The Korean Journal of Applied Statistics (2017) DOI: https://doi.org/10.5351/KJAS.2017.30.3.453
30(3), 453–462

Bootstrap estimation of the standard error of treatment effect with double propensity score adjustment

So Jung Lim[a] · Inkyung Jung[a,1]

[a]Department of Biostatistics and Medical Informatics, Yonsei University College of Medicine

(Received April 5, 2017; Revised May 23, 2017; Accepted May 23, 2017)

Abstract

Double propensity score adjustment is an analytic solution to address bias due to incomplete matching. However, it is difficult to estimate the standard error of the estimated treatment effect when using double propensity score adjustment. In this study, we propose two bootstrap methods to estimate the standard error. The first is a simple bootstrap method that involves drawing bootstrap samples from the matched sample using the propensity score as well as estimating the standard error from the bootstrapped samples. The second is a complex bootstrap method that draws bootstrap samples first from the original sample and then applies the propensity score matching to each bootstrapped sample. We examined the performances of the two methods using simulations under various scenarios. The estimates of standard error using the complex bootstrap were closer to the empirical standard error than those using the simple bootstrap. The simple bootstrap methods tended to underestimate. In addition, the coverage rates of a 95% confidence interval using the complex bootstrap were closer to the advertised rate of 0.95. We applied the two methods to a real data example and found also that the estimate of the standard error using the simple bootstrap was smaller than that using the complex bootstrap.

Keywords: propensity score, matching, observational study, bootstrap, standard error

1. 서론

관찰연구는 임상의학분야에서 흔히 수행되는 연구 형태 중 하나이다. 무작위배정(randomization)을 통하여 두 군 간 연구대상자들의 특성의 균형을 맞출 수 있는 임상시험과는 달리 관찰연구에서는 처리군의 기본 특성이 대조군의 기본 특성과 다른 상황이 존재할 수 있다. 따라서 관찰연구에서는 처리군과 대조군의 효과를 그대로 비교하면 잘못된 결론에 이를 수 있다. 관찰연구에서 생길 수 있는 이와 같은 편의를 줄이기 위하여 사용되는 방법으로 성향점수(propensity score)가 있는데, 이를 활용하여 매칭(matching), 층화(stratification) 또는 회귀분석(regression adjustment)을 할 수 있다 (D'Agostino, 1998). 성향점수란 관찰된 공변량이 주어졌을 때, 처리군에 배정될 조건부 확률로 정의되고 (Rosenbaum과 Rubin, 1983), 공변량의 정보만이 주어진 경우 처리군에 배정되었을 가능성의 척도로 생각할 수 있다 (D'Agostino, 1998).

[1]Corresponding author: Department of Biostatistics and Medical Informatics, Yonsei University College of Medicine, 50-1 Yonsei-ro, Seodaemun-gu, Seoul 03722, Korea. E-mail: ijung@yuhs.ac

성향점수 매칭(propensity score matching) 시 처리군과 성향점수가 유사한 대조군을 선정하는 경우, 처리군의 일부는 성향점수가 유사한 대조군을 찾지 못해 분석에서 제외되는 경우가 생긴다. 이로 인해 발생하는 편의가 Rosenbaum과 Rubin (1985)이 언급한 '불완전 매칭(incomplete matching)에 기인한 편의'이다. 최근에 Austin (2017)의 연구에서 이중 성향점수 보정(double propensity score adjustment) 방법이 위 문제의 분석적 해결방안으로 제시된 바 있다. 그러나 이중 성향점수 보정 방법을 이용한 처리효과 추정치의 표준오차는 이론적 추정치가 제시되지 않아 표준오차의 추정과 처리효과에 대한 추론에 어려움이 존재한다.

본 연구에서는 이중 성향점수 보정 방법을 이용한 처리효과 추정치의 표준오차 추정을 위한 방법으로 두 가지 붓스트랩 방법을 제안하고, 제안하는 두 가지 방법으로 추정된 표준오차의 정확도를 비교하고자 한다. 두 붓스트랩 방법으로는 Austin과 Small (2014)이 비복원추출로 성향점수 매칭을 하는 경우에 있어 처리효과 추정치의 표준오차 추정 방법으로 제시한 두 가지 방법을 활용하였다. 첫 번째 방법은 단순(simple) 붓스트랩으로 원자료를 이용하여 성향점수 매칭 후 매칭된 표본으로부터 붓스트랩 표본을 얻는 것이다. 두 번째 방법은 복합(complex) 붓스트랩으로 원자료에서 붓스트랩 표본을 먼저 생성하고 각 붓스트랩 표본에서 성향점수 매칭을 하는 것이다. 두 방법의 성능을 비교하기 위하여 다양한 상황을 가정하여 모의실험을 실시하였다. 연속형과 이분형 두 종류의 결과변수에 대하여 이중 성향점수 보정 방법을 이용해 처리효과를 추정했을 때 두 가지 붓스트랩 방법이 표준오차를 얼마나 정확히 추정하는지를 경험적 표준오차와의 차이와 95% 신뢰구간에 대한 포함확률(coverage probability)을 이용하여 비교하였다. 또한 유방암 환자의 유방초음파 검사 자료에 두 붓스트랩을 적용하여 처리효과 추정치의 표준오차를 추정 결과를 비교해 본다.

2. 성향점수 매칭과 이중 성향점수

2.1. 성향점수 매칭(propensity score matching)과 처리효과

성향점수 매칭은 유사한 성향점수를 가지는 각 군의 개체로 이루어진 짝들의 집합을 형성하는 것이다. Rosenbaum과 Rubin (1983)은 유사한 성향점수를 가진 개체들은 측정된 공변량 값들의 분포가 유사할 것이라고 언급했다. 이는 관찰연구에서 처리효과 추정 시 성향점수 매칭을 통해 무작위배정 실험에서와 유사하도록 가능한 한 편의를 줄일 수 있음을 의미한다. Rosenbaum과 Rubin (1983)이 제안한 성향점수는 관찰된 공변량들이 주어졌을 때, 해당 공변량 값들을 가지는 각 개체가 처리군에 배정될 조건부 확률로써 정의되며 식은 아래와 같다.

$$e(\boldsymbol{x}) = \Pr(z = 1|\boldsymbol{x}),$$

여기서 z는 처리군일 때는 1이고 대조군일 때는 0을, $\boldsymbol{x}$는 공변량들을 나타낸다. 성향점수 매칭 알고리즘에는 매칭된 쌍의 성향점수 차이의 평균을 최소화하는 optimal matching, 처리군의 개체와 가장 가까운 성향점수를 가지는 대조군의 개체를 매칭하는 nearest neighbor matching (NNM), 성향점수의 차이가 명시된 범위(specified caliper)보다 적은 경우에만 매칭 쌍이 되게 하는 caliper matching 등이 있고, 대개 비복원추출을 하여 대조군의 각 개체는 매칭 표본에 한번만 포함된다. Caliper matching을 하면 편의는 작아지나 분산이 커질 수 있고 matching되지 않는 처리군의 개체로 인해 표본수가 줄어들 수 있다. 반면, optimal matching과 NNM을 이용하면 처리군의 모든 개체를 매칭할 수 있지만, 매칭의 질을 통제하지 못하므로 편의가 커질 수 있다.

대조군과 처리군의 두 군이 존재하는 경우, 각 개체는 각 군에 대한 두 개의 잠재적 결과변수(potential outcomes)를 가진다. 그러나 각 개체는 처리군 혹은 대조군 중 하나의 군에만 속할 수 있으므로 특정

개체에서 오직 하나의 결과변수만이 관측된다. r_1과 r_0는 각각 처리를 받았을 때와 받지 않았을 때의 잠재적 결과변수를 나타낸다. $E(r_1 - r_0)$는 두 잠재적 결과변수들의 차이의 평균인 전체 자료 혹은 모집단(average effect of treatment in an entire sample or population; ATE)에서의 처리효과로 정의된다. 이와 관련된 처리효과의 측도로 처리군에서의 처리효과(average treatment effect for the treated; ATT)는 아래와 같다.

$$\text{ATT} = \tau_1 = E(r_1 - r_0|z = 1). \tag{2.1}$$

2.2. 이중 성향점수 보정(double propensity score adjustment)

이중 성향점수 보정 방법은 '불완전 매칭에 기인한 편의'를 줄이고자 완전 매칭을 하여 ATT를 추정하는 방법이다 (Austin, 2017). 이때 소수의 처리군 모두와 다수의 대조군 중 일부를 optimal matching 또는 NNM을 이용해 완전 매칭한다. 완전 매칭을 하는 경우 상이한 성향점수를 가진 짝이 매칭됨으로 인하여 ATT 추정치에 편의가 생길 수 있다. 이러한 편의를 줄이기 위하여 성향점수를 이용해 결과변수에 추가적인 보정을 가하는 방법을 이중 성향점수 보정 방법이라고 한다 (Austin, 2017).

먼저 매칭된 자료에서 대조군의 성향점수를 공변량으로 하여 결과변수를 예측하는 회귀모형인 $m_0(e(\boldsymbol{x}))$를 추정한다. $m_0(e(\boldsymbol{x}))$에 대한 식은 아래와 같다.

$$m_0(e(\boldsymbol{x})) = E(r_0|z = 0, e(\boldsymbol{x})). \tag{2.2}$$

여기서 $m_0(e(\boldsymbol{x}))$는 결과변수의 종류에 따라 선형 회귀 모형, 로지스틱 회귀 모형 등 다양한 모형을 적용시킬 수 있다. 추정된 (2.2)의 모형을 처리군에 적용하여 처리군에서의 처리를 받지 않았을 때의 잠재적 결과변수 r_0에 대한 추정치를 얻는다. 식 (2.1)의 ATT는 처리군에서 관측된 r_1과 추정된 r_0의 차이의 평균으로 추정되며 식은 아래와 같다.

$$\widehat{\text{ATT}} = \frac{1}{N}\sum_{i=1}^{N}(r_{1,i} - \hat{r}_{0,i}) = \frac{1}{N}\sum_{i=1}^{N}\left(r_{1,i} - \widehat{m_0(e(\boldsymbol{x_i}))}\right) \tag{2.3}$$

여기서, $e(\boldsymbol{x_i})$는 처리군 N명 중 i번째 개체의 성향점수 추정치이며 $r_{1,i}$는 처리를 받았을 때의 잠재적 결과변수 r_1의 관측 값이다. $\hat{r}_{0,i} = \widehat{m_0(e(\boldsymbol{x}_i))}$는 처리군의 i번째 개체가 만약 처리를 받지 않았을 때의 잠재적 결과변수 r_0의 추정치이다. 결과변수가 연속형일 경우에는 선형회귀모형으로 식 (2.2)를 추정하여 $\hat{r}_{0,i}$가 연속형이 되고, 결과변수가 이분형일 경우에는 로지스틱 회귀모형으로 식 (2.2)를 추정하여 $\hat{r}_{0,i}$가 확률이 된다.

2.3. 표준오차 추정을 위한 붓스트랩 방법

Simple 붓스트랩은 매칭된 자료에서 전형적으로 적용되는 방법으로 원자료가 아닌 매칭된 짝들을 붓스트랩하는 방법이다. 매칭된 짝들로 이루어진 집합이 짝 M_i, $i = 1, \ldots, N$로 이루어져 있을 경우 B개의 붓스트랩 표본들은 N개의 짝들의 집합 $A = \{M_1, M_2, \ldots, M_N\}$로부터 반복을 허용하여 추출된다. 따라서 각 붓스트랩 표본은 N개의 짝들로 구성된다. B개의 붓스트랩 표본들의 처리효과 추정치들의 표준편차는 원자료로부터 구해진 처리효과 추정치의 표준오차 추정치로 사용된다.

Complex 붓스트랩은 매칭 전의 원자료를 붓스트랩하는 방법이다. B개의 붓스트랩 표본들은 원자료로부터 반복을 허용하여 추출된다. 따라서 원자료가 M개의 개체들로 구성된 경우 각 붓스트랩 표본은 M개의 개체들로 구성된다. B개의 붓스트랩 표본으로부터 각각 성향점수 모형이 추정되고 매칭이 이루어진다. 이 B개의 매칭된 표본들의 처리효과 추정치들의 표준편차는 원자료로부터 구해진 처리효과 추

정치의 표준오차 추정치로 사용된다. Complex 붓스트랩은 simple 붓스트랩과 비교할 때 두 가지 추가적인 변동을 고려하는데, 이는 성향점수 모형(propensity score model; PSM)을 추정하는 변동과 매칭된 표본을 형성하는 변동이다. 또한 simple 붓스트랩은 원자료로부터 단 한 번의 성향점수 매칭을 시행하는 반면 complex 붓스트랩은 총 B번의 성향점수 매칭을 시행한다. 따라서 complex 붓스트랩의 경우 훨씬 더 강도 높은 컴퓨터 계산 과정이 필요하다.

3. 모의실험

3.1. 모의실험 설계

모의실험 설계 시 고려해야 하는 결과변수, 처리변수, 공변량 등은 Austin (2017)의 논문과 Austin과 Small (2014)의 논문을 참고하여 설정하였다. 결과변수는 연속형, 이분형으로 두 경우를 가정하고 처리변수와 결과변수에 영향을 미치는 여부를 고려하여 10개의 공변량들($x_1, \ldots, x_{10}$)을 생성하였다. 10개의 공변량들 중 7개($x_1, \ldots, x_7$)는 처리변수에 영향을 주고, 7개($x_4, \ldots, x_{10}$)는 결과변수에 영향을 준다고 가정하였다. 공변량들 간의 상관성을 고려할 수 있도록 동일한 특성 범주 안에 포함되는 공변량들 간의 상관계수를 다양하게 설정하여 다음과 같이 다변량 정규분포로부터 자료를 생성하였다.

$$x_T = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \sim \text{MVN}\left(0, \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}\right),$$

$$x_{TY} = \begin{pmatrix} x_4 \\ x_5 \\ x_6 \\ x_7 \end{pmatrix} \sim \text{MVN}\left(0, \begin{pmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{pmatrix}\right),$$

$$x_Y = \begin{pmatrix} x_8 \\ x_9 \\ x_{10} \end{pmatrix} \sim \text{MVN}\left(0, \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}\right), \qquad \rho = 0, 0.3, 0.5, 0.8.$$

처리변수 생성 시 공변량들의 계수를 다르게 하여 처리변수에 영향을 주는 강도를 설정하며, 영향을 주지 않는 공변량들은 계수값을 0으로 하여 식에서 제외시킨다. 처리변수 z_i의 생성은 다음과 같이 한다.

$$\begin{aligned} \text{logit}(p_{i,\text{treat}}) &= \beta_{0,\text{treat}} + \beta_W x_{1,i} + \beta_M x_{2,i} + \beta_S x_{3,i} + \beta_W x_{4,i} + \beta_M x_{5,i} + \beta_S x_{6,i} + \beta_{VS} x_{7,i}, \\ z_i &\sim \text{Bernoulli}(p_{i,\text{treat}}). \end{aligned}$$

위의 식에서 절편($\beta_{0,\text{treat}}$)은 모의실험 자료에서 처리군으로 배정될 개체들의 비율(prevalence)을 결정하게 되는데, 이번 연구에서는 그 비율을 0.05과 0.1이 되도록 설정하였다. 회귀계수 $\beta_W, \beta_M, \beta_S, \beta_{VS}$는 각각 $\log(1.25), \log(1.5), \log(1.75), \log(2)$로 설정하였다. 이는 각 회귀계수에 해당하는 공변량이 약한, 중간, 강한, 매우강한 정도로 처리변수에 영향을 주는 것을 뜻한다.

결과변수의 생성은 결과변수에 영향을 준다고 가정한 7개의 공변량($x_4, \ldots, x_{10}$)으로부터 생성한다. 연속형 결과변수인 경우 선형회귀모형을 사용하여 다음과 같은 식을 이용한다.

$$\begin{aligned} r_{z,i,\text{continuous}} = \beta_{\text{treat,continuous}} z_i &+ \beta_W x_{4,i} + \beta_M x_{5,i} + \beta_S x_{6,i} + \beta_{VS} x_{7,i} + \beta_W x_{8,i} + \beta_M x_{9,i} \\ &+ \beta_S x_{10,i} + \epsilon_i, \quad \epsilon_i \sim N\left(0, \sigma^2\right), \ \sigma = 3. \end{aligned}$$

위 식에서 회귀계수 $\beta_{\text{treat,continuous}}$은 처리군이 처리를 받았을 때와($z_i = 1$) 받지 않았을 때($z_i = 0$)의 결과변수 차이의 평균을 결정한다. 결과변수가 연속형인 경우 z_i와 다른 공변량 간의 교호작용이 존재하지 않는다면, 즉, 두 군에서 처리효과가 다르지 않는다면 $\beta_{\text{treat,continuous}}$는 곧 처리효과를 의미한다. 이번 연구에서는 ATT를 다음과 같이 설정하였다.

$$\text{ATT} = E(r_1 - r_0 | z = 1) = \beta_{\text{treat,continuous}} = 1.$$

결과변수가 이분형인 경우 로지스틱회귀모형을 사용하며 식은 아래와 같다.

$$\begin{aligned} \text{logit}(p_{i,\text{outcome}}) &= \beta_{0,\text{outcome}} + \beta_{\text{treat,binary}} z_i + \beta_W x_{4,i} + \beta_M x_{5,i} + \beta_S x_{6,i} + \beta_{VS} x_{7,i}, \\ &\quad + \beta_W x_{8,i} + \beta_M x_{9,i} + \beta_S x_{10,i}, \\ r_{z,i,\text{binary}} &\sim \text{Bernoulli}(p_{i,\text{outcome}}). \end{aligned}$$

위 식에서 절편($\beta_{0,\text{outcome}}$)은 만약 모든 개체가 처리를 받지 않을 경우의 주변 확률을 결정한다. 이번 연구에서는 주변 확률을 0.1로 설정하였다. 회귀계수 $\beta_{\text{treat,binary}}$는 처리군이 처리를 받았을 때와($z_i = 1$) 받지 않았을 때($z_i = 0$)의 결과변수 차이의 평균을 결정한다. 이번 연구에서는 실제 ATT는 다음과 같이 설정하였다.

$$\text{ATT} = E[\Pr(r_1 = 1) - \Pr(r_0 = 1) | z = 1] = 0.02.$$

성향점수 추정을 위한 PSM으로 로지스틱회귀모형을 사용하였다. 모형에 포함된 공변량들의 특성에 따라 조합을 달리하여 세 모형을 고려하였다. PSM1은 모든 변수($x_1, \ldots, x_{10}$)를 포함한 모형이다. PSM2는 결과변수에 영향을 주는 변수들($x_4, \ldots, x_{10}$)을 포함한 모형이다. PSM2가 처리할당에만 영향을 주는 변수들도 모형에 포함한 경우인 PSM1과 비교할 때 더 좋은 추정치를 제공한다는 연구결과가 존재한다 (Austin 등, 2007). PSM3는 회귀분석을 시행하여 결과변수에 유의한 영향을 미치는 것으로 확인된 S개의 유의한 변수($x_1, \ldots, x_S$)들을 포함한 모형이다.

생성된 표본의 각 개체로부터 NNM을 이용하여 성향점수 매칭을 한 후 이중 성향점수를 이용한 처리효과 추정방법을 이용하여 ATT를 추정하였다. 처리효과 추정치의 표준오차는 2.3절에서 설명한 두 가지 붓스트랩 방법으로 추정하였다. 3,000개의 개체로 이루어진 1,000개의 표본을 생성하였고, 표준오차 추정 시 100번의 붓스트랩을 시행하였다. 1,000개의 표본에서 직접 추정된 처리효과들의 표준편차를 경험적 표준오차로 정의하고 두 가지 붓스트랩 방법으로 추정된 표준오차 추정치와 비교하였다. 추가로 왈드 타입의 95% 신뢰구간을 구하여 포함확률을 비교하였다.

3.2. 모의실험 결과

연속형 결과변수에 대하여 처리군에 배정된 개체들의 비율을 0.05, 0.1로 가정하였을 경우 이중 성향점수 보정 방법을 이용하여 처리효과를 추정하고 그에 대한 표준오차를 두 가지 붓스트랩으로 추정한 모의실험의 결과는 Tables 3.1–3.2에 있다. 이분형 결과변수에 대한 결과는 Tables 3.3–3.4에 나타내었다. 모든 경우에서 complex 붓스트랩을 사용한 추정치가 경험적 표준오차와 매우 유사한 것을 확인 할 수 있었다. 반면 simple 붓스트랩을 사용한 추정치는 경험적 표준오차와 매우 상이하며 과소추정 되는 것을 확인 할 수 있었다. 연속형 결과변수의 경우에는 complex 붓스트랩으로 추정된 표준오차가 경험적 표준오차에 비해 다소 컸으나, 이분형 결과변수의 경우에는 다소 작았다. 공변량들 간의 상관성이나 PSM이 달라지는 경우에는 결과의 차이가 크지 않았다. 95% 신뢰구간의 포함확률을 또한 complex 붓스트랩을 사용한 경우에는 0.95에 매우 가까웠으나, simple 붓스트랩을 사용한 경우에는 0.95보다 훨씬 작은 0.85에 가까웠다.

Table 3.1. Estimated standard errors of average treatment effect for the treated (ATT) estimate for continuous outcome using double propensity score adjustment (prevalence = 0.05)

Continuous outcome : difference in means						
Correlation	Model	Estimated standard error			Coverage probability (Wald type)	
		Empirical	Simple	Complex	Simple	Complex
0	PSM 1	0.3591	0.2582	0.3774	0.844	0.965
	PSM 2	0.3509	0.2546	0.3760	0.862	0.956
	PSM 3	0.3481	0.2548	0.3774	0.864	0.957
0.3	PSM 1	0.3623	0.2653	0.3799	0.842	0.957
	PSM 2	0.3550	0.2592	0.3772	0.847	0.959
	PSM 3	0.3615	0.2596	0.3787	0.831	0.960
0.5	PSM 1	0.3793	0.2701	0.3826	0.836	0.957
	PSM 2	0.3603	0.2633	0.3782	0.838	0.959
	PSM 3	0.3680	0.2639	0.3832	0.833	0.949
0.8	PSM 1	0.3776	0.2753	0.3848	0.846	0.956
	PSM 2	0.3444	0.2662	0.3802	0.865	0.962
	PSM 3	0.3531	0.2673	0.3882	0.849	0.964

Table 3.2. Estimated standard errors of average treatment effect for the treated (ATT) estimate for continuous outcome using double propensity score adjustment (prevalence = 0.1)

Continuous outcome : difference in means						
Correlation	Model	Estimated standard error			Coverage probability (Wald type)	
		Empirical	Simple	Complex	Simple	Complex
0	PSM 1	0.2418	0.1819	0.2629	0.860	0.963
	PSM 2	0.2412	0.1794	0.2621	0.853	0.965
	PSM 3	0.2401	0.1796	0.2629	0.861	0.967
0.3	PSM 1	0.2518	0.1872	0.2647	0.866	0.958
	PSM 2	0.2405	0.1827	0.2626	0.856	0.968
	PSM 3	0.2447	0.1832	0.2653	0.851	0.965
0.5	PSM 1	0.2533	0.1894	0.2659	0.837	0.952
	PSM 2	0.2473	0.1840	0.2626	0.834	0.962
	PSM 3	0.2480	0.1846	0.2668	0.837	0.962
0.8	PSM 1	0.2601	0.1937	0.2697	0.846	0.951
	PSM 2	0.2469	0.1867	0.2635	0.848	0.962
	PSM 3	0.2430	0.1874	0.2717	0.858	0.970

4. 실제자료 분석

본 자료는 2011년 1월 1일부터 2012년 12월 31일까지 2년 동안 대학병원의 영상의학과에서 2011년 1월 1일 이전에 유방암을 진단받아 수술을 받은 환자 3060명을 대상으로 유방초음파 검사를 진행하여 이차성 유방암에 대한 병변 검출여부를 기록한 자료이다. 이 자료를 통해 유방초음파 검사 빈도가 높은 biannual군(처음 검사부터 평균 검사 주기가 1년 미만인 환자군)과 검사 빈도가 낮은 annual군(검사 주기가 약 1년인 환자군)에서 이차성 유방암에 대한 병변이 검출 될 확률이 차이가 있는지 알아보고자 한다. 결과변수는 이분형으로 유방초음파 검사를 통해 2년간 한번이라도 병변이 검출되었다면 '1'로 그렇

Table 3.3. Estimated standard errors of average treatment effect for the treated (ATT) estimate for binary outcome using double propensity score adjustment (prevalence = 0.05)

Binary outcome : risk difference						
Correlation	Model	Estimated standard error			Coverage probability (Wald type)	
		Empirical	Simple	Complex	Simple	Complex
0	PSM 1	0.0495	0.0320	0.0458	0.849	0.950
	PSM 2	0.0489	0.0314	0.0454	0.837	0.959
	PSM 3	0.0486	0.0314	0.0454	0.842	0.960
0.3	PSM 1	0.0479	0.0341	0.0479	0.866	0.960
	PSM 2	0.0462	0.0330	0.0471	0.866	0.970
	PSM 3	0.0461	0.0330	0.0475	0.866	0.965
0.5	PSM 1	0.0488	0.0350	0.0485	0.852	0.958
	PSM 2	0.0477	0.0336	0.0476	0.830	0.960
	PSM 3	0.0474	0.0337	0.0484	0.841	0.960
0.8	PSM 1	0.0472	0.0362	0.0494	0.863	0.962
	PSM 2	0.0473	0.0345	0.0479	0.843	0.959
	PSM 3	0.0465	0.0347	0.0494	0.860	0.975

Table 3.4. Estimated standard errors of average treatment effect for the treated (ATT) estimate for binary outcome using double propensity score adjustment (prevalence = 0.1)

Binary outcome : risk difference						
Correlation	Model	Estimated standard error			Coverage probability (Wald type)	
		Empirical	Simple	Complex	Simple	Complex
0	PSM 1	0.0365	0.0221	0.0313	0.846	0.947
	PSM 2	0.0362	0.0216	0.0310	0.824	0.955
	PSM 3	0.0354	0.0217	0.0310	0.833	0.959
0.3	PSM 1	0.0351	0.0235	0.0329	0.829	0.949
	PSM 2	0.0346	0.0227	0.0320	0.847	0.954
	PSM 3	0.0347	0.0228	0.0325	0.834	0.954
0.5	PSM 1	0.0354	0.0240	0.0340	0.822	0.947
	PSM 2	0.0337	0.0230	0.0323	0.843	0.946
	PSM 3	0.0344	0.0231	0.0333	0.836	0.949
0.8	PSM 1	0.0351	0.0246	0.0354	0.840	0.949
	PSM 2	0.0321	0.0234	0.0325	0.872	0.966
	PSM 3	0.0322	0.0236	0.0344	0.860	0.967

지 않으면 '0'으로 기록하였다. 독립변수로는 임상적으로 중요한 환자의 나이(Age), 수술로부터 유방초음파 검사까지의 기간(Time interval)과 수술 방법(Op method)을 고려하였다. Table 4.1에 자료의 기술통계량을 두 군으로 나누어 나타내었다. Age와 Time interval은 two-sample t-test로 Op method는 chi-square test로 비교하였는데, 두 군 간 매우 유의한 차이가 나는 것을 알 수 있다.

세 변수로 성향점수 모형을 추정하고 모의실험에서와 마찬가지로 NNM 방법을 사용하여 성향점수 매칭을 하였다. 매칭된 자료의 기술통계량은 Table 4.2에 있는데, Age와 Op method는 두 군간 유의한 차이가 없어졌지만, Time interval은 여전히 유의한 차이가 존재한다. 참고로, caliper matching (caliper = 0.05)을 하면 385쌍이 매칭이 되며 세 변수의 군 간 차이가 유의하지 않게 되지만, 자료의 손실이

Table 4.1. Descriptive statistics for breast ultrasonography data

Variable		Biannual ($n = 2390$)	Annual ($n = 670$)	p-value
Age		50.68 ± 9.39	55.02 ± 9.51	<0.0001
Time interval		1567.04 ± 731.20	3012.40 ± 1142.17	<0.0001
Op method	MRM	1187 (49.67%)	464 (69.25%)	<0.0001
	PM	1098 (45.94%)	187 (27.91%)	
	MRM + PM	105 (4.39%)	19 (2.84%)	

Table 4.2. Descriptive statistics for matched pairs of breast ultrasonography data

Variable		Biannual ($n = 670$)	Annual ($n = 670$)	p-value
Age		54.05 ± 9.20	55.02 ± 9.51	0.0590
Time interval		2297.51 ± 784.07	3012.40 ± 1142.17	<0.0001
Op method	MRM	446 (66.57%)	464 (69.25%)	0.6009
	PM	206 (30.75%)	187 (27.91%)	
	MRM + PM	15 (2.69%)	19 (2.84%)	

발생한다. 본 연구에서는 NNM을 사용하여 완전 매칭 후 이중 성향점수 보정 방법으로 처리 효과를 추정하고 제안하는 두 가지 붓스트랩을 이용하여 표준오차를 추정하였다. Annual군과 biannual군의 이차성 유방암 병변 검출 확률의 차이로 정의된 처리효과의 크기는 −0.0006으로 추정되었고, simple, complex 붓스트랩 방법으로 추정한 표준오차는 각각 0.0066, 0.0167로 추정되었다. 두 방법의 통계적 유의성에 대한 결론은 같지만, 모의실험에서와 마찬가지로 simple 붓스트랩을 이용하여 추정한 표준오차가 complex 붓스트랩으로 추정한 것보다 더 작음을 확인 할 수 있었다.

5. 고찰 및 결론

본 연구에서는 이중 성향점수 보정 방법을 이용한 처리효과 추정치의 표준오차 추정 방법으로 두 가지 붓스트랩을 적용하는 것에 대해 연구하였다. 이중 성향점수를 사용하는 방법은 optimal matching 또는 NNM으로 완전매칭을 할 때, 성향점수의 차이가 작지 않은 쌍이 만들어질 수 있어 생길 수 있는 편의를 줄이는 방법으로 제안되었지만, 처리효과 추정치의 표준오차의 이론적 추정치가 제시되어 있지 않다. 처리효과 추정을 위해, 대조군을 이용하여 추정된 모델로 처리군이 처리를 받지 않았을 때의 결과변수를 추정하고 이를 처리군의 관측값과 비교한다. 모형이 추정되는 군과 처리효과 추정에 사용되는 군이 다르므로 표준오차 추정에 어려움이 존재한다. 따라서 본 논문에서는 표준오차 추정을 위한 방법으로 두 가지 붓스트랩을 제시하고 다양한 상황에서 모의실험을 시행하여 어떤 방법이 더 정확한 표준오차 추정을 가능하게 하는지에 대해 연구하였다.

모의실험 결과, 매칭된 쌍의 표본으로부터 붓스트랩하여 표준오차를 추정하는 simple 붓스트랩 방법에 비해, 원자료를 먼저 붓스트랩하고 각 붓스트랩 표본에서 매칭 후 표준오차를 추정하는 complex 붓스트랩 방법이 경험적 표준오차와 더 가깝게 추정함을 알 수 있었다. 95% 신뢰구간에 대한 포함확률 또한 complex 붓스트랩 방법을 사용했을 때 0.95에 훨씬 가까웠다. 공변량들의 상관성이나 성향점수 매칭 시 사용하는 모형에 따라서는 결과에 큰 차이가 없었다. 유방초음파 검사 자료 분석에서도 simple 붓스트랩을 이용한 경우가 complex 붓스트랩을 이용한 경우보다 처리효과 추정치의 표준오차가 더 작게 추정됨을 확인하였다. 모의실험 결과를 토대로 complex 붓스트랩으로 추정한 표준오차의 정확도가 높

다 할 수 있으므로 simple 붓스트랩으로 추정한 표준오차로 처리효과의 유의성을 판단할 경우 실제 유의하지 않은 결과를 유의하다고 잘못 판단할 가능성이 존재할 것이다.

Complex 붓스트랩은 simple 붓스트랩에 비하여 더 큰 자료를 붓스트랩하여야 하고, 붓스트랩 표본의 개수만큼 성향점수 매칭을 시행하여야 한다. 따라서 훨씬 더 강도 높은 컴퓨터 계산 과정이 필요하다는 어려움이 있다. 하지만, 본 연구의 결과에 의하면 이중 성향점수 보정 방법을 사용하는 경우 simple 붓스트랩 보다는 complex 붓스트랩으로 처리효과 추정치의 표준오차를 추정하는 것이 더 바람직하다고 생각한다.

본 연구에서는 연속형과 이분형 두 종류의 결과변수에 대하여 이중 성향점수 보정 방법을 이용해 처리효과를 추정했을 때 두 가지 붓스트랩으로 표준오차를 추정하는 방법을 적용하였는데, 생존 자료에 적용하여 비교·평가해 보는 것도 흥미로운 연구가 될 것이다.

References

Austin, P. C. (2017). Double propensity-score adjustment: a solution to design bias or bias due to incomplete matching. *Statistical Methods in Medical Research*, **26**, 201–222.

Austin, P. C., Grootendorst, P., and Anderson, G. M. (2007). A comparison of the ability of different propensity models to balance measured variables between treated and untreated subject: a Monte Carlo study, *Statistics in Medicine*, **26**, 734–753.

Austin, P. C. and Small, D. S. (2014). The use of bootstrapping when using propensity-score matching without replacement: a simulation study, *Statistics in Medicine*, **33**, 4306–4319.

D'Agostino, Jr., R. B. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group, *Statistics in Medicine*, **17**, 2265–2281.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects, *Biometrika*, **70**, 41–55.

Rosenbaum, P. R. and Rubin, D. B. (1985). The bias due to incomplete matching, *Biometrics*, **41**, 103–116.

이중 성향점수 보정 방법을 이용한 처리효과 추정치의 표준오차 추정: 붓스트랩의 적용

임소정[a] · 정인경[a,1]

[a]연세대학교 의과대학 의학정보통계학과

(2017년 4월 5일 접수, 2017년 5월 23일 수정, 2017년 5월 23일 채택)

요 약

성향점수 매칭은 관찰연구에서 처리효과 추정 시 혼란변수에 의한 편의를 줄이기 위해 자주 사용되는 방법이다. 매칭을 위해 처리군에 대응되는 대조군 선정 시 처리군의 일부가 탈락되는 경우가 발생할 수 있는데, 이로 인해 편의가 발생할 수 있다. 최근, Austin (2017)의 연구에서 이중 성향점수 보정(double propensity score adjustment) 방법을 사용하는 것이 이에 대한 해결책이 될 수 있음을 제시하였다. 하지만, 처리효과 추정치의 표준오차는 이론적 추정치가 제시되지 않아 추정에 어려움이 있다. 본 연구에서는 이중 성향점수 보정 방법을 이용한 처리효과 추정치의 표준오차 추정을 위하여 두 가지 붓스트랩 방법을 제안한다. 첫 번째는 원 자료에서 성향점수 매칭 후 매칭 된 표본에서 붓스트랩 표본을 얻는 방법(simple 붓스트랩)이고, 두 번째는 원 자료에서 붓스트랩을 먼저 시행하고 각 붓스트랩 표본에서 성향점수 매칭을 하는 방법(complex 붓스트랩)이다. 두 방법의 성능을 비교하기 위하여 다양한 상황을 가정하여 모의실험을 시행한 결과 complex 붓스트랩 방법이 경험적 표준오차와 더 가까운 값으로 추정함을 알 수 있었다. 95% 신뢰구간의 포함확률도 complex 방법을 사용했을 때 0.95에 훨씬 가까웠다. 실제 자료에 적용하였을 때에도 simple 방법은 complex 방법에 비해 표준오차를 작게 추정하였다.

주요용어: 성향점수, 매칭, 관찰연구, 붓스트랩, 표준오차

[1]교신저자: (03722) 서울시 서대문구 연세로 50-1, 연세대학교 의과대학 의학정보통계학과.
E-mail: ijung@yuhs.ac