마도 연구자로서 겪는 커다란 어려움 가운데 하나는 동시에 많은 사람들로부터 연구에 필요한 자료를 얻어내는 일일 것이다. 설문조사의 경우 극단적으로 낮은 응답율과 그에 따른 자료의 부족으로 인해 연구자가 연구를 중도에 그만두게 되는 경우도 발생된다.

대표적인 모수통계분석기법으로서 최소자승법(OLS: ordinary least square) 회귀분석은 모집단의 분포에 대한 가정을 전제로 하는 모수적통계기법이다. OLS 회귀분석(regression analysis)은 정책과 행정을 비롯한 수많은 연구 분야에서 활용되고 있다. 특히 계량적 정책분석과 정책평가 기법의 하나로서 정책대안의 미래 예측, 정책영향의 평가 등 폭넓게 이용되고 있다.

회귀분석의 광범위한 활용의 배경에는 회귀분석의 가장 큰 강점인 "가정위배에 강력히 저항하는 속성(robust against regression assumptions)"이 자리하고 있다. 일반적으로 회귀분석은 그 가정이(예: 정규성, 등분산성의 가정 등등) 위배되는 상황에도 충분히 강력하여(robust) 최고불편추정치(BLUE: best linear unbiased estimates)는 아니지만 여전히 불편추정치(unbiased estimates)를 산출한다(Berry, 1993; Pedhazur, 1982). 물론 회귀분석의 가정들이 충족되지 못하는 상황 하에서의 모수적 통계에 기초한 추정치, 즉 OLS 공식을 통해 계산된 회귀계수 등 모수 추정치들의 정확성에 대해 의문이 제기되기도 하지만(구체적인 내용은 Lewis-Beck, 1980: 30) 회귀분석의 불편추정치 산출 속성은 계량적 연구(quantitative research)에 있어서 회귀분석을 매우 강력한 분석기법의 하나로 만들기에 충분하다.

이러한 강점에도 불구하고 OLS 회귀분석이 지닌 방법론적 한계 가운데 하나는 회귀분석을 통해 얻어진 계수들이 표본의 크기에 상당히 민감하며, 따라서 안정적인 회귀계수들을 산출하기 위해서는 표본의 크기가 충분히 커야 한다는 점이다.[1] 다수의 통계학자들은 "표본수:변수의 개수" 비율이 적어도 10:1은 되어야 안정적인 회귀계수들을 얻을 수 있다는 것에 동의한다(Nunnally & Bernstein, 1994; Green, 1991; Harris, 1975). 그러나 많은 문항들을 포함하고 있는 설문조사의 낮은 응답률을 고려할 때 10:1의 비율을 확보하는 것이 쉬운 일은 아니다. 또한 이러한 경험적 법칙들(rule-of-thumb)은 회귀분석에 필요한 표본의 크기에 관해 개략적인 가이드라인만을 제시할 뿐, 정확한 기준을 제시하지 못하고 있다.

그렇다면, 제한된 수의 표본에 대한 회귀분석을 통해 얻어진 계수들의 안정도(stability)

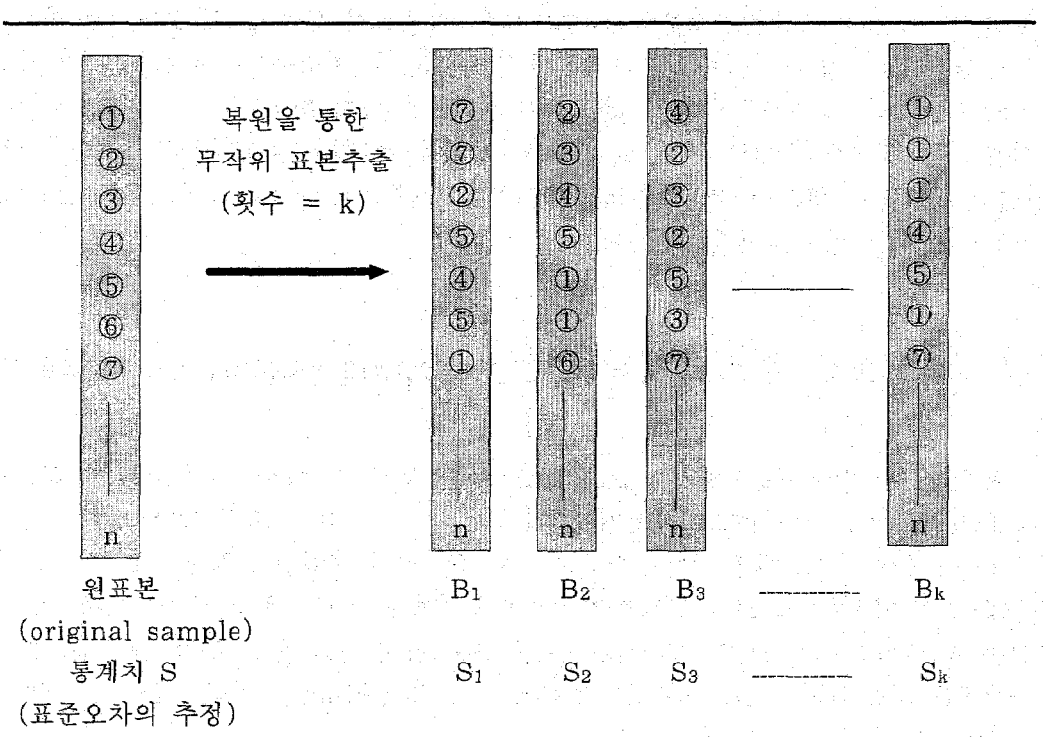
1) Claudy(1972)와 Schmidt(1972)의 통계적 시뮬레이션을 통한 연구결과는 회귀계수의 안정도가 표본의 크기에 얼마나 민감한가를 단적으로 보여주고 있다.

를 추정할 수 있는 대안적인 방법은 없는가? 모집단의 분포에 대한 지식이 없는 경우 회귀계수의 안정도를 측정할 수 있는 비모수적 통계기법은 없는가? 회귀분석은 표본의 크기에 어느 정도 민감한가? 이 연구의 목적은 표본크기가 제한된 상황 하에서의 전통적인 OLS 회귀분석에 의한 모수추정치들이 갖는 불안정성의 문제를 극복하기 위한 대안으로 플러그인(plug-in) 통계기법의 하나인 회귀모형 부스트래핑(bootstrapping regression models) 기법의 활용가능성을 살펴보는 것이다.

Ⅱ. 회귀모형 부스트래핑(bootstrapping regression models)

비모수통계기법인 부스트래핑(bootstrapping)은 1979년 Efron에 의해 처음 소개된 후 표본크기의 한계를 극복하고 모수통계기법의 비현실성을 극복하기 위한 대안으로서 폭넓게 이용되고 있다. 부스트래핑 기법의 핵심은 모집단의 속성을 추론하기 위하여 하나의 표본을 통계적으로 활용하는데 있다. 경험적 연구들을 통하여 Efron은 부스트래핑을 통해 얻어진 모수 추정치들의 시뮬레이션 분포(simulated distribution)는 실제 모수의 분포에 가까운 근사치를 제공한다는 것을 발견하였다(Efron & Tibshirani, 1993; Efron, 1981, 1979).

부스트래핑은 반복표본추출(resampling) 기법의 하나로서 n개의 관측치를 포함하는 표본을 가상의 모집단(virtual population) 으로 활용하는 기법이다. 구체적으로, 이 기법은 모집단으로 부터 추출된 표본을(표본속 관측치의 개수 = n) 가상의 모집단으로 취급하고, 이 가상의 모집단으로부터 복원(replacement)의 과정을 거치면서, 원표본의 크기와 동일한 n개의 관측치 들을 포함하는 표본을 추출해 낸다. 이것이 하나의 부스트래핑 표본(bootstrapping sample)을 구성한다. 중요한 것은 표본추출에 있어 복원의 과정을 거치므로 동일한 관측치 들이 하나의 부스트래핑 표본 속으로 여러 번 추출될 수도 있으며, 극단적인 경우 하나의 부스트래핑 표본이 모두 동일한 관측치로 구성될 수도 있다는 점이다. 다음으로 이 부스트래핑 표본에 대해 연구하고자 하는 통계치를 계산한다. 마지막으로 이상의 과정을 보통 200회 이상 반복함으로써 통계치들이 시뮬레이션 분포(simulated distribution)를 이루게 된다. 그 결과, 평균과 표준오차, 신뢰구간 등을 실제로 시뮬레이션 분포 상에서 계산할 수 있게 된다.



〈그림 1〉 부스트래핑의 과정

〈그림 1〉은 부스트래핑의 과정을 도식화한 것이다. 그림에서 보듯이, 부스트래핑은 n개의 관측치를 포함하고 있는 원표본을 마치 모집단처럼 취급하고, 이 모집단으로부터 복원(replacement)의 과정을 거치면서, n개의 표본을 포함하는 부스트래핑 표본을 k만큼 반복해서 추출해 낸다. 여기서 부스트래핑 표본의 크기(k)는 클수록 안정적인 결과를 산출한다. 그림에서 제시된 것처럼 복원의 원리를 활용함으로 하나의 부스트래핑 표본 속에 동일한 관측치들이 동시에 여러 번 포함될 수 있다.

전통적인 OLS 표준오차 추정공식이 모집단의 분포에 대한 가정을 전제로 하는 모수적통계기법인데 반하여 회귀모형 부스트래핑은 모집단의 분포에 대한 엄격한 가정을 필요로 하지 않는 비모수통계기법이다. 따라서 표본의 크기가 작고 회귀분석의 가정들이 충족되는지 알 수 없는 경우, 공식에 의한 OLS 모수 추정치들의 정확성을 검증하기 위한 대안적 기법으로 회귀모형 부스트래핑이 제시될 수 있다.

회귀모형의 부스트래핑은 k개의 부스트래핑 표본 각각에 대해 회귀분석을 수행하고 이를

통해 얻어진 k개의 회귀계수들을 대상으로 시뮬레이션 분포를 형성한다. 이 분포상에서 회귀계수들의 평균과 표준오차를 실제로 계산하고 더 나아가 신뢰구간을 계산한다. 즉, 각 부스트래핑 표본에 대해 회귀분석을 시행하면 결정계수(R^2), 표준화회귀계수(β), 비표준화회귀계수(b) 등 회귀계수들이 산출된다. 이 과정을 k번 반복함으로써 회귀계수들의 시뮬레이션 분포를 얻게 된다. 계수 값들의 k개 분포 속에서 해당 계수의 표준편차에 대한 실제 계산이 가능해 진다. 그러므로 회귀계수의 정확도를 측정하는 기준인 표준오차(standard error)가 전통적인 OLS 회귀분석처럼 공식을 통해 "추정"되는 것이 아니라 실제 분포로 부터 "계산"되는 것이다. Efron을 비롯한 연구자들은 신뢰할만한 값을 얻기 위해서는 적어도 부스트래핑 표본추출을 200회 이상(즉, $k \geq 200$) 반복하여야 한다고 주장한다(Sprent, 1998; Efron & Tibshirani, 1993; Efron, 1981). 마지막으로, 시뮬레이션 분포상의 회귀계수의 표준편차를 OLS 회귀분석의 공식에 의해 추정된 표준오차와[2] 비교함으로써, OLS 표준오차 추정치의 안정성과 정확성을 평가할 수 있게 된다.

Efron과 Tibshirani(1993)는 회귀모형의 부스트래핑을 위한 두 가지 알고리즘을 제시하였다. 구체적으로 독립변수와 종속변수 값들을 동시에 쌍 (y_i x_i)으로 부스트래핑 하는 "쌍 부스트래핑(pairs bootstrapping)"과 잔차에 대한 부스트래핑을 하는 "잔차 부스트래핑(residuals bootstrapping)"의 두 가지 알고리즘을 개발하였는데, 잔차부스트래핑에 비해 쌍부스트래핑이 기술적으로 쉬우며 부스트래핑 표본의 크기가 커질수록 두 기법의 통계치들이 서로 근접하게 된다(Efron & Tibshirani, 1993). Efron(1981)에 의하면 쌍 부스트래핑은 잔차부스트래핑에 비해 회귀분석의 가정들에 대한 위배에 대해 덜 민감하며 회귀모형의 선정에 적합한 기법이다. 따라서 본 연구에서는 쌍부스트래핑에 분석의 초점을 맞추고자 한다.

n개의 관측치를 포함하는 원표본(original sample)이 주어진 상황 하에서, 회귀분석을 위한 쌍부스트래핑의 구체적인 절차는 다음과 같다.

---

2) 전통적인 OLS 회귀분석에서 비표준화 회귀계수 (b_i)의 표준오차는 아래의 공식으로 부터 추정되는데,

$$\hat{SE}_{b_i} = \frac{sd_{y_s}}{sd_{x_i}} \sqrt{\frac{1 - R^2_{y_s}}{n - k - 1} \times \frac{1}{1 - R^2_{x_i}}}$$

$\hat{SE}_{b_i}$는 회귀계수 bi의 표준오차를 의미하며, n은 표본의 크기, k는 독립변수의 개수를 의미한다. sd_{ys} and sd_{xi} 는 각각 종속변수의 표준편차와 독립변수의 표준편차를 의미한다. R^2_{ys}는 회귀모형의 결정계수이며, $R^2_{x_i}$는 i번째 독립변수를 종속변수로 하고 나머지 독립변수들을 독립변수로 하여 회귀분석을 하고, 여기서 얻어진 결정계수이다. 이 공식에서 보듯이, 독립변수의 개수를 비롯한 다른 모든 값들이 고정된 상태에서 표본의 크기가 커질수록 표준오차는 감소하는 것을 확인할 수 있다.

(1) 원표본(original sample)으로 부터, 복원을 거치면서 표본의 크기와 동일한 n개의 관측치들을 무작위로 추출하여 하나의 부스트래핑 표본을 구성한다.
(2) (1)에서 얻어진 부스트래핑 표본에 대해 회귀분석을 하고, 모수추정치들을 산출한다.
(3) (1)과 (2)의 과정을 독립적으로 200회 이상 반복한다. 그 결과, 각 모수추정치의 가상적인 분포(simulated distribution)를 얻게 된다.
(4) 이 분포로 부터 해당 모수추정치의 표준오차, 즉 표준편차를 계산한다. 이 표준편차는 OLS의 공식에 의해 추정된 표준오차와는 다르게 분포로 부터 계산된 실제 표준오차가 된다.
(5) 분포의 표준오차를 OLS의 추정표준오차와 비교한다.

Ⅲ. 연구설계: 통계적 시뮬레이션(statistical simulation)

회귀모형 부스트래핑 기법의 적용가능성을 검증하기 위해 이 연구는 Monte Carlo 통계적 시뮬레이션을 수행하였다. 본 연구의 특성상 실제 데이터를 활용하여 다양한 조건하에서 실험을 수행하기는 매우 어려웠으며, 따라서 연구방법으로 통계적 시뮬레이션이 선정되었다. 이 연구에서 시뮬레이션의 목적은 정해진 모수들을 투입조건(input condition)으로 가상의 데이터를 제작하고, 이 가상의 데이터에 대해 회귀분석의 쌍부스트래핑 기법을 적용한 후 얻어진 회귀계수들이 실제 모수의 근사치를 산출하는가를 평가하는 것이다. 회귀분석의 쌍부스트래핑 기법을 시뮬레이션 하기 위해 MS Excel 프로그램내의 Visual Basic을 이용하여 독립적인 시뮬레이션 프로그램을 제작하였다. 프로그램에 크게, 독립변수의 생성, 종속변수의 생성, 회귀분석, 부스트래핑의 네 가지 주요 구성요소들을 포함하였다. 연구를 위한 시뮬레이션의 구체적인 과정은 다음과 같다.

제일 먼저 Steinmann(1977)과 그 동료들의 연구에 토대를 두고 아래의 회귀모형이 시뮬레이션을 위한 기본모형(baseline model)으로 선정되었다.

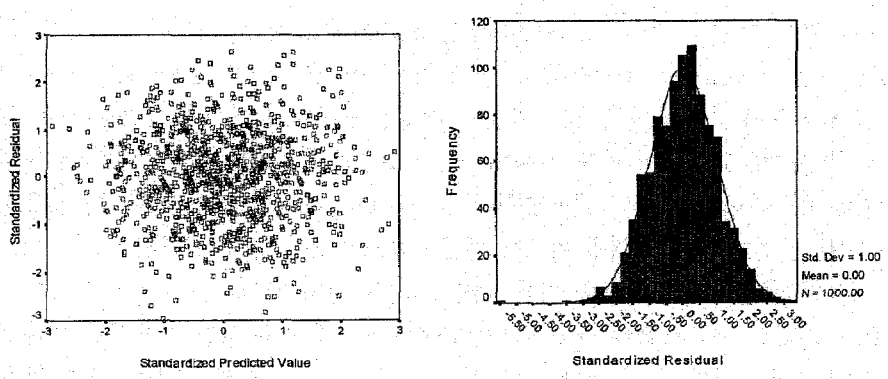
$$Y = 0.22X_1 - 0.49X_2 + 0.24X_3 + 0.13X_4 + 0.31X_5 + 0.60X_6 + 0.17X_7 \quad (R^2 = 0.85)$$

위 회귀모형은 7개의 독립변수와 각 변수의 표준화 회귀계수 및 결정계수(R-square)를 나타내고 있다. 양(+)의 회귀계수와 음(-)의 회귀계수를 모두 포함하고 있으며, 또한 다양한 크기의 계수 값을 포함함으로서 본 연구의 시뮬레이션을 위해 적절한 기본모형을 제공하였다. Steinmann(1977)의 연구와 동일하게 직각설계(orthogonal design) 조건에 따라 독립변수들 간의 상관관계는 "0"이었다.

이러한 기본모형의 조건들을 충족시키는 독립변수 값들을 생성하기 위해 MS Excel 프로그램의 무작위수 발생기(random number generator)를 이용하는 경우의 가장 큰 문제점은 정해진(pre-specified) 상관관계가 존재하는 데이터를 발생시킬 수 없다는 점이다. 이러한 문제점을 극복하고 또한 향후 상관관계를 가정한 연구에 있어서의 활용을 위해 일본 게이요 대학에서 개발한 MS Excel 추가(add-in) 프로그램인 NtRand를 이용하여 무작위로 생성된 7개의 독립변수들이 서로 상관계수가 0이 되도록 데이터를 발생시킬 수 있었다. 각 독립변수는 1에서 10사이의 범위를 갖도록 생성되었으며, 평균은 5 표준편차는 2.5로 모두 동일하게 생성되었다. 시뮬레이션을 위해 1000개의 가상의 관측치들이 생성되었다.

다음으로 독립변수의 값들이 만들어진 상황하에서 종속변수 값들의 생성을 위해 우선 $(0, \sigma^2)$의 정규분포를 이루는 오차(ε)를 무작위로 발생시켰다. 그 후 가중치(ω)를 곱해 오차의 크기를 조절함으로써 $R^2 = 0.85$를 만족시키는 종속변수의 값들을 발생시킬 수 있었다. 구체적으로, 위의 회귀모형에 (1000개의 사례들) 각각의 독립변수의 값들을 대입하여 합계를 구하고, 여기에 "가중치*오차"를 더함으로써 가상적인 1000개의 종속변수 값들이 만들어 질 수 있었다. 이렇게 생성된 종속변수의 값은 1에서 100사이에 분포되었으며, 잔차의 평균은 0이고, 표준편차는 4.5였다$(0, 4.5^2)$.

이러한 과정을 거쳐 독립변수와 종속변수의 값 모두를 포함하는 1000개의 가상적인 사례들을 생성하였으며, 이 사례 들을 시뮬레이션을 위한 가상의 모집단으로 이용하였다. 마지막으로, 부스트래핑 기법의 적용을 위한 표본추출에 앞서 1000개의 관측치들에 대해 회귀분석을 하였고 그 결과가 기본모형에 부합하는가를 검증하였다. 검증결과, 이상에서 제시된 기본모형과 1000개의 시뮬레이션 데이터로 부터 얻어진 회귀모형간에는 상당한 수준의 유사성이 존재하였다. 〈그림 2〉는 시뮬레이션 데이터를 가지고 회귀분석 가정들을 검증한 결과를 보여준다. 데이터가 등분산성(homoscedasticity)과 정규성(normality)을 만족하고 있음을 보여준다.



〈그림 2〉 1000개 데이터의 등분산성과 정규성의 검증

이상에서 만들어진 1000개의 데이터를 모집단으로 가정하고, 다음과 같은 세 가지 다른 표본추출 기법들을 적용하였다:

1) 단순무작위 표본추출 (one random sample): 모집단으로부터 n개의 사례들을 포함하는 한 번의 무작위 표본 추출로 전통적인 표본추출 방식.
2) 반복무작위 표본추출(200 random samples): 모집단으로부터 n개의 사례들을 포함하는 무작위 표본추출의 200회 반복, 즉, 1) 방법의 200회 반복적인 적용.
3) 부스트래핑(200 bootstrapping samples) 200회: 1)에서 추출된 표본을 가상적인 모집단으로 하여 n개의 사례들을 포함하는 표본을 복원을 통해 추출하고 이 과정을 200회 반복.

2)의 반복무작위 표본추출을 통해 얻어진 회귀계수들은 모수추정에 있어 BLUE를 제공한다. 즉, 200개 회귀계수값들의 평균은 모수에 매우 근접하는 추정치를 제공하며, 이들의 표준편차는 모집단 속에서의 모수의 표준편차에 매우 근접한다(Lewis-Beck, 1980). 1)의 단순무작위 표본추출의 경우 회귀계수와 회귀계수의 표준오차가 공식을 통해 추정된 반면, 2)의 반복무작위 표본추출과 3)의 부스트래핑의 경우 200번 독립적으로 회귀계수를 계산하고 이 값들의 분포 상에서 회귀계수의 표준오차 즉 표준편차가 실제로 계산된다.

결국, 이상의 세 가지 표본추출 기법을 비교함으로써, 회귀계수의 표준오차와 신뢰구간을 추정하는 대안적 기법인 회귀모형 부스트래핑의 결과를 전통적인 OLS 회귀분석의 표준오차 추정치와 비교하고, 더 나아가, 2)의 반복표본추출의 결과와도 비교할 수 있었다. 이러한 과정을 통하여 부스트래핑 기법의 정확성에 대한 평가가 가능하였다.

Ⅳ. 시뮬레이션 결과 및 분석

〈표 1〉과 〈표 2〉는 표본의 크기를 달리하면서 세 가지 표본추출기법을 적용한 결과를 보여준다. 표준화회귀계수(standardized regression coefficient)인 β_2 와 β_6의 추정치와 이 추정치들의 표준오차를 보여준다. β_2는 가장 큰 음의 계수 값을 지녔고, β_6는 가장 큰 양의 계수 값을 지녔기에 논의의 편의상 이 두 회귀계수들에 대한 비교를 제시하였다. 각각 200개의 표본을 활용하는 반복무작위 표본추출과 부스트래핑의 경우 추정치의 평균을 나타낸다.

〈표 1〉 β_2와 β_2의 표준오차

(모집단의 β_2 = -0.49)

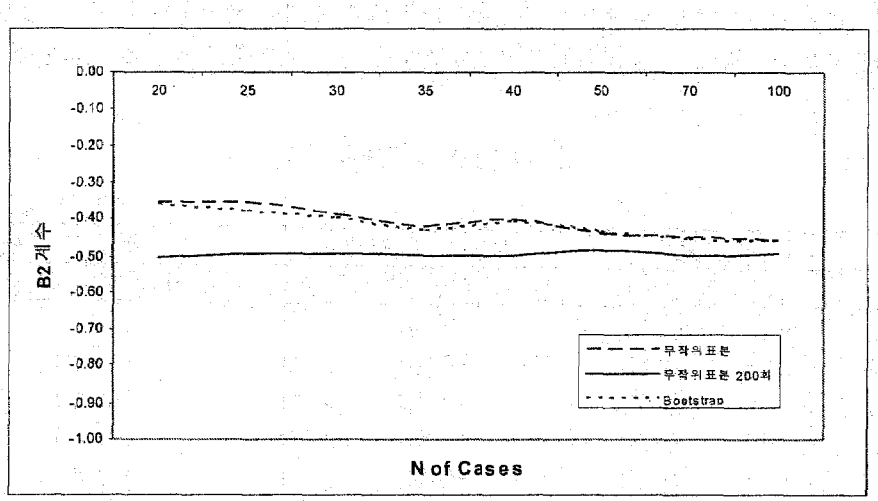
	단일표본		200회 반복표본		부스트래핑	
n	추정치	표준오차	추정치(평균)	표준오차	추정치(평균)	표준오차
20	-0.36	0.11	-0.50	0.11	-0.36	0.18
25	-0.36	0.10	-0.49	0.10	-0.38	0.14
30	-0.39	0.09	-0.49	0.09	-0.40	0.11
35	-0.42	0.08	-0.50	0.09	-0.43	0.10
40	-0.41	0.08	-0.50	0.08	-0.41	0.09
50	-0.44	0.07	-0.48	0.06	-0.44	0.08
70	-0.45	0.05	-0.50	0.06	-0.45	0.06
100	-0.46	0.04	-0.49	0.05	-0.46	0.04

〈표 2〉 β_6와 β_6의 표준오차

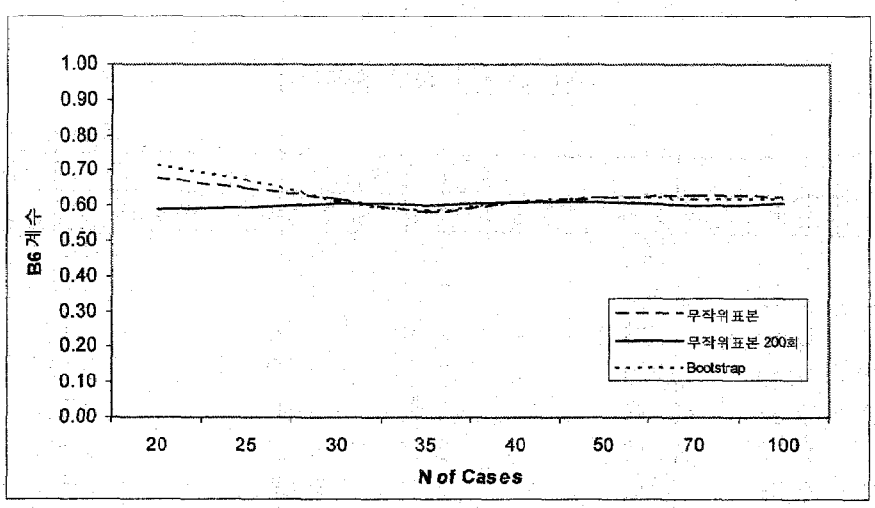
(모집단의 β_6 = 0.60)

	단일표본		200회 반복표본		부스트래핑	
n	추정치	표준오차	추정치(평균)	표준오차	추정치(평균)	표준오차
20	0.68	0.12	0.59	0.12	0.71	0.17
25	0.65	0.10	0.60	0.11	0.67	0.14
30	0.62	0.09	0.60	0.10	0.61	0.11
35	0.58	0.08	0.60	0.09	0.59	0.10
40	0.61	0.08	0.61	0.08	0.61	0.08
50	0.62	0.06	0.61	0.07	0.62	0.06
70	0.69	0.05	0.60	0.06	0.68	0.06
100	0.67	0.04	0.61	0.05	0.67	0.05

〈표 1〉과 〈표 2〉를 그래프로 나타내면 각각 〈그림 3〉과 〈그림 4〉와 같다. 구체적으로 〈그림 3〉과 〈그림 4〉는 각각 세 가지 통계기법을 통해 얻어진 β_2와 β_6의 추정치를 표본의 크기에 따라 비교한 결과를 보여주며, 〈그림 5〉와 〈그림 6〉은 각각 β_2와 β_6 추정치의 표준오차를 비교한 그래프이다. 세 가지 기법들 간의 추세비교의 편의를 위해 완만화(smoothed line) 기능을 적용하였다.



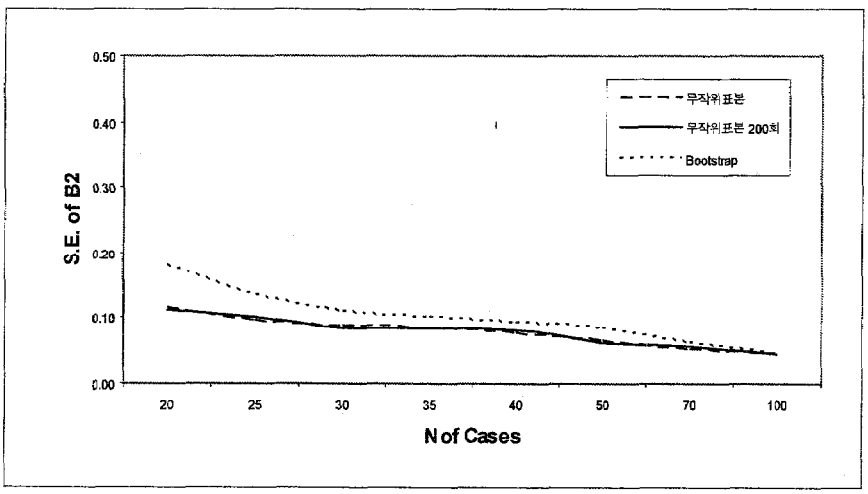
〈그림 3〉 β_2의 비교



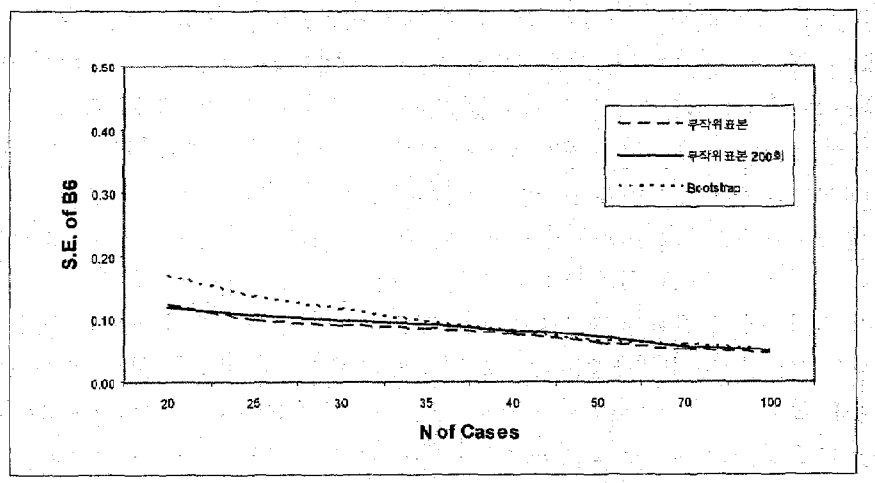
〈그림 4〉 β_6계수의 비교

〈그림 3〉과 〈그림 4〉에서 제시된 것처럼 세 기법 모두 β_2와 β_6 추정에 있어 유사한 결과를 산출하였다. 표본의 크기가 커질수록 모수인 β_2 =-0.49와 β_6 =0.60에 근사치를 산출하였다. 반복표본추출 200회의 결과는 예상했던 것처럼 표본의 크기에 관계없이 BLUE(best linear unbiased estimators) 추정치를 제공하였다. 단순무작위 표본추출과 부스트래핑의 결과 사이에는 실질적인 차이가 존재하지 않았다. 완만화 기능의 적용에도 불구하고 단순무작위 표본추출과 부스트래핑의 결과는 표본의 크기에 따라 부분적인 요동(fluctuation)을 하였다. 이것은 단순무작위 표본추출이 지닌 무작위성(randomness)으로 인해 추출되는 표본에 따라 회귀분석의 결과가 다르게 나타날 수 있음을 보여준다. 더욱 중요한 결과로서, 부스트래핑이 단순무작위 표본추출과 동일한 곡선을 그리는 것은 부스트래핑 기법이 200개의 부스트래핑 표본을 이용하였음에도 불구하고 가상적인 모집단(virtual population)으로 사용하고 있는 원표본(original sample)의 통계적 속성으로 부터 완전히 자유로울 수 없음을 보여주었다.

〈그림 5〉와 〈그림 6〉에서 제시된 것처럼 β_2와 β_6의 표준오차를 추정하거나 계산함에 있어 세 기법들 간에 유사한 결과를 산출하였다. 표본의 크기가 30 이상이 (30 〈 n) 되면서 단순무작위 표본추출과 부스트래핑간의 차이는 매우 작아지고 있음을 알 수 있다. 또한, 표본의 크기가 커질수록 단순무작위 표본추출과 반복표본추출의 결과가 서로 근접하였다. 이것은 OLS 공식에 의한 회귀계수 표준오차의 추정치가 모수의 표준편차를 추정함에 있어 근사치를 제공하기 때문이다.



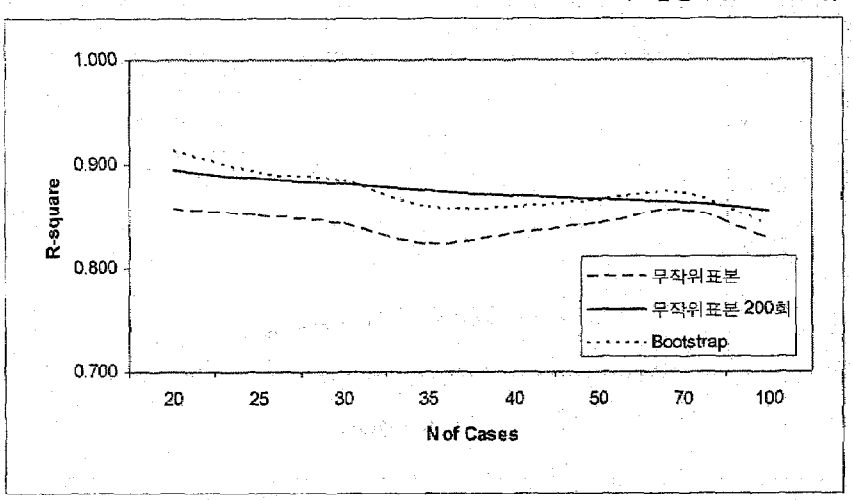
〈그림 5〉 β_2의 표준오차 비교



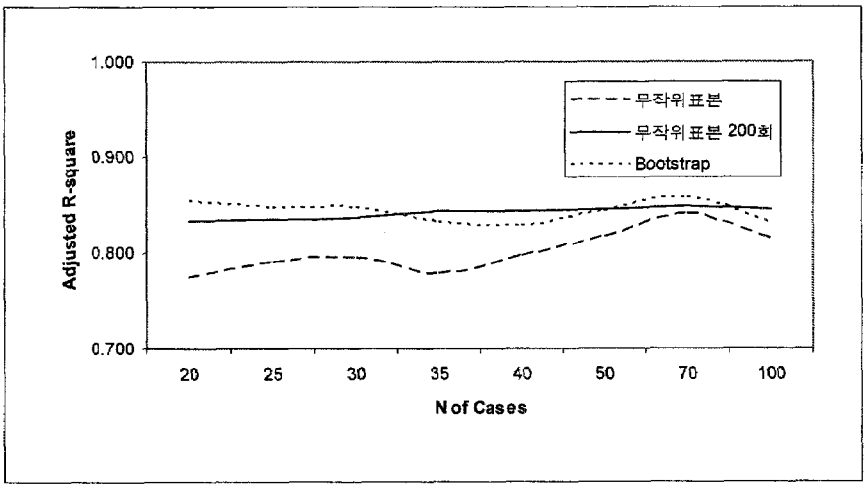
〈그림 6〉 β_6의 표준오차 비교

한편, 표본의 크기가 작은 경우 부스트래핑은 단순무작위 표본추출에 비해 표준오차를 과대추정(overestimate)하는 경향을 보였고, 이것은 "독립변수의 개수: 사례 수"의 비율이 매우 낮은 경우 부스트래핑이 왜곡된 추정치(biased estimates)를 산출할 수 있음을 보여준다(유사한 연구결과는 Freedman & Peters. 1984).

(모집단의 R^2 = 0.85)



〈그림 7〉 R^2의 비교



〈그림 8〉 조정(adjusted) R^2의 비교

〈그림 7〉과 〈그림 8〉은 세 가지 표본추출 기법들 간의 표본크기의 변화에 따른 결정계수(R^2)와 조정결정계수(adjusted R^2)의 변화를 보여준다. 각각 200개의 표본을 활용하는 반복무작위표본추출과 부스트래핑의 경우 결정계수와 조정결정계수의 평균치를 나타낸다.

〈그림 7〉에서 제시된 것처럼, 세 기법 모두 우하향하는 곡선을 산출하였다. 즉, 세 기법 모두 표본의 크기가 커질수록 R^2가 작아지고 모집단의 R^2인 0.85에 접근해갔다. 이것은 여타 조건이 동일할 때 독립변수의 개수가 표본크기에 근접할수록 회귀모형의 적합도(goodness of fit) 지표인 R^2가 과대추정(over-fitting)되는 현상을 반영한다(Pedhazur, 1982).

구체적으로 단일표본에 의한 R^2의 추정은 표본의 무작위성(randomness)으로 인해 정확한 추정치를 제공하지 못하며 또한 표본크기에 따른 변동폭(variability)도 큰 것으로 나타났다. 이것은 OLS 회귀분석을 기반으로 하는 계량적 연구에 있어 회귀모형선정의 기준으로 R^2가 이용되는데 한계가 있음을 보여준다. 특히, 표본의 크기가 작은 경우 이러한 Type I 오류의[3] 가능성은 더욱 커짐을 알 수 있다. 이와는 대조적으로 부스트래핑 기법은 모든 표본 크기에 걸쳐 반복무작위 표본추출과 유사한 결과를 산출함으로써 단순무작위 표본추출에 비해 R^2의 예측오차를 줄이는데 크게 기여하는 것으로 나타났다(유사한 결과는 Ohtani, 2000). 이러한 결과는 모집단의 R^2에 대한 추정에 있어 표본의 크기에 관계없이 회귀모형

3) 이 경우 영가설(null hypothesis)은 "Ho: $R^2 = 0$"이다. 따라서 특히 표본크기가 작은 경우 모델적합도가 순전히 R^2와 adjusted R^2에 의해 측정된다면 과대추정의 문제로 인하여 Type I 오류의 가능성, 즉 올바른 영가설의 기각 가능성이 커지게 된다.

부스트래핑을 통한 R^2 추정치가 단일표본에 의한 R^2의 추정치보다 상대적으로 정확한 값을 제공하고 있음을 보여준다.

〈그림 8〉은 세 기법들 간에 조정결정계수(adjusted R^2)를 비교한 결과를 보여주고 있다. R^2의 경우와 마찬가지로 부스트래핑 기법이 단순무작위표본추출에 비해 조정결정계수의 예측에 따른 오차를 크게 줄이는데 기여하고 있음을 보여준다.

반복무작위 표본추출에 의한 조정 R^2는 표본의 크기에 상관없이 모집단의 $R^2 = 0.85$에 근접한 반면, 단순무작위 표본추출의 경우 무작위성으로 인해 표본에 따라 조정 R^2에 상당한 차이가 존재하였으며, 또한 표본의 크기가 작을수록 그 변동 폭은 컸으며 모든 표본크기에 걸쳐 과소추정(underestimation)의 경향을 보였다. 이것은 조정 R^2가 OLS 회귀분석에 있어 R^2 과대예측(overestimation)의 문제를 조정하기(adjustment) 위해 고안된 것임에도 불구하고, 실제로 완벽한 조정기능을 제공하지는 못하고 있음을 보여준다(유사한 결과는 Ohtani, 1994). 이와는 반대로 부스트래핑의 경우 200개 부스트래핑 표본들로부터 얻어진 조정결정계수의 평균이 반복무작위 표본추출에 의한 추정치에 근접함으로써 단일표본 조정결정계수의 과소추정 문제를 해결하기 위한 하나의 대안으로 이용될 수 있음을 의미한다.

한편 〈그림 7〉과 〈그림 8〉에서 부스트래핑이 단순무작위 표본추출과 유사한 곡선을 그리는 것은 위에서 설명된 것처럼 부스트래핑의 결과가 가상적인 모집단(virtual population)으로 사용하고 있는 원표본(original sample)의 통계적 속성을 상당부분 반영하고 있음을 나타낸다(Sprent, 1998).

V. 결 론

Efron(1979)에 의해 개발된 회귀모형 부스트래핑 기법은 회귀계수와 회귀계수의 표준오차 및 신뢰구간 등의 통계치 들을 계산함에 있어, 전통적인 OLS 회귀분석의 수학적 공식을 이용하는 것이 아니라 표본분포(sampling distribution)상에서 실제로 이러한 통계치들을 계산한다. 회귀모형 부스트래핑은 전통적인 회귀분석과는 달리 회귀분석 가정들에 대한 충족을 전제로 하지 않는 비모수통계기법이다. 따라서 표본의 크기가 작고 모집단의 분포를 알지 못하는 경우에도 회귀계수의 표준오차와 신뢰구간을 산출하는데 적극적으로 이용될 수 있을 것이라는 가정 하에서 회귀모형 부스트래핑 기법의 적용가능성을 평가하였다.

연구결과에 따르면 회귀분석에 있어서 부스트래핑의 적용가능성은 부분적으로 지지되는 것으로 나타났다. 구체적으로 살펴보면 회귀모형 부스트래핑은 기존의 단일표본에 기초한 OLS 회귀분석에 비해 R^2와 조정-R^2의 계산에 있어서 상대적으로 우수한 기법인 것으로 확

인되었다. 그러나 회귀계수(β_1)의 추정과 회귀계수의 표준오차의 추정에 있어서는 기존의 OLS 공식에 의한 추정결과와 커다란 차이가 없는 것으로 나타났다. 이러한 결과는 작은 표본으로부터 얻어진 회귀모형의 적합도 추정의 정확성을 증대시키기 위하여 부스트래핑 기법이 적극적으로 이용될 수 있음을 보여준다.

이러한 연구결과에도 불구하고 이 연구는 부분적으로 다음과 같은 한계를 지니고 있다. 첫째, 회귀분석의 가정들을 충족시키는 데이터가 시뮬레이션을 통해 생성되었고 이 데이터를 모집단으로 활용하였다. 따라서 회귀분석의 가정에 위배되는 조건하에서의 부스트래핑의 정확성에 대한 검증을 할 수 없었다. 부스트래핑이 전통적인 OLS 회귀분석의 대안으로 제시되고 있는 가장 중요한 근거들 중의 하나가 모집단의 분포에 대한 가정을 필요로 하지 않는 것임을 고려할 때, 가정에 위배된 상황 하에서의 실험이 필요할 것이다. 둘째, 다양한 조건들 하에서 시뮬레이션을 통한 좀 더 광범위한 비교분석이 이루어지지 못하였다. 이러한 한계를 극복하고 회귀모형 부스트래핑 기법의 활용가능성을 높이기 위해서는 미래의 연구들이 회귀계수, R^2, 독립변수의 개수, 변수들 간의 상관관계 변화 등 다양한 조건하에서의 실험으로 확대되어야 할 것이다.

【참고문헌】

Berry, W. D. (1993). Understanding regression assumptions (Vol. 92): Sage.

Claudy, J. G. (1972). "A comparison of five variable weighting procedures," *Educational and Psychological Measurement*, 32, 311-322.

Efron, B. (1979). "Bootstrap methods: Another look at the Jackknife," Annals of Statistics, 7, 1-26.

_____. (1981). "Nonparametic estimates of standard error: the jack-knife, the bootstrap, and other method," Biometrika, 68, 589-599.

Efron, B., & Tibshirani, R. (1993). An introduction to the bootstrap. New York: Chapman & Hall.

Freedman, D. A., & Peters, S. C. (1984). "Bootstrapping a regression equation: Some empirical results," *Journal of the American Statistical Association*, 79(385), 97-106.

Green, S. B. (1991). "How many subjects it take to do a regression analysis?," Multivariate Behavioral Research, 26(3), 499-510.

Harris, R. J. (1975). Primer of multivariate statistics. New York: Academic Press.

Lewis-Beck, M. (1980). Applied Regression. (Vol. 22). Beverly Hills: Sage.

Najjar, D., & Iost, B. A. (2003). "The computer-based bootstrap method as a tool to select a relevant surface roughness parameter," Wear, 5-6, 450-460.

Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric Theory. (Third ed.). New York: McGraw-Hill.

Ohtani, K. (2000). "Bootstrapping R-square and adjusted R-square in regression analysis," Economic Modelling, 17, 473-483.

_____. (1994). "The density functions of R-square and adjusted R-square, and their risk performance under asymmetric loss in misspecified linear regression models," Economic Modelling, 11, 463-471.

Pedhazur, E. (1982). Multiple Regression in Behavioral Research. (2nd ed.): Harcourt Brace Jovanovich, Inc.

Pedhazur, E. J., & Schmelkin, L. P. (1991). Measurement, Design, and Analysis. (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Schmidt, F. L. (1972). "The reliability of differences between linear regression weights in applied differential psychology," *Educational and Psychological Measurement*, 32, 879-886.

Sprent, P. (1998). Data driven statistical methods. London: Chapman & Hall.
Steinmann, D. O., Smith, T. H., Jurdem, L. G., & Hammond, K. R. (1977). "Application of social judgment theory in policy formation: an example," *Journal of Applied Behavioral Science*, 13, 69-88.
Tabachnick, B. G., & Fidell, L. S. (1996). Using multivariate statistics: HarperCollins College Publishers.

심 준 섭: 미국 State University of New York at Albany에서 행정학 박사학위("Exploration of Alternative Designs for Judgment Analysis Application in Public Policy Formulation," 2002)를 취득하고, 현재 중앙대학교 공공정책학부 전임강사로 재직 중이다. 주요 연구관심 분야는 의사결정론, 협상론, 정책평가, 조사방법론이며, 주요 논문으로는 "Governors' powers: Conceptual issues and measurement"(2004), "Avoidance of anticipated regret: The ordering of prostate specific antigen tests"(2004), "Why do primary care physicians in the United States and France order prostate specific antigen tests for asymptomatic patients?"(2003), "Does choosing a treatment depend on making a diagnosis? U.S. and French physicians decision making about acute otitis media"(2002) 등이 있다(E-mail : jsshim@cau.ac.kr).

Bootstrapping Regression Models: A Statistical Simulation

Shim, Jun Seop

The purpose of this study is to present the bootstrap technique in the areas of public administration and policy as an alternative method to deal with statistical estimation problems in the conventional OLS regression analysis resulting from a limited number of observations. In particular, pairs bootstrapping that involves choosing random samples in paris from the original data set was applied. To assess and compare the accuracy and stability of the conventional OLS and bootstrap estimates, repeated random sampling that provides precise estimates of the coefficients was also conducted. The results of a Monte Carlo simulation showed that the bootstrap technique provided considerably accurate estimates of the parameters including R^2 and the adjusted R^2. One tentative conclusion can be drawn that the bootstrap technique could be used to solve the overfitting problem of R^2 in the OLS regression analysis.

▌Key words : bootstrapping, multiple regression analysis, resampling, Monte Carlo simulation