Double propensity score를 이용한 처리효과 추정치의 표준오차 추정: 붓스트랩의 적용

연세대학교 대학원
의학전산통계학협동과정
의학통계학전공
임 소 정

Double propensity score를 이용한 처리효과 추정치의 표준오차 추정: 붓스트랩의 적용

지도 정 인 경 교수

이 논문을 석사 학위논문으로 제출함

2016년 6월 일

연세대학교 대학원

의학전산통계학협동과정

의학통계학전공

임 소 정

연세대학교
YONSEI UNIVERSITY

임소정의 석사 학위논문을 인준함

심사위원 정인경 인

심사위원 남정모 인

심사위원 송기준 인

연세대학교 대학원

2016년 6월 일

차 례

표 차 례

국 문 요 약

Double propensity score를 이용한 처리효과 추정치의 표준오차 추정: 붓스트랩의 적용

성향점수 매칭은 관찰연구에서 처리효과 추정 시 혼란변수 효과를 줄이기 위해 사용되는 방법이다. 매칭을 이용하여 처리 군에 대응되는 대조군 선정 시 처리군의 일부가 탈락되는 경우가 발생하게 된다. 이로 인해 발생하는 편의가 Rosenbaum과 Rubin (1985)이 언급한 '불완전 매칭에 기인한 편의'이다. 최근에 Austin (2014)의 연구에서 double propensity score가 위 문제의 분석적 해결방안으로 제시된 바 있다. 그러나 double propensity score를 이용한 처리효과 추정치의 표준오차는 이론적 추정치가 제시되지 않아 추정에 어려움이 존재한다.

본 연구는 double propensity score를 이용한 처리효과 추정치의 표준오차 추정을 위한 방법으로 두 가지 붓스트랩을 제시하여 다양한 상황에서 두 방법의 퍼포먼스를 비교해 보았다.

붓스트랩 방법으로는 Austin과 Small (2014)이 제시한 두 가지 방법을 사용했다. 첫 번째 방법(simple 붓스트랩)은 원자료를 이용하여 성향점수 매칭 후 붓스트랩 표본을 얻는다. 두 번째 방법(complex 붓스트랩)은 원자료에서 붓스트랩 표본을 얻어 각 붓스트랩 표본으로부터 성향점수 매칭을 한다. 결과변수와 처리변수에 영향을 미치는 여부와 변수들 간의 상관성을 고려하여 공변량을 생성했다. 성향점수 매칭에 이용된 변수들의 조합을 달리했으며 처리군의 비율이 0.05, 0.1인 경우에 결과를 비교해 본다. 결과변수가 연속형과 이분형일 경우를 가정하여 1000번의 모의실험을 했다. 또한 유방암 환자의 유방초음파 검사 자료에 두 붓스트랩을 적용하여 결과를 비교해 본다.

모의실험을 통해 다양한 상황을 가정하여 퍼포먼스를 비교 한 결과 complex 붓스트랩을 사용한 경우가 simple 붓스트랩을 사용한 경우보다 모든 상황에서 경험적 표준오차와 유사한 표준오차가 추정 돼 퍼포먼스가 더 좋은 것을 확인 할 수 있었다. Simple 붓스트랩을 사용한 경우 표준오차가 과소추정 됐으며 실제 자료 분석에서도 이와 같은 결과를 얻었다. 95% 신뢰구간에 대한 포함확률이 complex 붓스트랩을 사용한 경우에 95%에 가까웠으나 simple 붓스트랩을 사용한 경우에서는 95%보다 작으며 상이한 결과를 보였다. 따라서 double propensity score를 이용해 추정한 효과의

표준오차 추정 시 complex 붓스트랩을 사용하여 표준오차를 추정하는 것이 정확도를 높일 수 있는 방법이 될 것으로 기대된다.

핵심되는 말 : 성향점수 매칭, double propensity score, 붓스트랩, 표준오차, 모의실험

제 1장 서론

1.1 연구 배경 및 목적

무작위 실험(Randomized controlled trial; RCT)에서는 처리군에 배치되는 여부가 관측된 혹은 관측되지 않은 특성들에 의해 혼재되지 않고, 두 군의 결과변수를 바로 비교하여 처리효과 추정이 가능하다. 임상의학분야에서 수행되는 관찰연구에서는 처리군에 배치되는 여부가 연구 대상자의 특성에 의해 영향을 받는다. 이로 인해 각 군으로의 분류단계에서부터 처리군의 기본 특성이 대조군의 기본 특성과 다른 상황이 존재한다. 관찰연구에서는 이러한 구조적 차이를 고려하여 처리효과를 추정해야 한다. 성향점수 매칭(propensity score matching)은 관찰연구에서 처리효과 추정 시 처리효과와 관련된 공변량들의 효과를 줄여 각 군의 특성의 차이를 줄이기 위해 사용되는 방법이다.

매칭 시 처리군과 성향점수가 유사한 대조군만을 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 (2014)의 연구에서 double propensity score가 위 문제의 분석적 해결방안으로 제시된 바 있다. 그러나 double propensity score를 이용한 처리효과 추정치의 표준오차는 이론적 추정치가 제시되지 않아 추정에 어려움이 존재한다.

본 연구에서는 double propensity score를 이용한 처리효과 추정치의 표준오차 추정을 위한 방법으로 두 가지 붓스트랩 방법을 제시하고, 두 가지 방법으로 추정된 표준오차의 정확도를 비교하고자 한다.

1.2 연구 내용 및 방법

본 연구에서는 모의실험을 통해 표본을 생성하고 몬테칼로 방법을 적용시켜 표준오차 추정치들의 평균을 구한다. Double propensity score를 이용한 처리효과 추정치의 표준오차를 추정하기 위한 방법으로 Austin과 Small (2014)이 제안한 simple, complex 붓스트랩을 사용한다. 표본의 공변량은 결과변수와 처리변수에 영향을 미치는 여부와 변수들 간의 상관성을 고려하여 특정 분포 하에서 임의로 생성한다. 성향점수 매칭에 이용된 변수들의 조합을 달리하고 처리군의 비율도 달리하여 결과를 비교한다. 결과변수가 연속형과 이분형일 경우를 가정하여 double propensity score를 이용해 처리효과를 추정했을 때 두 가지 붓스트랩 방법이 표준오차를 얼마나 잘 추정하는지 비교한다. 표준오차 추정의 정확도를 비교하기 위해 붓스트랩으로 추정된 표준오차와 경험적 표준오차의 차이와 95% 신뢰구간에 대한 포함확률(coverage probability)을 비교하고자 한다.

1.3 논문의 구성

제 1장에서는 연구의 배경 및 목적과 연구 내용 및 방법을 소개한다. 제 2장에서는 본 연구의 이론적 배경이 되는 성향점수와 double propensity score, 표준오차 추정을 위한 붓스트랩 방법에 대해 살펴본다. 3장에서는 모의실험을 통해 다양한 상황을 가정하여 두 가지 붓스트랩 방법을 비교·평가한다. 4장에서는 실제자료에서 두 가지 붓스트랩을 사용하여 표준오차를 추정했을 때의 결과를 비교한다. 마지막으로 5장에서는 결론 및 고찰에 대해 논의한다.

제 2장 이론적 배경

2.1 성향점수(Propensity score)

Rosenbaum과 Rubin (1983)이 제안한 성향점수는 관찰된 공변량들이 주어졌을 때, 해당 공변량 값들을 가지는 각 개체가 처리군에 배치 될 조건부 확률로써 정의되며 식은 아래와 같다

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

z는 처리군일 때는 1이고 대조군일 때는 0을, $\boldsymbol{x}$는 공변량들을 나타낸다.

처리효과 추론 시 강한 무관성가정이 성립 할 때, 공변량들이 주어진 경우 두 군의 차이는 성향점수 매칭을 통해 추정한다. 매칭 후에는 처리군과 대조군에서 관찰된 공변량들의 분포가 유사하게 되므로 이들에 의해 발생하는 편의를 제거 할 수 있다. 이러한 것은 무작위 실험에 근접한 연구 설계의 형태를 갖출 수 있도록 한다.

성향점수의 특성은 균형점수(balancing score)라는 점으로, 성향점수가 주어진 상황에서 공변량들의 조건부 분포는 각 군에서 같으며 이를 표현하면 아래와 같다

$$\boldsymbol{x} \perp z | e(\boldsymbol{x}).$$

강한 무관성가정이 성립 할 때, 임의의 성향점수 값에서 두 군 간의 평균 차이는 처리효과의 불편추정량이다.

2.1.1에서는 본 연구에서 중점적으로 다루는 처리군에서의 처리효과에 대한 개념을 소개한다. 2.1.2에서는 성향점수 매칭을 통한 분석의 가정인 강한 무관성가정에 대해 소개한다. 2.1.3에서는 성향점수 매칭을 하는 이유와 활용하는 방법을 살펴보고, 성향점수가 일치하는 여부와 짝을 탈락시키는 여부에 따른 매칭 방법에 대해 소개한다.

2.1.1 처리효과

처리할당이 두 가지가 존재하는 경우, 각 개체가 두 가지 처리할당에 대한 두개의 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 ATE(average effect of treatment in an entire sample or population)로 전체 자료 혹은 모집단에서의 처리효과로 정의된다. 이와 관련된 처리효과의 측도로 처리군에서의 처리효과(average treatment effect for the treated; ATT)는 아래와 같다

$$ATT = \tau_1 = E(r_1 - r_0 | z = 1).$$

2.1.2 강한 무관성가정(Strongly ignorable assignment assumption)

무작위 실험은 처리할당이 무작위로 이루어진다는 점에서 관찰연구와 구분된다. 이로 인해 두 가지 차이점이 발생 하는데 먼저 무작위 실험의 경우 성향점수는 알려진 함수로써 하나로 특정 할 수 있다. 관찰연구의 경우 성향점수를 모형을 통해 추정해야 한다. 두 번째로 적절하게 시행된 무작위 실험의 경우 성향점수 함수에 사용된 공변량들이 처리할당과 결과변수에 영향을 미치는 모든 변수들로 알려져 있다. 이러한 무작위 실험에서 각 개체의 공변량들($\boldsymbol{x}$)이 주어졌을 때, 처리할당(z)은 결과변수들(r_1, r_0)과 조건부독립인 가정을 만족하며 이를 표현하면 아래와 같다

$$(r_1, r_0) \perp z | \boldsymbol{x}.$$

공통영역의 가정(common support assumption)은 두 군의 처리할당 확률 분포가 공통의 영역 내에 있다는 가정이며 이를 표현하면 아래와 같다

$$0 < \Pr(z = 1 | \boldsymbol{x}) < 1.$$

위의 두 가정이 충족될 경우, 강한 무관성가정(strongly ignorable assignment assumption)을 만족한다고 할 수 있다. 만약 관찰연구에서도 강한 무관성가정을 만족한다면, 즉, 적절한 공변량들이 주어진다면 성향점수를 통제하는 것으로 편의가 없는 처리효과 추정치를 산출할 수 있다.

2.1.3 성향점수 매칭(Propensity score matching)

처리군과 대조군의 비교에 있어 혼란변수(confounding)의 존재는 편의를 발생시킨다. 성향점수를 이용한 매칭은 관찰된 공변량들에 의한 혼란변수 효과를 통제하기 위해 사용된다. 매칭을 통해 유사한 성향점수를 가지는 각 군의 개체로 이루어진 짝들의 집합을 형성할 수 있다. Rosenbaum과 Rubin (1983)은 유사한 성향점수를 가진 개체들은 측정된 공변량 값들의 분포가 유사할 것이라고 언급했다. 이는 무작위 실험과 같이 편의가 없는 처리효과를 추정 할 수 있음을 의미한다.

성향점수를 추정하기 위해서 로지스틱 회귀모형이나 로지스틱 다층모형, 프로빗 모형 등 이분형 변수를 종속변수로 하는 모형을 사용한다. 추정된 성향점수를 이용해 편의가 없는 처리효과를 추정하는 다양한 방식이 존재한다. 대표적인 방법들로, 짝을 이루는 매칭, 성향점수가 유사한 그룹을 생성하는 하위분류(subclassification), 성향점수를 공변량 보정에 사용하는 방법이 있다. 이 외에도 성향점수의 역수를 가중치로 하여 각 군에 배치될 확률이 동일하게 적용되는 인위적 표본을 생성하는 가중치 방법으로 처리효과 추정에 있어 편의를 줄일 수 있다.

매칭 방법의 경우, 매칭된 짝의 성향점수가 일치하는 여부와 짝을 탈락 시키는 여부에 따라 종류를 나눌 수 있다. 정확 매칭(exact matching)은 성향점수가 일치하는 짝을 매칭하는 방법이다. 소수의 처리군과 다수의 대조군을 매칭 할 때 성향점수가 상이한 매칭 짝의 경우 정확 매칭을 위해 분석에서 제외시킬 수 있다. 이와 같이 매칭을 할 때 분석에서 탈락되는 짝이 존재하는 여부에 따라 완전 매칭(exact matching)과 불완전 매칭으로 나눌 수 있다. 완전 매칭은 모든 매칭된 짝을 분석에 포함하는 방법으로 optimal matching, nearest neighbor matching(NNM)방법 등이 있다. 불완전 매칭은 성향점수가 상이한 매칭 짝의 경우 정확 매칭을 위해 분석에서 제외하는 방법으로 caliper matching 등이 있다.

2.2 Double propensity score

Double propensity score는 Austin (2014)에 의해 제안된 방법으로 '불완전 매칭에 기인한 편의'를 줄이고자 완전 매칭을 하여 ATT를 추정한다. 이때 소수의 처리군 모두와 다수의 대조군 중 일부를 성향점수를 이용해 완전 매칭 한다. 완전 매칭을 하는 경우 상이한 성향점수를 가진 짝이 매칭됨으로 인한 residual confounding이 생길 수 있다. Double propensity score에서는 residual confounding을 줄이기 위하여 성향점수를 이용해 결과변수에 추가적인 보정을 가한다.

2.2.1에서는 본 연구에서 사용하는 분석방법인 double propensity score에서 해결하고자 하는 '불완전 매칭에 기인한 편의'에 대해 소개한다. 2.2.2에서는 double propensity score를 이용하여 처리효과를 추정하는 방법에 대해 소개한다.

2.2.1 불완전 매칭(Incomplete matching)에 기인한 편의

정확 매칭을 위해 불완전 매칭을 할 경우 성향점수가 높은 이들이 매칭 될 개체를 찾기 힘들기 때문에 주로 탈락된다. 마찬가지로 ATT추정 시 처리군 일부를 탈락시킬 경우 처리효과의 일반화(generalizability)에 문제가 생길 수 있다. 특히 연구 목적이 처리군 모두를 대상으로 효과를 알아보고자 하는 것일 때 가장 전형적인 개체를 탈락시키는 우를 범할 가능성이 있다. 이렇게 처리군에 속한 개체의 일부가 분석에 포함되지 못할 경우 생기는 편의가 Rosenbaum과 Rubin (1985)이 언급한 '불완전 매칭에 기인한 편의'이다.

Double propensity score를 사용하면 완전 매칭을 하게 되므로 위에 언급한 현상들을 지양 할 수 있다.

2.2.2 매칭된 자료의 처리효과 추정

소수의 처리군과 다수의 대조군이 존재 할 경우, 추정된 성향점수를 이용하여 두 군을 1:1로 완전 매칭 한다. 먼저 처리군의 성향점수를 $e_1(\boldsymbol{x})$, 대조군의 성향점수를 $e_0(\boldsymbol{x})$라고 정의한다. $e_0(\boldsymbol{x})$만을 공변량으로 하여 결과변수를 예측하는 회귀모형인 $m_0(e(\boldsymbol{x}))$를 추정한다. $m_0(e(\boldsymbol{x}))$에 대한 식은 아래와 같다

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

$m_0(e(\boldsymbol{x}))$는 결과변수의 종류에 따라 선형 회귀 모형, 로지스틱 회귀 모형 등 다양한 모형을 적용 시킬 수 있다. 위 모형을 사용하여 처리군의 개체가 처리를 받지 않았을 때의 잠재적 결과변수의 r_0의 추정치를 얻는다. 처리군의 개체가 처리를 받았을 때의 잠재적 결과변수 r_1는 처리군의 개체에서 관측된 결과변수 값을 사용한다. 처리군에서 얻어진 r_0, r_1에 대한 추정치 차이의 평균을 계산하여 처리효과가 추정되며 식은 아래와 같다

$$\widehat{ATT} = \frac{1}{N}\sum_{i=1}^{N}(r_{1,i} - m_0(e_1(\boldsymbol{x}_i))) = E(r_1) - E(m_0(e_1(\boldsymbol{x}))).$$

여기서, $e_1(\boldsymbol{x}_i)$는 처리군 N명 중 i번째 개체의 성향점수 추정치이며 $r_{1,i}$는 처리를 받았을 때의 잠재적 결과변수 r_1의 관측 값이다. $m_0(e_1(\boldsymbol{x}_i))$는 처리군의 i번째 개체가 만약 처리를 받지 않았을 때의 잠재적 결과변수 r_0의 추정치이다.

결과변수가 연속형일 경우 $m_0(e(\boldsymbol{x}))$는 선형 회귀모형을 사용하여 추정치가 연속형으로 나타난다. 결과변수가 이분형일 경우 $m_0(e(\boldsymbol{x}))$는 로지스틱 회귀모형을 사용하여 추정치가 확률로 나타난다. ATT추정 시 관측된 값 대신에, $e_1(\boldsymbol{x})$만을 공변량으로 하여 결과변수를 예측하는 회귀모형인 $m_1(e(\boldsymbol{x}))$을 추정하여 처리군의 결과변수 추정치를 계산 할 수도 있다. 결국 이 추정치들의 평균은 관측 값들의 평균과 같다.

2.3 표준오차 추정을 위한 붓스트랩 방법

매칭 후 유사한 성향점수를 가진 개체들은 측정된 공변량 값들의 분포가 유사할 것이다. 이로 인해 매칭된 개체들끼리는 연관성이 존재하게 된다. 이러한 연관성은 처리효과 추정치의 표준오차 추정에 있어 다양한 견해를 갖게 한다. Schafer과 Kang (2008)은 성향점수를 이용한 매칭으로 처리효과 추정 시 통계적 유의성을 판단함에 있어 독립 표본을 가정한 추론을 제시했다. 반면 몬테칼로 모의실험을 이용한 연구들의 결과를 보면 매칭 후 연관성을 고려한 표본을 가정한 경우가 처리효과 추정치들의 표집분포의 특성을 더 잘 반영한 것으로 나타난 바 있다(Austin 2009; Austin 2011; Gayat, Resche-Rigon, Mary and Porcher 2012; Austin 2013). 2.3.1과 2.3.2에서는 성향점수 매칭 후 처리효과 추정치의 표준오차를 추정하기 위한 두 가지 붓스트랩 방법(Austin and Small 2014)을 소개한다. 2.3.3에서는 두 가지 신뢰구간 추정 방법을 소개한다.

2.3.1 Simple 붓스트랩

Simple 붓스트랩은 매칭된 자료에서 전형적으로 적용되는 방법으로 원자료를 붓스트랩 시키는 것이 아닌 매칭된 짝들을 붓스트랩 시키는 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.

2.3.2 Complex 붓스트랩

Complex 붓스트랩은 매칭전의 원자료를 붓스트랩 시키는 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 모델이 추정되고 매칭이 이루어진다. 이 B개의 매칭된 표본들의 처리효과 추정치들의 표준편차는 원자료로부터 구해진 처리효과 추정치의 표준오차 추정치로 사용된다. Complex 붓스트랩은 Simple 붓스트랩과 비교할 때 두 가지 추가적인 변동을 고려하는데, 이는 성향점수 모델(propensity score model; PSM)을 추정하는 변동과 매칭된 표본을 형성하는 변동이다. 또한 simple 붓스트랩은 원자료로부터 하나의 매칭 알고리즘을 시행하는 반면 complex 붓스트랩은 B개의 추가적인 성향점수 매칭 알고리즘을 시행한다. 따라서 더 강도 높은 컴퓨터 계산 과정이 필요하다.

2.3.3 신뢰구간 추정방법

신뢰구간 추정을 위한 두 가지 방법은 다음과 같다. 첫 번째로 표준정규분포 이론을 근거로 한 왈드 타입(Wald type) 95% 신뢰구간으로 $\hat{\theta} \pm 1.96 se(\hat{\theta})$을 사용할 수 있다. $\hat{\theta}$은 처리효과의 추정치를, $se(\hat{\theta})$은 표준오차의 붓스트랩 추정치를 의미한다. 두 번째로 백분위수에 근거한 비모수적 방법을 사용하여 95% 신뢰구간을 추정 할 수 있다. B개의 붓스트랩 표본으로부터 얻은 처리효과 추정치들의 2.5번째와 97.5번째의 백분위수를 사용한다.

제 3장 모의실험

이 장에서는 2.3에서 소개한 두 가지 붓스트랩 방법으로 추정된 표준오차의 정확도를 비교하고 결과에 일관성이 있는지 확인하기 위해 다양한 상황을 가정하여 모의실험을 시행하였다. 먼저 성향점수 매칭 후 처리효과 추정 시 2.2의 double propensity score를 적용했을 경우의 표준오차 추정 결과를 비교한다. 추가 연구로 double propensity score를 적용하지 않고 처리효과를 추정 할 경우의 표준오차 추정 결과를 알아본다. 붓스트랩으로 추정된 표준오차와 경험적 표준오차의 차이, 95% 신뢰구간에 대한 포함확률을 비교하여 추정된 표준오차의 정확도를 평가한다.

3.1 모의실험 설계

표본의 생성은 Austin과 Small (2014)의 논문과 Austin (2014)의 논문에서 제시하는 데이터 생성과정을 참고하여 수행하였다.

처리변수와 결과변수에 영향을 미치는 여부를 고려하여 10개의 공변량들($x_1 \cdots x_{10}$)을 평균이 0이고 분산이 1인 표준정규분포로부터 생성하였다. 10개의 공변량들 중 7개는 처리할당에 영향을 주고($x_1 \cdots x_7$) 7개는 결과변수에 영향을 준다($x_4 \cdots x_{10}$). 각 공변량은 약한, 중간, 강한, 매우강한 정도로 처리할당 혹은 결과변수에 영향을 준다. 공변량들 간의 상관성이 존재하지 않는 경우에 공변량들의 분포는 아래와 같다

$$x_m \sim N(0,1),\ (m=1,\ldots,10).$$

공변량들의 상관성이 존재하는 경우엔 동일한 특성 범주 안에 포함된 공변량들 간의 피어슨의 상관계수(ρ)가 0.3, 0.5, 0.8이 되도록 아래와 같이 설정하였다

$$\boldsymbol{x_T}=\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}\sim MVN(\mathbf{0},\Sigma),\ \Sigma=\begin{bmatrix}1\,\rho\,\rho\\\rho\,1\,\rho\\\rho\,\rho\,1\end{bmatrix},\ (\rho=0.3,0.5,0.8),$$

$$\boldsymbol{x_{TY}}=\begin{pmatrix}x_4\\x_5\\x_6\\x_7\end{pmatrix}\sim MVN(\mathbf{0},\Sigma),\ \Sigma=\begin{bmatrix}1\,\rho\,\rho\,\rho\\\rho\,1\,\rho\,\rho\\\rho\,\rho\,1\,\rho\\\rho\,\rho\,\rho\,1\end{bmatrix},\ (\rho=0.3,0.5,0.8),$$

$$\boldsymbol{x_Y}=\begin{pmatrix}x_8\\x_9\\x_{10}\end{pmatrix}\sim MVN(\mathbf{0},\Sigma),\ \Sigma=\begin{bmatrix}1\,\rho\,\rho\\\rho\,1\,\rho\\\rho\,\rho\,1\end{bmatrix},\ (\rho=0.3,0.5,0.8).$$

처리변수 생성 시 처리변수에 영향을 주도록 공변량들의 계수를 다르게 하여 처리변수에 영향을 주는 강도를 설정하며, 영향을 주지 않는 공변량들은 계수값을 0을 주어 식에서 제외시킨다. 처리변수 생성에 관한 식은 아래와 같다

$$logit(p_{i,treat})=\beta_{0,treat}+\beta_W x_1+\beta_M x_2+\beta_S x_3+\beta_W x_4+\beta_M x_5+\beta_S x_6+\beta_{VS} x_7,$$

$$z_i\sim Bernoulli(p_{i,treat}).$$

위의 식에서 절편($\beta_{0,treat}$)은 모의실험 자료에서 식을 통해 계산된 각 개체의 처리할당 확률의 평균을 결정한다. 이번 연구에서는 처리할당 확률의 평균을 0.05, 0.1로 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 처리할당에 영향을 주는 것을 뜻한다. 개체특정적인 모수 $p_{i,treat}$에 대한 Bernoulli 분포를 이용하여 처리변수(z_i)를 생성한다.

결과변수는 연속형, 이분형으로 두 경우를 가정하며 결과변수의 생성은 처리변수의 생성과 동일하게 진행된다. 결과변수가 연속형인 경우 선형회귀모형을 사용한다. 결과변수 생성에 관한 식은 아래와 같다

$$r_{z,i,continuous} = \beta_{treat,continuous} z_i + \beta_W x_4 + \beta_M x_5 + \beta_S x_6 + \beta_{VS} x_7 + \beta_W x_8 + \beta_M x_9 + \beta_S x_{10} + \epsilon_i ,$$

$$\epsilon_i \sim N(0,\sigma^2),\ \sigma = 3.$$

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

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

결과변수가 이분형인 경우 로지스틱회귀모형을 사용하며 식은 아래와 같다

$$logit(p_{i,outcome}) = \beta_{0,outcome} + \beta_{treat,binary} z_i + \beta_W x_4 + \beta_M x_5 + \beta_S x_6 + \beta_{VS} x_7 + \beta_W x_8 + \beta_M x_9 + \beta_S x_{10} ,$$

$$r_{z,i,binary} \sim Bernoulli(p_{i,outcome}).$$

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

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

성향점수 추정을 위해 PSM으로 로지스틱회귀모형을 사용하였다. 모델에 포함된 공변량들의 특성에 따라 조합을 달리하여 모델을 비교하였다. PSM1은 모든 변수를 포함한 모형이다. PSM2는 결과변수에 영향을 주는 변수들을 포함한 모형이다. PSM2가 처리할당에만 영향을 주는 변수들도 모형에 포함한 경우인 PSM1과 비교할 때 더 좋은 추정치를 제공한다는 연구결과가 존재한다(Austin, Grootendorst and Anderson 2007). PSM3는 회귀분석을 시행하여 결과변수에 유의한 영향을 미치는 변수가 S개일 때 유의한 변수($x_1,\cdots,x_S$)들을 포함한 모형이다.

표 1. 성향점수 모형

PSM 모형	모형 식
PSM1	$logit[\Pr(z=1)] = \alpha_{1,0} + \alpha_{1,1}x_1 + \alpha_{1,2}x_2 + \alpha_{1,3}x_3 + \alpha_{1,4}x_4 + \alpha_{1,5}x_5 + \alpha_{1,6}x_6 + \alpha_{1,7}x_7 + \alpha_{1,8}x_8 + \alpha_{1,9}x_9 + \alpha_{1,10}x_{10}$
PSM2	$logit[\Pr(z=1)] = \alpha_{2,0} + \alpha_{2,4}x_4 + \alpha_{2,5}x_5 + \alpha_{2,6}x_6 + \alpha_{2,7}x_7 + \alpha_{2,8}x_8 + \alpha_{2,9}x_9 + \alpha_{2,10}x_{10}$
PSM3	$logit[\Pr(z=1)] = \alpha_{3,0} + \cdots + \alpha_{3,s}x_s,\ (s=1,\ldots,S)$

처리효과는 생성된 표본의 각 개체로부터 추정된 성향점수를 이용하여 매칭을 한 후 2.2.2의 double propensity score에서 제시된 매칭된 자료의 처리효과 추정방법을 이용하여 추정한다. Double propensity score를 적용하지 않을 경우에는 매칭 후 결과변수에 추가적인 보정을 하지 않고 추정하며 본 논문에서는 이를 성향점수 매칭을 적용하는 경우로 지칭했다. 본 모의실험에서는 매칭 방법으로 완전 매칭 중 NNM을 사용한다. 표준오차는 생성된 표본으로부터 2.3.1의 simple 붓스트랩과 2.3.2의 complex 붓스트랩을 이용하여 추정된다.

3.2 모의실험 결과

표 2～5는 2.2의 double propensity score를 적용한 결과이고 표 6～9는 성향점수 매칭을 적용한 결과이다. 결과변수의 종류와 처리할당 비율을 달리하여 모의실험을 수행하였다. 모집단으로부터 1000개의 표본을 생성하며 각 표본은 3000개의 개체로 이루어진다. 모집단은 3.1에 기술된 과정에 따라 생성된 공변량, 처리변수, 결과변수를 가지는 개체들로 이루어진다. 표준오차 추정 시 100번의 붓스트랩을 시행한다.

경험적 표준오차(empirical standard error; SE(Empirical))는 1000개의 표본으로부터 추정된 처리효과들의 표준편차이다. 붓스트랩으로 추정된 표준오차(estimated standard error; SE(Bootstrap))가 경험적 표준오차와 유사할수록 정확한 추정이 가능함을 의미한다.

먼저 처리효과 추정이 잘 되었는지 확인하기 위해 1000개의 표본으로부터 구해진 처리효과 추정치의 평균과 편의를 나타냈다. 표준오차 추정의 정확도를 비교하기 위해 1000개의 표본으로부터 2.3의 붓스트랩 방법으로 1000개의 표준오차를 추정하여 평균을 내고 이를 경험적 표준오차와 비교했다. 그리고 왈드 타입의 95% 신뢰구간을 구하여 포함확률을 비교했다.

3.2.1 Double propensity score를 이용한 경우

먼저 처리효과 추정 결과를 비교해 보면 결과 변수가 연속형일 때 처리효과 추정치의 평균이 1과 가까워 편의가 매우 작게 나타났다. 결과변수가 이분형일 때 처리효과 추정치의 평균이 0.02와 가까워 편의가 매우 작게 나타났다. 결과변수가 연속형일 때 공변량들 간의 상관성이 증가할수록 편의가 커지는 양상을 보였으며 이분형일 때는 일부 경우에서 이러한 양상을 보였다. 특히 PSM1모형을 사용한 경우에 위와 같은 양상이 잘 드러났다. 대부분의 경우에서 처리할당 확률이 0.1에서 0.05로 감소할 때 편의가 작아졌으며 PSM1모형을 사용한 경우에 비해 PSM2, PSM3를 사용한 경우에서 편의가 작게 나타났다.

표준오차 추정 결과를 비교해 보면 complex 붓스트랩을 사용한 추정치가 경험적 표준오차와 매우 유사한 것을 확인 할 수 있었다. 반면 simple 붓스트랩을 사용한 추

정치는 경험적 표준오차와 매우 상이하며 과소추정 되는 것을 확인 할 수 있었다. 처리 할당 확률, 공변량들 간의 상관성, PSM의 모든 경우를 고려한 24개의 조합에서 경험적 표준오차에 대한 추정된 표준 오차의 평균 비율을 계산했다. 결과변수가 연속형인 경우 simple, complex 붓스트랩에서 0.74, 1.07이고 이분형인경우에는 0.69, 0.97이었다. 평균 비율이 1에 가까울수록 경험적 표준오차와 더 유사한 값이 추정됨을 나타낸다. 특히 처리 할당 확률이 0.1에서 0.05로 감소할 때 simple 붓스트랩을 사용한 표준오차의 추정치가 과소추정 되는 정도가 더 커졌다. 또한 결과변수의 종류에 따라 complex 붓스트랩으로 추정된 표준오차와 경험적 표준오차와의 관계가 달랐다. 연속형일 때 모든 경우에서 complex 붓스트랩으로 추정된 표준오차가 경험적 표준오차보다 컸으나 이분형일 때는 작은 경우가 존재하는 것을 확인 할 수 있었다. 공변량들 간의 상관성, PSM이 달라질 경우에는 결과의 차이가 크지 않았다.

포함확률을 비교해 보면 complex 붓스트랩을 사용한 경우 포함확률이 0.95에 가까웠다. Simple 붓스트랩을 사용한 경우 포함확률이 0.85에 가까웠다.

3.2.2 성향점수 매칭을 이용한 경우

먼저 처리효과 추정 결과를 비교해 보면 결과 변수가 연속형일 때 처리효과 추정치의 평균이 1과 가까워 편의가 매우 작게 나타났다. 결과변수가 이분형일 때 처리효과 추정치의 평균이 0.02와 가까워 편의가 매우 작게 나타났다. 결과변수가 연속형일 때 공변량들 간의 상관성이 증가할수록 편의가 커지는 양상을 보였으며 이분형일 때는 일부 경우에서 이러한 양상을 보였다. 특히 PSM1모형을 사용한 경우에 위와 같은 양상이 잘 드러나며, 결과변수가 연속형, 이분형일 때 모두에서 처리 할당 확률이 0.1일 때 공변량들 간의 상관성이 높은 상황에서 편의가 큰 경우가 나타났다. 대부분의 경우에서 처리할당 확률이 0.1에서 0.05로 감소할 때 편의가 작아졌으며 PSM1모형을 사용한 경우에 비해 PSM2, PSM3를 사용한 경우에서 편의가 작게 나타났다. 전체적으로 성향점수 매칭을 이용한 처리효과 추정치의 편의가 double propensity score를 이용한 편의보다 더 컸다.

표준오차 추정 결과를 비교해 보면 전체적으로 두 붓스트랩 방법을 사용한 추정치가 경험적 표준오차와 유사한 값을 가지는 것을 확인 할 수 있었다. 특히 결과변수가 연속형일 때 simple 붓스트랩을 사용한 추정치가 complex 붓스트랩을 사용한 추정치보다 경험적 표준오차와 더 유사한 값을 가지는 것을 확인 할 수 있었다. 처리 할당

확률, 공변량들 간의 상관성, PSM의 모든 경우를 고려한 24개의 조합에서 경험적 표준오차에 대한 추정된 표준 오차의 평균 비율을 계산했다. 결과변수가 연속형인 경우 simple, complex 붓스트랩에서 1.03, 1.07이고 이분형인경우에는 0.95, 0.97이었다. 이는 double propensity score를 이용한 경우와 상이한 결과이다. 또한 결과변수의 종류에 따라 붓스트랩으로 추정된 표준오차와 경험적 표준오차와의 관계가 달랐다. 결과변수가 연속형일 때 대부분의 경우에서 두 붓스트랩으로 추정된 표준오차가 경험적 표준오차보다 컸으나 이분형일 때는 작은 경우가 존재했다. 특히 simple 붓스트랩으로 추정된 표준오차는 결과변수가 이분형일 때 대부분의 경우에서 경험적 표준오차보다 작았다. 처리 할당 확률, 공변량들 간의 상관성, PSM이 달라질 경우에는 결과의 차이가 크지 않았다.

포함확률을 비교해 보면 두 붓스트랩을 사용한 경우 모두 포함확률이 0.95에 가까워 결과의 차이가 크지 않았다. 예외적으로 결과변수가 연속형, 이분형일 때 모두에서 처리 할당 확률이 0.1일 때 공변량들 간에 상관성이 높은 상황에서 포함확률이 0.95에서 벗어나는 경우가 나타났다. 이는 앞서 처리 할당 확률이 0.1일 때 처리효과 추정치의 편의가 큰 경우가 나타난 것과 관련이 있다.

표 2. Double propensity score 에서 결과변수가 연속형인 경우 처리효과와 표준오차 추정치 및 포함확률(처리할당비율=0.05)

Continuous outcome : difference in means								
Correlation	Model	Empirical			SE(Bootstrap)		Coverage probability (Wald type)	
		Mean	Bias	SE(Empirical)	SE(Simple)	SE(Complex)	SE(Simple)	SE(Complex)
0	PSM 1	0.9948	0.0052	0.3591	0.2582	0.3774	0.844	0.965
	PSM 2	1.0051	0.0051	0.3509	0.2546	0.3760	0.862	0.956
	PSM 3	0.9994	0.0006	0.3481	0.2548	0.3774	0.864	0.957
0.3	PSM 1	0.9925	0.0075	0.3623	0.2653	0.3799	0.842	0.957
	PSM 2	1.0021	0.0021	0.3550	0.2592	0.3772	0.847	0.959
	PSM 3	0.9931	0.0069	0.3615	0.2596	0.3787	0.831	0.960
0.5	PSM 1	0.9621	0.0379	0.3793	0.2701	0.3826	0.836	0.957
	PSM 2	0.9686	0.0314	0.3603	0.2633	0.3782	0.838	0.959
	PSM 3	0.9672	0.0328	0.3680	0.2639	0.3832	0.833	0.949
0.8	PSM 1	0.9428	0.0572	0.3776	0.2753	0.3848	0.846	0.956
	PSM 2	0.9604	0.0396	0.3444	0.2662	0.3802	0.865	0.962
	PSM 3	0.9533	0.0467	0.3531	0.2673	0.3882	0.849	0.964

표 3. Double propensity score에서 결과변수가 이분형인 경우 처리효과와 표준오차 추정치 및 포함확률(처리할당비율=0.05)

Binary outcome : risk difference								
Correlation	Model	Empirical			SE(Bootstrap)		Coverage probability (Wald type)	
		Mean	Bias	SE(Empirical)	SE(Simple)	SE(Complex)	SE(Simple)	SE(Complex)
0	PSM 1	0.0216	0.0016	0.0495	0.0320	0.0458	0.849	0.950
	PSM 2	0.0209	0.0009	0.0489	0.0314	0.0454	0.837	0.959
	PSM 3	0.0216	0.0016	0.0486	0.0314	0.0454	0.842	0.960
0.3	PSM 1	0.0157	0.0043	0.0479	0.0341	0.0479	0.866	0.960
	PSM 2	0.0211	0.0011	0.0462	0.0330	0.0471	0.866	0.970
	PSM 3	0.0199	0.0001	0.0461	0.0330	0.0475	0.866	0.965
0.5	PSM 1	0.0161	0.0039	0.0488	0.0350	0.0485	0.852	0.958
	PSM 2	0.0182	0.0018	0.0477	0.0336	0.0476	0.830	0.960
	PSM 3	0.0193	0.0007	0.0474	0.0337	0.0484	0.841	0.960
0.8	PSM 1	0.0132	0.0068	0.0472	0.0362	0.0494	0.863	0.962
	PSM 2	0.0189	0.0011	0.0473	0.0345	0.0479	0.843	0.959
	PSM 3	0.0206	0.0006	0.0465	0.0347	0.0494	0.860	0.975

표 4. Double propensity score에서 결과변수가 연속형인 경우 처리효과와 표준오차 추정치 및 포함확률(처리할당비율=0.1)

Continuous outcome : difference in means								
Correlation	Model	Empirical			SE(Bootstrap)		Coverage probability (Wald type)	
		Mean	Bias	SE(Empirical)	SE(Simple)	SE(Complex)	SE(Simple)	SE(Complex)
0	PSM 1	0.9870	0.0130	0.2418	0.1819	0.2629	0.860	0.963
	PSM 2	1.0057	0.0057	0.2412	0.1794	0.2621	0.853	0.965
	PSM 3	0.9941	0.0059	0.2401	0.1796	0.2629	0.861	0.967
0.3	PSM 1	0.9687	0.0313	0.2518	0.1872	0.2647	0.866	0.958
	PSM 2	0.9910	0.0090	0.2405	0.1827	0.2626	0.856	0.968
	PSM 3	0.9777	0.0223	0.2447	0.1832	0.2653	0.851	0.965
0.5	PSM 1	0.9477	0.0523	0.2533	0.1894	0.2659	0.837	0.952
	PSM 2	0.9677	0.0323	0.2473	0.1840	0.2626	0.834	0.962
	PSM 3	0.9511	0.0489	0.2480	0.1846	0.2668	0.837	0.962
0.8	PSM 1	0.9310	0.0690	0.2601	0.1937	0.2697	0.846	0.951
	PSM 2	0.9609	0.0391	0.2469	0.1867	0.2635	0.848	0.962
	PSM 3	0.9434	0.0566	0.2430	0.1874	0.2717	0.858	0.970

표 5. Double propensity score에서 결과변수가 이분형인 경우 처리효과와 표준오차 추정치 및 포함확률(처리할당비율=0.1)

Binary outcome : risk difference								
Correlation	Model	Empirical			SE(Bootstrap)		Coverage probability (Wald type)	
		Mean	Bias	SE(Empirical)	SE(Simple)	SE(Complex)	SE(Simple)	SE(Complex)
0	PSM 1	0.0185	0.0015	0.0365	0.0221	0.0313	0.846	0.947
	PSM 2	0.0200	0.0000	0.0362	0.0216	0.0310	0.824	0.955
	PSM 3	0.0203	0.0003	0.0354	0.0217	0.0310	0.833	0.959
0.3	PSM 1	0.0157	0.0043	0.0351	0.0235	0.0329	0.829	0.949
	PSM 2	0.0176	0.0024	0.0346	0.0227	0.0320	0.847	0.954
	PSM 3	0.0172	0.0028	0.0347	0.0228	0.0325	0.834	0.954
0.5	PSM 1	0.0139	0.0061	0.0354	0.0240	0.0340	0.822	0.947
	PSM 2	0.0184	0.0016	0.0337	0.0230	0.0323	0.843	0.946
	PSM 3	0.0175	0.0025	0.0344	0.0231	0.0333	0.836	0.949
0.8	PSM 1	0.0126	0.0074	0.0351	0.0246	0.0354	0.840	0.949
	PSM 2	0.0172	0.0028	0.0321	0.0234	0.0325	0.872	0.966
	PSM 3	0.0174	0.0026	0.0322	0.0236	0.0344	0.860	0.967

표 6. 성향점수 매칭에서 결과변수가 연속형인 경우 처리효과와 표준오차 추정치 및 포함확률(처리할당비율=0.05)

Continuous outcome : difference in means								
Correlation	Model	Empirical			SE(Bootstrap)		Coverage probability (Wald type)	
		Mean	Bias	SE(Empirical)	SE(Simple)	SE(Complex)	SE(Simple)	SE(Complex)
0	PSM 1	1.0157	0.0157	0.3586	0.3615	0.3775	0.945	0.964
	PSM 2	1.0158	0.0158	0.3507	0.3573	0.3760	0.951	0.958
	PSM 3	1.0111	0.0111	0.3482	0.3570	0.3772	0.951	0.956
0.3	PSM 1	1.0596	0.0596	0.3615	0.3697	0.3800	0.943	0.951
	PSM 2	1.0315	0.0315	0.3552	0.3611	0.3772	0.948	0.961
	PSM 3	1.0248	0.0248	0.3607	0.3616	0.3780	0.947	0.960
0.5	PSM 1	1.0720	0.0720	0.3788	0.3770	0.3823	0.941	0.941
	PSM 2	1.0148	0.0148	0.3608	0.3653	0.3782	0.937	0.957
	PSM 3	1.0185	0.0185	0.3684	0.3661	0.3823	0.929	0.951
0.8	PSM 1	1.1333	0.1333	0.3741	0.3839	0.3854	0.934	0.934
	PSM 2	1.0340	0.0340	0.3440	0.3674	0.3810	0.959	0.967
	PSM 3	1.0360	0.0360	0.3526	0.3694	0.3823	0.962	0.969

표 7. 성향점수 매칭에서 결과변수가 이분형인 경우 처리효과와 표준오차 추정치 및 포함확률(처리할당비율=0.05)

Binary outcome : risk difference								
Correlation	Model	Empirical			SE(Bootstrap)		Coverage probability (Wald type)	
		Mean	Bias	SE(Empirical)	SE(Simple)	SE(Complex)	SE(Simple)	SE(Complex)
0	PSM 1	0.0253	0.0053	0.0494	0.0440	0.0457	0.949	0.952
	PSM 2	0.0224	0.0024	0.0490	0.0432	0.0454	0.944	0.962
	PSM 3	0.0232	0.0032	0.0487	0.0432	0.0454	0.951	0.961
0.3	PSM 1	0.0269	0.0069	0.0483	0.0471	0.0480	0.957	0.962
	PSM 2	0.0249	0.0049	0.0463	0.0453	0.0474	0.958	0.970
	PSM 3	0.0240	0.0040	0.0463	0.0456	0.0477	0.956	0.967
0.5	PSM 1	0.0343	0.0143	0.0482	0.0486	0.0486	0.945	0.937
	PSM 2	0.0240	0.0040	0.0481	0.0468	0.0481	0.950	0.960
	PSM 3	0.0258	0.0058	0.0481	0.0468	0.0488	0.945	0.957
0.8	PSM 1	0.0433	0.0233	0.0469	0.0503	0.0497	0.943	0.941
	PSM 2	0.0278	0.0078	0.0480	0.0482	0.0487	0.953	0.956
	PSM 3	0.0312	0.0112	0.0472	0.0480	0.0500	0.962	0.977

표 8. 성향점수 매칭에서 결과변수가 연속형인 경우 처리효과와 표준오차 추정치 및 포함확률(처리할당비율=0.1)

Continuous outcome : difference in means								
Correlation	Model	Empirical			SE(Bootstrap)		Coverage probability (Wald type)	
		Mean	Bias	SE(Empirical)	SE(Simple)	SE(Complex)	SE(Simple)	SE(Complex)
0	PSM 1	1.0256	0.0256	0.2424	0.2560	0.2638	0.953	0.964
	PSM 2	1.0213	0.0213	0.2421	0.2522	0.2619	0.959	0.959
	PSM 3	1.0118	0.0118	0.2408	0.2523	0.2631	0.959	0.967
0.3	PSM 1	1.0913	0.0913	0.2522	0.2619	0.2659	0.941	0.952
	PSM 2	1.0388	0.0388	0.2415	0.2550	0.2633	0.960	0.963
	PSM 3	1.0344	0.0344	0.2444	0.2553	0.2661	0.965	0.964
0.5	PSM 1	1.1446	0.1446	0.2556	0.2660	0.2676	0.920	0.920
	PSM 2	1.0478	0.0478	0.2486	0.2587	0.2638	0.954	0.961
	PSM 3	1.0443	0.0443	0.2504	0.2591	0.2684	0.946	0.959
0.8	PSM 1	1.2458	0.2458	0.2568	0.2711	0.2692	0.859	0.851
	PSM 2	1.0989	0.0989	0.2520	0.2604	0.2670	0.948	0.943
	PSM 3	1.1000	0.1000	0.2459	0.2617	0.2762	0.946	0.955

표 9. 성향점수 매칭에서 결과변수가 이분형인 경우 처리효과와 표준오차 추정치 및 포함확률(처리할당비율=0.1)

Binary outcome : risk difference								
Correlation	Model	Empirical			SE(Bootstrap)		Coverage probability (Wald type)	
		Mean	Bias	SE(Empirical)	SE(Simple)	SE(Complex)	SE(Simple)	SE(Complex)
0	PSM 1	0.0253	0.0053	0.0364	0.0305	0.0312	0.950	0.951
	PSM 2	0.0228	0.0028	0.0364	0.0298	0.0310	0.942	0.951
	PSM 3	0.0233	0.0033	0.0354	0.0299	0.0309	0.948	0.957
0.3	PSM 1	0.0364	0.0164	0.0344	0.0322	0.0325	0.929	0.929
	PSM 2	0.0257	0.0057	0.0347	0.0316	0.0323	0.949	0.951
	PSM 3	0.0262	0.0062	0.0349	0.0317	0.0325	0.938	0.952
0.5	PSM 1	0.0457	0.0257	0.0346	0.0331	0.0332	0.886	0.892
	PSM 2	0.0317	0.0117	0.0345	0.0320	0.0328	0.924	0.939
	PSM 3	0.0325	0.0125	0.0352	0.0322	0.0335	0.925	0.935
0.8	PSM 1	0.0609	0.0409	0.0337	0.0341	0.0337	0.784	0.776
	PSM 2	0.0391	0.0191	0.0331	0.0328	0.0334	0.909	0.923
	PSM 3	0.0417	0.0217	0.0332	0.0328	0.0349	0.901	0.927

제 4장 실제자료 분석

이 장에서는 유방암 환자의 유방초음파 검사 자료를 사용하여 double propensity score로 처리효과를 추정하고 그 유의성을 판단해보고자 한다. 유의성 판단을 위한 처리효과 추정치의 표준오차를 추정하는 방법으로 앞서 소개한 두 가지 붓스트랩 방법을 이용한다. 모의실험 결과에 의하면 simple 붓스트랩은 과소 추정되는 경향이 있었다. 따라서 실제 분석에서도 simple 붓스트랩을 이용한 표준오차가 complex 붓스트랩을 이용한 표준오차보다 더 작게 추정되는지 확인하고 이로 인해 유의성이 어떻게 달라지는지 확인하고자 한다.

4.1 자료 설명

유방암으로 수술을 받은 환자는 대개 치료 종료 1년 후 재발성 암을 생각해야 한다(Mendelson 1992). 만약 유방초음파 검사를 1년에 한번이상 하는 것이 재발성 암을 예방하는데 도움이 된다면 환자들에게 검사 주기에 대한 조언을 할 수 있을 것이다. 또한 유방암 환자의 재발 위험은 수술 후 2~3년이 가장 많고 그 이후에는 재발 가능성이 상대적으로 적다. 그러므로 분석 등록(enroll)시점과 수술시점의 차이인 수술로부터의 날 수가 짧을수록 재발 위험이 높기 때문에 이를 연구에 고려해야한다. 이 자료는 2011년 1월 1일부터 2012년 12월 31일까지 2년 동안 2011년 1월 1일 이전에 유방암을 진단받아 수술을 받은 환자 3060명을 대상으로 유방초음파 검사를 진행하여 이차성 유방암(second breast cancer)에 대한 병변 검출여부를 기록한 자료이다.

본 연구에서는 이 자료를 통해 유방초음파 검사 빈도가 높은 biannual군과 검사 빈도가 낮은 annual군에서 이차성 유방암에 대한 병변이 검출 될 확률이 차이가 있는지 알아보고, 그 결과가 simple 붓스트랩과 complex 붓스트랩을 사용한 경우에 따라 달라지는지 확인하고자 한다. 이 자료에서 Group은 환자의 검사 날짜를 토대로 처음 검사부터 이후 검사까지의 기간들을 계산하여 평균 F/U기간이 1년 이하인 환자는 biannual군, 약 1년 주기로 시행된 환자는 annual군으로 정의했다. DetectionbyUS는 유방초음파 검사를 통해 2년간 한번이라도 병변이 검출됐다면 그 환자는 병변이 있다

고 했다. 독립변수로는 임상적으로 중요한 Age(연구 시작 시점의 나이), Time_interval_day(수술로부터 유방초음파 검사까지의 날수의 평균)와 OP(수술방법으로 MRM, PM, PM&MRM 중 하나)를 고려하여 분석했다.

표 10 에는 성향점수 매칭 모형을 적합 시키기 위해 사용된 변수와 이에 대한 설명이 있다. 표 11 에는 매칭 전 자료의 각 변수에 대한 기술통계량과 유의확률이 있다. 먼저 Group변수의 군에 따라 환자 수(n)를 나타냈다. 명목형 변수는 각 변수 값에 속한 환자의 수(퍼센트), 연속형 변수는 평균±표준편차로 나타내었다. 유의확률은 두 군에서 변수들의 분포가 다른지를 검정한 결과로 명목형 변수는 chi-square test, 연속형 변수는 유의확률 1은 t-test를 유의확률 2는 Wilcoxon rank-sum test를 수행한 결과이다. 기술통계량을 살펴보면 매칭 전 두 군에서 변수 값들이 상이한 것을 확인할 수 있다. 모든 변수에 대한 유의확률 1,2가 0.0001보다 작으므로 유의수준 0.05하에서 두 군에서 변수들의 분포가 다르다고 할 수 있다.

표 10. 유방초음파 검사 자료의 변수 설명

변수	척도	변수 설명	변수값 설명
Group	명목형	유방초음파 검사 주기에 따른 군	0: Biannual 1: Annual
DetectionbyUS	명목형	유방초음파 검사 기간 동안 병변이 발견된 여부	0: 없다 1: 있다
Age	연속형	연구 시작 시점의 나이(만)	
Time_interval_day	연속형	수술로부터 유방초음파 검사까지의 날수의 평균	
OP	명목형	수술 방법	1:MRM, 2: PM, 3: PM & MRM

표 11. 유방초음파 검사 자료의 기술통계량

변수	전체 환자		유의확률1	유의확률2
	Biannual (n=2390)	Annual (n=670)		
Age	50.68±9.39	55.02±9.51	<0.0001	<0.0001
Time_interval_day	1567.04±731.20	3012.40±1142.17	<0.0001	<0.0001
OP			<0.0001	
MRM	1187(49.67)	464(69.25)		
PM	1098(45.94)	187(27.91)		
MRM + PM	105(4.39)	19(2.84)		

4.2 분석 결과

유방초음파 검사 자료를 성향점수 매칭 모형에 적합 시킨 후 성향점수에 따라 두 군의 환자를 매칭하여 매칭된 자료를 만들었다. 매칭 방법은 NNM, caliper matching을 사용했다. 먼저 매칭된 자료의 각 변수에 대한 기술통계량을 토대로 두 군에서 각 변수의 분포가 유사하도록 매칭이 잘 되었는지 확인한다. 그리고 처리효과를 추정하여 그 유의성을 유의확률과 신뢰구간으로 확인한다.

표 12 에는 NNM을 사용하여 매칭된 자료의 각 변수에 대한 기술통계량과 유의확률이 있다. 기술통계량의 설명은 표 11과 같다. 유의확률은 두 군에서 변수들의 분포가 다른지를 검정한 결과로 명목형 변수는 Bowker's test, 연속형 변수는 유의확률 1은 paired t-test를 유의확률 2는 Wilcoxon signed-rank test를 수행한 결과이다. 기술통계량을 살펴보면 매칭 전 두 군의 변수 값들과 비교했을 때 상이한 정도가 감소된 것을 확인 할 수 있다. Age와 OP에 대한 유의확률 1,2 가 0.05보다 크므로 유의수준 0.05하에서 두 군에서 변수들의 분포가 다르다고 할 수 없다. Time_interval_day의 유의확률 1,2가 0.0001보다 작으므로 유의수준 0.05하에서 두 군에서 변수의 분포가 다르다고 할 수 있다.

표 13 에는 caliper=0.05인 caliper matching을 사용하여 매칭된 자료의 각 변수에 대한 기술통계량과 유의확률이 있다. 기술통계량의 설명은 표 11과 같다. 유의확률의 설명은 표 12와 같다. 기술통계량을 살펴보면 매칭 전 두 군의 변수 값들과 비교했을

때 상이한 정도가 감소된 것을 확인 할 수 있다. 또한 NNM을 사용하여 매칭된 자료와 비교하여도 상이한 정도가 감소된 것을 확인 할 수 있다. 모든 변수에 대한 유의확률 1,2가 0.05보다 크므로 유의수준 0.05하에서 두 군에서 변수들의 분포가 다르다고 할 수 없다.

표 12. NNM 후 유방초음파 검사 자료의 기술통계량

변수	매칭된 환자		유의확률1	유의확률2
	Biannual (n=670)	Annual (n=670)		
Age	54.05±9.20	55.02±9.51	0.0590	0.0577
Time_interval_day	2297.51±784.07	3012.40±1142.17	<0.0001	<0.0001
OP			0.6009	
MRM	446(66.57)	464(69.25)		
PM	206(30.75)	187(27.91)		
MRM + PM	18(2.69)	19(2.84)		

표 13. Caliper matching 후 유방초음파 검사 자료의 기술통계량

변수	매칭된 환자		유의확률1	유의확률2
	Biannual (n=385)	Annual (n=385)		
Age	54.21±9.74	53.83±10.29	0.5812	0.5905
Time_interval_day	2322.17±994.28	2349.98±965.71	0.2300	0.1370
OP			0.4954	
MRM	245(63.64)	247(64.16)		
PM	130(33.77)	124(32.21)		
MRM + PM	10(2.60)	14(3.64)		

표 14 는 double propensity score를 사용하여 추정된 처리효과와 simple, complex 붓스트랩을 통해 추정한 표준오차 그리고 왈드 타입의 유의확률을 나타낸 결과이다. 매칭 방법으로는 NNM, caliper matching 방법이 사용됐으며 표준오차 추정 시 1000번의 붓스트랩을 시행했다. 왈드 타입의 유의확률은 추정된 처리효과를 표준오차로 나눈 값을 누적 표준 정규분포 함수에 넣어 계산했다. 분석 결과를 보면 표준오차는 매칭 방법이 NNM일 때는 simple 붓스트랩의 경우 0.0066, complex 붓스트랩의 경우 0.0167로 추정됐다. Caliper matching일 때는 simple 붓스트랩의 경우 0.0069, complex 붓스트랩의 경우 0.0109로 추정됐다. 매칭 방법에 상관없이 simple 붓스트랩을 이용하여 추정된 표준오차가 complex 붓스트랩을 이용하여 추정된 표준오차보다 더 작게 추정됨을 확인 할 수 있었다. 이를 통해 모의실험 결과와 실제 자료 분석의 결과가 일치함을 확인 할 수 있다. 처리효과는 매칭 방법이 NNM일 때는 -0.0006, caliper matching일 때는 -0.0002로 추정되며 둘 다 음수이고 유사한 값을 가졌다. 유의확률은 모든 경우에서 0.05보다 크므로 유의수준 0.05하에서 처리효과가 존재한다고 할 수 없다. 즉, biannual군과 annual군에서 이차성 유방암에 대한 병변이 검출 될 확률이 차이가 있다고 할 수 없다.

표 14. 매칭 방법에 따른 처리효과 추정치와 붓스트랩 방법에 따른 표준오차 추정치, 유의확률

매칭 방법	처리효과	표준오차		유의확률	
		Simple	Complex	Simple	Complex
NNM (n=670*2)	-0.0006	0.0066	0.0167	0.4638	0.4857
Caliper matching (n=385*2)	-0.0002	0.0069	0.0109	0.4884	0.4927

표 15는 simple, complex 붓스트랩 방법을 통해 추정한 신뢰구간을 매칭 방법별로 나타낸 결과이다. 신뢰구간의 추정은 2.3.3에서 기술된 신뢰구간 추정방법인 왈드 타입 신뢰구간과 백분위수에 근거한 비모수적 방법을 사용했다. 분석 결과를 보면 simple 붓스트랩을 이용하여 추정된 신뢰구간의 길이가 complex 붓스트랩을 이용하여 추정된 신뢰구간의 길이보다 더 짧게 추정됨을 확인 할 수 있다. 또한 두 붓스트랩 방법에서 모든 신뢰구간이 0을 포함하지 않으므로 유의수준 0.05하에서 처리효과가 존재한다고 할 수 없다.

표 15. 붓스트랩 방법에 따른 신뢰구간

	Simple	Complex
NNM		
Wald type 95% CI	(-0.0135, 0.0122)	(-0.0333, 0.0321)
Wald type 99% CI	(-0.0175, 0.0163)	(-0.0436, 0.0424)
Empirical 95% CI	(-0.0135, 0.0129)	(-0.0431, 0.0224)
Empirical 99% CI	(-0.0175, 0.0162)	(-0.0610, 0.0303)
Caliper matching		
Wald type 95% CI	(-0.0137, 0.0133)	(-0.0216, 0.0212)
Wald type 99% CI	(-0.0180, 0.0176)	(-0.0284, 0.0279)
Empirical 95% CI	(-0.0234, 0.0026)	(-0.0161, 0.0261)
Empirical 99% CI	(-0.0260, 0.0078)	(-0.0219, 0.0337)

이를 통해 simple 붓스트랩을 사용해서 표준오차를 추정할 경우 표준오차가 작게 추정되므로 유의확률이 작게 추정되며, 신뢰구간의 길이가 짧아 0을 포함하지 않을 확률이 높은 것을 확인 할 수 있다. 결론적으로 simple 붓스트랩을 사용한 표준오차 추정방법은 실제 유의하지 않은 효과를 유의하다고 결론 내릴 가능성이 존재하는 방법임을 알 수 있다.

제 5장 결론 및 고찰

본 연구에서는 double propensity score를 이용한 처리효과 추정치의 표준오차 추정 방법으로 simple, complex 붓스트랩을 적용하는 것에 대해 연구했다.

현재 제안된 double propensity score에서는 처리효과 추정치의 표준오차의 이론적 추정치가 제시되어 있지 않다. 처리효과 추정을 위해, 대조군을 이용하여 추정된 모델로 처리군이 처리효과를 받지 않았을 때의 결과변수를 추정하고 이를 처리군의 관측값과 비교한다. 모델이 추정되는 군과 처리효과 추정에 사용되는 군이 다르므로 표준오차 추정에 어려움이 존재한다. 따라서 본 논문에서는 표준오차 추정을 위한 방법으로 두 가지 붓스트랩을 제시하여 다양한 상황에서 어떤 방법이 더 정확한 표준오차 추정을 가능하게 하는지에 대해 연구했다.

모의실험을 통해 먼저 double propensity score를 사용하여 처리효과를 추정한 결과 편의가 매우 작은 값을 가지며 추정이 잘 되는 것을 확인 할 수 있었다. 다음으로 double propensity score를 사용했을 때 두 가지 붓스트랩 방법을 통해 추정된 표준오차를 비교한 결과 complex 붓스트랩을 사용하여 추정한 표준오차가 경험적 표준오차와 유사한 값을 가지며 추정이 잘 되는 것을 확인 할 수 있었다. 반면 simple 붓스트랩을 사용하여 추정한 표준오차는 경험적 표준오차와 상이한 값을 가지며 과소추정됐다. 95% 신뢰구간을 이용한 포함확률은 complex 붓스트랩을 사용한 경우에 0.95에 가까워 정확한 추정을 했으나 simple 붓스트랩을 사용한 경우에는 0.85에 가까워 정확한 추정을 하지 못했다. 위와 같은 결과는 다양한 상황을 가정했을 때에도 계속해서 유지 됐다.

모의실험에서 성향점수 매칭을 사용하는 경우에는 simple 붓스트랩과 complex 붓스트랩을 사용할 때 모두 정확한 표준오차 추정을 가능하게 하는 것을 확인 할 수 있었다. 표본의 연관성을 고려할 때 붓스트랩을 사용하여 표준오차를 추정하는 경우 simple 붓스트랩에서와 같이 매칭된 짝들로 이루어진 집합을 붓스트랩 할 수 있다 (Konietschke and Pauly 2014). 성향점수 매칭을 사용할 경우에는 결과변수가 연속형일 때는 근소한 차이이지만 simple 붓스트랩이 complex 붓스트랩보다 표준오차를 추정함에 있어 더 높은 정확도를 보였다. 그러나 앞에서 double propensity score를 사용할 경우에는 추정된 결과변수의 연관성을 가정 할 수 있음에도 불구하고 simple 붓스트랩을 사용하였을 때 표준오차가 과소 추정됐다. 이러한 결과가 나온 이유는

double propensity score를 사용한 경우에는 성향점수 매칭을 사용한 경우와 달리 회귀모형 $m_0(e(\boldsymbol{x}))$를 사용하여 처리군의 r_0를 추정함에 있다. 이로 인해 r_0에 대한 추정치들의 변동이 관측 값들의 변동보다 상대적으로 작아 simple 붓스트랩을 사용하였을 때 표준오차가 과소 추정됐을 것으로 추측된다. Complex 붓스트랩은 처리효과 추정 방법이 달라져도 표준오차를 추정함에 있어 높은 정확도를 유지했다.

실제 자료 분석에서도 double propensity score를 이용한 처리효과 추정치의 표준오차를 추정한 결과, simple 붓스트랩을 이용한 경우가 complex 붓스트랩을 이용한 경우보다 표준오차가 더 작게 추정되어 유의확률에 차이를 보였다. 모의실험 결과를 토대로 complex 붓스트랩으로 추정한 표준오차가 정확도가 높다고 가정한다면 simple 붓스트랩으로 추정한 표준오차로 처리효과의 유의성을 판단할 경우 실제 유의하지 않은 결과를 유의 하다고 잘못 판단할 가능성이 존재한다고 본다.

위 내용들을 종합하여 볼 때 double propensity score를 사용하는 경우 complex 붓스트랩을 이용하여 표준오차를 추정하는 것이 정확도를 높일 수 있는 방법이 될 것으로 기대한다.

참고문헌

Austin, P. C. 2009. “Type I error rates, coverage of confidence intervals, and variance estimation in propensity-score matched analyses". The international journal of biostatistics, 5(1).

Austin, P. C. 2011. “Comparing paired vs non-paired statistical methods of analyses when making inferences about absolute risk reductions in propensity-score matched samples". Statistics in medicine, 30(11): 1292-1301.

Austin, P. C. 2013. “The performance of different propensity score methods for estimating marginal hazard ratios". Statistics in medicine, 32(16): 2837-2849.

Austin, P. C. 2014. “Double propensity-score adjustment: a solution to design bias or bias due to incomplete matching". Statistical methods in medical research.

Austin, P. C., Grootendorst, P., Anderson, G. M. 2007. “A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study". Statistics in medicine, 26(4): 734-753.

Austin, P. C., Small, D. S. 2014. “The use of bootstrapping when using propensity-score matching without replacement: a simulation study". Statistics in medicine, 33(24): 4306-4319.

Gayat, E., Resche-Rigon, M., Mary, J. Y., Porcher, R. 2012. “Propensity score applied to survival data analysis through proportional hazards models: a Monte Carlo study". Pharmaceutical statistics, 11(3): 222-229.

Konietschke, F., Pauly, M. 2014. “Bootstrapping and permuting paired t-test type statistics". Statistics and Computing, 24(3): 283-296.

Mendelson, E. B. 1992. “Evaluation of the postoperative breast". Radiol Clin North Am, 30(1): 107-38.

Rosenbaum, P. R., Rubin, D. B. 1983. “The central role of the propensity score in observational studies for causal effects". Biometrika, 70(1): 41-55.

Rosenbaum, P. R., Rubin, D. B. 1985. “The bias due to incomplete matching". Biometrics, 41(1): 103-116.

Schafer, J. L., Kang, J. 2008. “Average causal effects from nonrandomized studies: a practical guide and simulated example". Psychological methods, 13(4): 279-313.

ABSTRACT

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

Lim, So Jung
Dept. of Biostatistics and Computing
The Graduate School
Yonsei University

Propensity score matching is often used to minimize the confounding that occurs in observational studies to estimate the treatment effect. When no suitable control subject is available, some treated subjects are excluded from the matched sample, which leads to biased effect estimation. Rosenbaum and Rubin (1985) introduced the term 'bias due to incomplete matching' to describe this. Double propensity score adjustment (Austin 2014) is an analytic solution to address bias due to incomplete matching. When using double propensity score adjustment, however, theoretical estimation of the standard error has not been proposed; hence, it is difficult to estimate the standard error of the estimated treatment effect.

Therefore, we propose two bootstrap methods (Austin and Small 2014) to estimate the standard error. The first method is the simple bootstrap, which involves drawing bootstrap samples from matched pairs in the matched sample. The second method is the complex bootstrap, which involves drawing bootstrap samples from the original sample and estimating the propensity score separately in each bootstrap sample.

Through simulation, we examined the performances of two bootstraps when using double propensity score adjustment with a series of Monte Carlo simulations. Covariates were generated by considering the correlations of covariates and whether the variable affects the treatment selection and/or the

outcome. The proportions of subjects who were treated were 0.05 and 0.1. Additionally, we examined the results modifying the propensity score models. Considering continuous and binary outcomes, we simulated 1000 datasets, each consisting of 3000 subjects. In addition, we applied two bootstrap methods to breast cancer patients with US data and compared the results of the simple bootstrap with that of the complex bootstrap.

The simulation results of this study showed that in the various scenarios, the complex bootstrap resulted in estimates of the standard error that were closer to the empirical standard error of the sampling distribution of estimated effects. On the contrary, the estimates of the standard error using the simple bootstrap were underestimated. In the actual data analysis, the estimate of the standard error using the simple bootstrap was smaller than that using the complex bootstrap. The empirical coverage rates using the complex bootstrap tended to be closer to the advertised rate of 0.95. Additionally, the simple bootstrap tended to result in lower empirical coverage rates compared with the complex bootstrap.

In conclusion, the complex bootstrap can be an alternative for accurate variance estimation when using double propensity score adjustment.

Key words : propensity score matching, standard error, double propensity score adjustment, bootstrap, Monte Carlo simulations