교육과정평가연구
The Journal of Curriculum and Evaluation
2019, Vol. 22, No. 2, pp. 269~291

경향점수를 활용한 인과효과 추정 방법 비교[1] : 대응, 가중, 층화, 이중경향점수 보정

장 현 진(한국교육과정평가원 전문연구원)*
정 혜 경(한국교육과정평가원 부연구위원)
민 경 석(세종대학교 교수)**

<요 약>

본 연구에서는 경향점수 활용 방식으로 최근 개발된 이중경향점수 보정 방법을 소개하고, 기존의 경향점수를 활용한 대응, 가중, 층화 방법과 비교하여 어떠한 특징과 차별점이 있는지를 논의하는데 그 주요 목적을 두었다. 또한, 실증자료 분석을 바탕으로 통계적 추론을 위한 관심 모수에 따라 경향점수를 다르게 활용할 경우, 추정 가능한 인과효과 모수(ATE 또는 ATT)와 그에 따른 평균차이, 효과크기, 효과크기의 신뢰구간 등을 바탕으로 주요 결과를 논의하였다.

네 가지 경향점수 활용방법에 따른 비교·분석을 위해 경기교육종단연구(GEPS2012) 자료를 활용하여 수학 사교육 효과를 검증하였다. 분석 결과로서 대응 방법이 상대적으로 자료의 특성에 따라 활용이 제한적이며, 그에 따른 결과의 해석에서도 주의가 필요한 것으로 확인되었다. 이중경향점수 보정 방법은 교육적으로 특수한 집단을 관심 모집단으로 삼거나 처지·비교 집단이 매우 다른 분포를 보일 때 처치집단에 대한 ATT 인과효과 추정에 유용하다.

주제어: 인과효과, 경향점수, 이중경향점수 보정, ATT

Ⅰ. 서론

다양한 교육활동의 효과성을 연구하는 것은 투입된 교육의 내용과 그 결과의 인과관계를 추론하여 규명하는 것과 밀접하게 연관된다. 예를 들어, 자유학기제 도입이 학생들의 진로선택에 긍정적인 영향을 미치는지, 사교육이 학업성취도 향상에 효과적인지 등에 대한 연구는

1) 이 논문은 첫 번째 저자의 석사학위 논문을 일부 발췌하여 수정·보완한 것이며, 초기 원고는 2018 한국교육평가학회(한국교육과정평가원 공동 개최) 춘계학술대회에서 발표되었음을 밝힘

* 제1저자, jangzin@kice.re.kr
** 교신저자, minkyungseok@sejong.ac.kr

특정 교육처치가 갖는 인과효과를 추정하는 데 주요 관심이 있다. 일반적으로, 인과관계의 성립을 위해서는 1) 원인변수의 시간적 선행, 2) 원인 및 결과변수 간 상관(correlation), 3) 혼재요인의 통제라는 세 가지 조건이 모두 충족되어야 한다(Shadish, Cook & Campbell, 2002). 앞선 두 가지 조건을 충족하기는 상대적으로 쉬우나, 원인변수와 결과변수에 동시에 영향을 미치는 혼재요인에 대한 통제는 엄격한 수준에서 불가능하다. 특히, 교육연구를 비롯한 사회과학 연구에서는 그 가능성이 더욱 낮다.

연구유형 중 변수 간의 인과관계를 가장 확실하게 추론할 수 있는 방법은 임의할당(random assignment)을 통해 사전 동등성을 확보하는 실험연구(experimental design)를 실시하는 것이나, 이는 사회과학 분야에서 보편화되기 어려우므로 일반적으로 이미 수집된 성취도 자료, 설문조사 등의 관찰자료를 활용하여 연구를 진행한다. 이 경우 사전 동등성이 확보되어있지 않기 때문에 선택편의(selection bias)의 문제가 발생한다. 즉, 관찰자료에 근거한 연구로 대표되는 비실험 연구에서는 집단별 구성원의 특성이 체계적으로 다를 가능성이 매우 높다(정혜경, 2012).

선택편의로 인해 발생할 수 있는 문제를 최소화하기 위해 다양한 통계방법들이 제시되어 왔으며, 그 중 하나가 경향점수(propensity score) 방법이다. Rosenbaum & Rubin(1983)에 의해 처음 제안된 경향점수는 관찰된 공변인들이 주어졌을 때, 해당 공변인 값들을 가지는 각 사례가 처치집단에 배치될 조건부 확률로 정의된다. 이는 로지스틱 회귀모형(logistic regression), 프로빗 회귀모형(probit regression) 등으로 추정된 후(김준엽, 정혜경, Seltzer, 2008), 대응(matching), 가중(weighting), 층화(subclassification)를 비롯하여 최근 논의되는 이중경향점수 보정(double propensity-score adjustment) 방법(Austin, 2017) 등을 통해 인과효과 추정에 활용된다.

이 연구는 경향점수 활용방법에 따른 인과효과 추정치의 특성을 확인하기 위해, 실증적 자료를 활용하여 경향점수 대응, 가중, 층화, 이중경향점수 보정 방법을 적용한 결과에 어떠한 차이가 있는지 이론적·경험적 관점에서 비교한다. 특히 본 연구에서는 교육학 연구에서는 상대적으로 덜 알려진 이중경향 보정 방법을 소개하고, 기존의 경향점수를 활용한 대응, 가중, 층화 방법과 비교하여 어떠한 특징과 차별점이 있는지를 논의하는데 주요 목적을 두었다.

인과효과 연구에서 관심 모집단의 특성과 연관지어 분석 결과를 해석하고 추론하는 것은 중요하며, 이에 연구자는 통계적 추론(inference)를 위한 관심 모수(estimand; parameter of interest)를 정의하고 이에 대한 추정치(estimates)를 산출하게 된다. 일반적으로 인과추정의 관심 모수인 평균 처치효과(ACE: Average Causal Effect)는 크게 두 가지로 구분하는데, 하나는 '모집단 전체'에 대한 평균 처치 효과(ATE: Average Treatment Effect)이고 다른 하나는 '처치집단'에 대한 평균 처치효과(ATT: Average Treatment effect for the Treated)이다(Schafer & Kang, 2008; Morgan & Winship, 2007). 이 연구에서는 이와 같은 맥락에서 이중경향

점수 보정 방법이 어떻게 처치집단에 대한 인과 효과(ATT) 추정에 유용한지를 이론적 검토와 실증 자료 분석에 기반하여 설명하고자 시도하였다. 일반적으로 평균 처치효과를 추정하기 위해서는 원자료의 모든 사례가 활용되는 반면, 처치집단에 대한 평균 처치효과 추정에서는 인과효과 추정에 활용되는 처치집단의 범위가 중요한 의미를 갖는다. 다만, 구분되는 두 가지 인과효과(ATE와 ATT)의 설정은 연구자의 관심뿐만 아니라 자료구조(사전 동등성, 처치집단의 사례수 등)에 영향을 받는다. 따라서 본 연구에서는 실증 자료 분석을 통해 네 가지 경향점수 활용방법(대응, 가중, 층화, 이중경향점수 보정)에 따라 추정 가능한 인과효과 모수와 그에 대한 평균차이, 효과크기, 효과크기의 신뢰구간 등을 기반으로 한 인과효과 추정 결과의 해석을 중심으로 방법론적 논의를 제시했다.

Ⅱ. 이론적 배경

1. 경향점수 방법

가. 경향점수의 개념

경향점수를 활용한 인과효과 추론 방법은 Rubin(1974)의 인과모형(RCM: Rubin's causal model)을 따른다. 이 모형에서 처치효과는 '특정한 사례가 처치를 받았을 때의 잠재적 결과(potential outcomes)와 처지를 받지 않았을 때 보이게 될 잠재적 결과의 차이($Y_i^t - Y_i^c$)'로 정의된다(Holland, 1986). 그러나 현실에서는 두 잠재적 결과 중 처치 배정에 따라 어느 하나만 관찰되기 때문에 나머지 하나는 항상 결측치(missing data)가 된다. Holland(1986)는 이를 '인과효과 추론의 근본적인 문제(fundamental problem of causal inference)'라고 하였다. 결국 교육연구를 비롯한 사회과학 연구에서 처치 및 비교집단의 평균차이는 서로 다른 특성을 갖는 두 집단의 평균차이가 된다. 이 때 관심 추정 모수 중 전체 모집단에 대한 처치효과인 ATE는 $E[Y_i^t - Y_i^c]$로 표현되며, 처치집단에 대한 처치효과인 ATT는 $E[Y_i^t - Y_i^c \mid T_i = 1]$로 구분된다(Schafer & Kang, 2008).

이러한 모수를 추정하는데 있어 이상적인 방법은 임의할당(random assignment)을 통해 집단간 사전 동등성을 확보하는 것이다. 그러나 대부분 비실험연구(non-experimental design) 또는 준실험연구(quasi-experimental design)로 진행되는 사회과학 연구의 경우, 엄격한 수준에서 임의할당이 이루어질 수 없으므로 선택편의(selection bias)가 발생하게 된다.

실험적 통제가 어려운 사회과학 연구에서 선택편의의 문제를 해결하고 보다 타당한 연구결

과를 도출하기 위해서는 통계적 통제가 유용하다. 이에 도구변인법, 회귀-불연속 설계, 이중차분법 등의 다양한 통계 방법들이 제안되어 왔으며, 그 중 하나가 경향점수 방법이다. 경향점수란 사례들이 처치집단이나 비교집단에 속할 가능성에 영향을 주는 모든 변수들, 즉 공변인(covariates)의 값이 주어졌을 때, 해당 공변인 값을 가진 사례가 처치집단에 배치될 조건부 확률이다(Rosenbaum & Rubin, 1983, 식(1) 참조).

$$ps(x) = \Pr(T_i = 1|\boldsymbol{x}) \quad (1)$$

여기서 ps(x)는 경향점수, $T_i = 1$ 은 처치집단에의 할당을 의미하며, $\boldsymbol{x}$는 개별 사례가 가지고 있는 각 변수들의 벡터 정보로 처치에 선행하는 공변인이다(Becker & Ichino, 2002). 경향점수는 로지스틱 회귀모형, 프로빗 모형 등을 통해 추정될 수 있으며, 이를 대응, 층화, 가중 등의 방법을 통해 인과효과 추정에 활용한다(민경석, 2008; Hirano & Imbens, 2001; Rosenbaum & Rubin, 1984; Rosenbaum & Rubin, 1985).

나. 경향점수 활용방법

경향점수 활용방법은 처치집단과 대응되는 비교/통제집단을 설정하는 통계적 방법으로, 비교적 널리 알려진 대응, 가중, 층화 세 가지 절차를 먼저 논의하고자 한다. 특히 각각의 방법이 앞서 언급한 ATE와 ATT를 추정하는 데 있어 어떠한 연관이 있는지를 중요하게 다루고자 한다.

1) 대응

집단간 동등성 확보를 위하여 가장 보편적으로 사용되는 대응 방법은 경향점수가 동일하거나 유사한 처치집단의 사례와 비교집단의 사례를 한 쌍으로 묶는 방법이다. 이 때, 대응되지 않은 사례들은 분석대상에서 제외된다(Rosenbaum & Rubin, 1983). 결국 유사한 경향점수를 갖는 사례들이 새로운 처치 및 비교집단을 구성하게 되며, 두 집단은 유사한 경향점수 분포를 보이게 된다. 처지집단과 비교집단의 사례를 짝짓는 방법에 따라 일대일/일대다 대응으로, 중복 허용여부에 따라 복원/비복원 대응 등으로 구분하며, 짝을 짓는 기준에 따라 최근린/라디우스/최적화 대응 등으로 구분할 수 있다. 그 중 이 연구에서 활용한 최근린(nearest neighbor matching) 대응은 그리디(greedy) 알고리즘을 이용하여 처치집단의 사례 중 가장 유사한 경향점수를 가지는 비교집단의 사례를 일대일로 짝 짓는 방법이다. 이 방법은 처치집단의 사례를 모두 분석대상에 포함시킬 수 있다는 장점이 있으나, 가장 근접한 경향점수를 갖는다고 판

단한 처치집단의 경향점수와 비교집단의 경향점수가 큰 차이를 보일 수 있다는 한계를 갖는다(최강식, 2007). 이에 따라 처치집단과 비교집단의 경향점수 차이를 일정수준으로 유지하기 위하여 캘리퍼(caliper)를 설정한 대응방법을 적용하였다(식(2) 참조).

$$C(P(x)_t) = \| ps(x)_t - ps(x)_c \| < \epsilon, \quad c \in N_0 \tag{2}$$

$ps(x)_t$: 처치집단 사례의 경향점수
$ps(x)_c$: 비교집단 사례의 경향점수

최근린 대응에서는 처치집단 사례의 $ps(x)_t$와 가장 작은 차이를 보이는 $ps(x)_c$를 갖는 비교집단의 사례가 짝을 이루며, 캘리퍼는 대응 사례 간 차이의 범위를 설정하여 해당 범위 안에서 대응 짝을 찾도록 강제함으로써 최근린 방식에 비해 좀 더 엄격하다 할 수 있다. 즉, 실험설계와 유사한 상황을 만들어 낼 수 있다는 장점이 있다. 또한, 통계모형에 대한 의존성을 낮추고, 극단적인 추정을 사전에 예방할 수 있으며, 선택편의를 효율적으로 축소할 수 있다(정혜경, 2012). 그러나 자료의 구조나 변수의 유형에 따라 제한을 많이 받는다는 한계를 가진다.

관심 모수 추정과 관련하여 경향점수에 기반한 대응 방법을 통해 ATE와 ATT 추정 가능 여부는 경향점수 분포의 겹침 정도와 관련이 있으며, 대응된 자료가 각각 전체 모집단과 처치집단의 특성을 대응 후에도 유지하고 있어야 한다(Schafer & Kang, 2008). 즉, 두 집단간 분포의 불균형(imbalance) 정도가 크지 않을 경우 대응에 기반한 모수 추정치의 해석이 용이할 수 있으며, 특히 대응 방법에 기반한 ATT 추정이 가능하기 위해서는 대체로 비교집단의 규모가 충분히 클 뿐만 아니라 '전형적으로 처치를 받을만한 특징을 지닌' 집단을 비교집단에서 포함하고 있어야 한다.

2) 가중

가중 방법은 경향점수의 역수를 가중치로 적용하여 처치집단과 비교집단에 배치될 확률을 동일하게 적용하는 방식으로, 이를 활용하면 공변인의 균형을 이루는 유사 모집단을 형성할 수 있다(Morgan & Todd, 2008). 일반적인 표집설계의 맥락에서 살펴보면 처치집단에는 경향점수가 높은 사례가, 비교집단에는 경향점수가 낮은 사례가 과대표집 되었을 가능성이 크므로, 과대표집에 대한 균형을 맞추고자 각 집단에 다른 가중치를 부여하여 처치효과를 추정한다(Hirano & Imbens, 2001; Austin, 2011). 이 때 가중치(ω_i) 부여 방식은 연구자의 연구 목적과 관련하여 ATE와 ATT 중 어떠한 인과효과 모수를 추정하고자 하는지에 따라 다를 수 있다. 다음 식 (3)은 ATE를 추정할 때의 가중치 부여 방식으로, 경향점수가 높을수록 처치집단

에 속한 사례는 작은 가중치를, 비교집단의 경우 큰 값의 가중치를 갖게 함으로써 모집단 전체의 특성을 대변할 수 있도록 장치한다.

$$\omega_i = T_i \frac{1}{ps(x)_i} + (1 - T_i) \frac{1}{1 - ps(x)_i} \quad (3)$$

T_i= 처치 여부

$ps(x)_i$ = i번째 사례에 대한 경향점수

반면, 연구자가 ATT를 추정하고자 한다면 처치집단에서는 경향점수 가중치를 부여하지 않고 1을 일관적으로 부여하는 대신, 비교집단에 대해 $ps(x)_i/(1-ps(x)_i)$의 가중치를 적용하게 된다(Guo & Fraser, 2010, pp. 161-162; Morgan & Winship, 2007, p. 103). 이 방식은 ATT를 추정할 경우 경향점수 가중 방식을 활용하여 처치를 받을 확률이 큰 비교집단 사례에 대해 좀 더 큰 비중의 가중을 부여하게 된다.

가중 방법은 표집된 사례 수를 거의 그대로 활용할 수 있고, 원래의 자료 구조를 그대로 활용할 수 있다는 장점이 있으며, 특히 특정 구간에서 데이터가 충분치 않을 경우 ATE 추정에 유용하다(Morgan & Winship, 2007, pp. 98-100). 즉, 경향점수의 역확률을 활용하여 통계적 의미에서 실험적 설계 상황을 설정하게 함으로써 데이터가 관심 모집단 특성을 갖추도록 하는데 용이하다. 그러나 극단적 가중치의 영향을 받을 수 있다는 한계가 존재하는데, 이를 보완하는 방법으로 10보다 큰 가중치는 제외하고 분석하거나 경향점수의 상·하위 각 2.5%에 해당하는 사례를 제외하고 분석하는 방법 등이 소개되었다(Rubin, 2001).

3) 층화

층화 방법은 추정된 경향점수를 이용하여 사례들을 동질적인 집단으로 나누어 층내, 층간 분석을 통해 인과효과를 추정하는 방법이다. 이 방법은 추정된 경향점수를 정렬하여 몇 개의 하위집단으로 구분한 후 하위집단 내 또는 전체에 대해 인과효과를 추정한다. 전체 인과효과의 경우 가중평균 방법(각 층에 속한 전체 사례수에 비례하는 가중치 설정)을 사용하여 각 하위집단에서 추정된 처치효과를 전체 처치효과로 통합한다(Rosenbaum & Rubin, 1984). 또는 각 하위집단의 처치효과 추정치를 층에 포함된 처치 사례 수에 비례하게 가중치를 부여하여 통합한다(장은진 외, 2013). 추정 모수와 관련하여 전자는 ATE를, 후자는 ATT를 추정하게 된다(Schafer & Kang, 2008). 층화 방법 적용 시, 보편적으로 5~10개의 하위집단을 구성하는데, 이는 경향점수를 5개 하위집단으로 층화했을 때 경향점수 모형에 포함된 관찰된 혼재요인으로 인한 편의(bias)를 90% 가까이 제거할 수 있음을 확인한 Rosenbaum & Rubin(1984)의

연구에 근거한다.

층화 방법은 구분된 층에 따른 차별적인 처치효과를 산출할 수 있으므로 보다 많은 정보를 확인할 수 있는 장점이 있지만, 상대적으로 많은 사례 수가 필요하며 자료의 구조나 연구 설계가 복잡할 때에는 적용이 어렵다는 한계가 있다(이진실, 2016).

다. 이중경향점수 보정

여기에서는 경향점수 활용방법 중 비교적 최신 방법에 해당하는 이중경향점수 보정 방법에 대해 좀 더 심층적으로 논의하고자 한다.

1) 개념

앞서 언급한 경향점수 대응 방법은 대응의 조건을 엄격하게 할수록 대응에서 탈락하는 사례의 수가 커진다는 한계를 갖는다. 연구 목적이 처치집단 전체를 대상으로 처치효과를 알아보고자 하는 것일 때 '처치를 받을 만한' 가장 전형적인 사례가 탈락되는 상황이 발생할 여지가 있으며(임소정, 정인경, 2017), 이 경우 결과 추정치를 바탕으로 어떠한 모집단에 대한 통계적 추론이 적절한지가 모호해 진다. 이러한 맥락에서 논의되는 것이 Rosenbaum & Rubin (1985)이 언급한 '불완전 대응에 기인한 편의(the bias due to incomplete matching)'로, 타겟 편의(target bias) 또는 설계 편의(design bias)라고도 일컫는다(Austin, 2017). 이는 처치집단에 속한 사례의 일부가 분석에 포함되지 못할 경우 생기는 편의를 의미하며, 인과효과 추정치에 기반한 결과의 해석 및 일반화에 대한 문제가 대두된다. Austin(2017)은 불완전 대응에 기인한 편의를 감소시키는 방법으로 이중경향점수 보정 방법을 제안하였다. 이중경향점수 보정 방법은 일차적으로 대응 방법을 적용하여 처치집단에서 대응된 짝에서 발생한 편의를 조정하기 위해 경향점수를 이용하여 결과 값에 추가적인 보정을 가한다. 처치집단의 모든 사례를 포함하며, 비교집단의 모든 사례는 통계적 보정에 의한 추정치로 이루어진다는 특징을 갖는다. 따라서 이중경향점수 보정은 ATT 추정에 초점을 둔 경향점수 활용 방법이다.

좀 더 구체적으로 살펴보면, 이중경향점수 보정 방법은 대응 이후 한 번의 통계적 조정 과정을 더 거친다는 특징을 갖는다. 경향점수 대응 후 후속 회귀 조정을 실시하여 추가적인 보정을 가하는 방법에 대한 논의는 Rubin & Thomas(2000)의 연구를 주목해 볼 필요가 있다. Rubin과 Thomas는 경향점수 대응과 예측변수를 연속적인 결과변수로 설정하기 위한 추가 조정을 실시하는 방법에 대한 연구로 잔류혼재요인에 의한 편의를 줄이기 위해 경향점수 대응 표본 내에서 회귀 조정을 사용하는 방법과 경향점수와 예측된 주요 공변수를 대응시키는 방법을 비교하였다. 연구 결과, Rubin은 대응과 후속 회귀 조정을 조합하는 방법이 둘 중 하나의 방법을 선택적으로 사용하는 것보다 나음을 제안하였다. 또한, 이중경향점수 보정 방법

과 접근이 유사한 Abadie & Imbens(2011)의 연구에서는 대응된 데이터에서 누락된 잠재적 결과를 추정하기 위해 회귀모형을 사용하였으며, 이러한 방법이 단순한 대응 방법에 비해 편의를 감소시킴을 증명하였다.

2) 추정 방법 및 장점

이중경향점수를 활용하여 처치효과를 추정하는 방법은 다음과 같다. 첫째, 로지스틱모형으로 경향점수를 추정하고, 처치집단 내 모든 사례에 대해 비교집단 중 일부 사례와 매칭짝을 대응시킨다. 둘째, 새로이 구성된 매칭 집단에서 비교집단 사례를 사용해서 식(4)와 같이 비교집단의 경향점수 $ps(x)$만을 공변인으로 하여 성과변수를 예측하는 단순 회귀모형인 $m_0(ps(x))$를 추정한다. 이 때 성과변수의 종류에 따라 선형 회귀모형, 로지스틱 회귀모형 등 다양한 모형을 적용시킬 수 있다.

$$m_0(ps(x)) = E(Y_i^c | ps(x)) = E(Y_i^c | T=0, ps(x)) = E(Y | T=0, ps(x)) \quad (4)$$

셋째, 위 모형을 처치집단에 적용하여 처치집단의 사례가 처치를 받지 않았을 때의 잠재적 성과변수인 Y_i^c의 추정치를 얻는다. 즉, $m_0(ps(x))$로 추정한 회귀식에 처치집단 각 사례의 값을 대입하는 통계적 절차를 통해 처치집단 사례가 비교집단에 소속될 경우의 가상적 성과변수 추정치를 도출한다. 처치집단의 사례가 처치를 받았을 때의 성과변수 Y_i^t은 처치집단에서 관찰된 성과변수 값을 그대로 사용한다. 이후 처치집단에서 얻어진 Y_i^t와 Y_i^c에 대한 추정치 차이의 평균을 계산하여 식(5)로 처치효과를 추정한다.

$$\text{평균처치효과(ATT)} = \frac{1}{n}\sum_{i=1}^{n}(Y_i^t - m_0(ps(x_i)_t)) = \frac{1}{n}\sum_{i=1}^{n} Y_i^t - \frac{1}{n}\sum_{i=1}^{n} m_0(ps(x_i)_t) \quad (5)$$

$$= E(Y_i^t) - E(m_0(ps(x_i)_t))$$

$ps(x_i)_t$= 처치집단 i번째 사례의 경향점수 추정치

Y_i^t= 처치집단의 성과변수

$m_0(ps(x_i)_t)$= 처치집단의 i번째 사례가 처치를 받지 않았을 경우의 잠재적 비교변수 Y_i^c의 추정치

결국, 이중경향점수 보정 방법을 활용한 인과효과 추정에서는 관찰값 Y_i^t과 추정값 $\widehat{Y_i^c}$이

모두 처치집단으로부터 산출되기에 대응에 따른 오차요인이 제거된다.

인과효과 연구에 있어 경향점수가 극단적으로 높거나 낮은 처치집단 내 사례를 포함하는 것은 매우 중요한 의미를 갖는다. 만약 그러한 사례를 포함시키지 않는다면 처치집단 내 일부 그룹에 대한 결과가 배재되는 등 편의가 발생하게 된다. 이러한 이유에서 이중경향점수 보정 방법은 편의를 최소화하고 처치집단의 특성을 누락 없이 반영하는 통계적 접근이라고 볼 수 있다.

Austin(2017)의 연구에서는 몬테 카를로 시뮬레이션(monte carlo simulation)을 통해 이중경향점수의 효과성을 검증하였으며, 연구 결과에 기반하여 이중경향점수의 두 가지 장점을 제시한다. 첫째, 이중경향점수의 활용은 불완전한 대응으로 인한 편의를 줄인다. 최근린법 또는 최적화법으로 추정된 경향점수를 사용하고 후속 회귀 보정을 가할 때, 즉 이중경향점수를 적용하였을 때 후속 회귀계수가 최소 편의로 추정됨을 확인하였다. 이는 전체대응을 통해 모든 처치의 사례를 분석에 포함시켰기 때문이다. 둘째, 이중경향점수는 상이한 경향점수를 가진 짝이 대응되었음을 고려하여 추가적인 보정을 가하기 때문에 잔류혼재 요인(residual confounding errors)이 줄어든다. 한편, 비교집단의 결과가 추정치에 해당하므로 통계모형에 대한 의존도가 높다는 한계를 갖는다.

종합적으로 보았을 때, 이중경향점수 보정 방법은 연구자의 연구 설계의 결과로 발생하는 처치집단 사례수의 손실 등 편의를 일으킬 요인 없이 비교집단의 성과변수 전체를 사용하는 통계모형에 의해 산출하므로(Austin, 2017), 두 집단의 분포가 크게 다르고 비교집단의 사례수에 대응하는 처치집단의 사례수가 충분하지 않은 상황에서 유용하게 활용할 수 있을 것으로 기대된다.

Ⅲ. 연구 방법

1. 분석자료

이 연구는 경향점수를 활용한 인과효과 추정을 위하여 경기도교육연구원에서 실시한 경기교육종단연구(GEPS2012)의 1~3차년도 자료를 사용하였다(경기도교육연구원, 2017; 성기선 외, 2013). 네 가지 경향점수 활용방법을 적용하여 고등학생의 수학 사교육 효과성을 검증하는 본 연구에서는 활용하는 변수에 대해 결측치를 포함하고 있는 경우 분석 대상에서 제외하였으며, 최종적으로 고등학생 3,244명의 자료를 분석에 활용하였다.

이 연구는 성과변수로 고등학생의 '3차년도 수학성취도'를, 처치변수로 '사교육 참여 여부'를 사용하였다. 경향점수를 추정하기 위한 공변인에 해당하는 예측변수는 월평균 가구소득,

부모학력, 성별, 수학수업이해정도, 수학수업태도, 수학교과흥미, 2차년도 수학성취도 등의 7개 변수이다. 예측변수 선정기준은 처치여부가 정해지기 전에 측정된 변수들을 선정해야한다는 '처치 전 기준(pretreatment criterion)'과 처치변수의 원인인 동시에 성과변수의 원인인 변수들을 선택해야 한다는 '공통원인 기준(common cause criterion)'을 따랐다(김지현, 2016). 이에 기준하여 1·2차년도 자료를 사용하였으며, 종속 및 처치변수와의 상관정도를 고려하였다. 주요변수의 빈도 및 기술통계는 <표 Ⅲ-1>과 같다.

<표 Ⅲ-1> 전체자료 주요변수 기술통계

구분	변수[2]	사례수	평균	표준편차	최솟값	최댓값
성과변수	3차년도 수학성취도	3,244	40.48	24.89	0	100
처치변수	수학 사교육 참여 여부	3,244	0.44	0.50	0	1
예측변수	월평균 가구소득	3,244	5.40	2.38	1	11
	부모학력	3,244	2.87	0.96	1	6
	성별(남학생=1)	3,244	1.53	0.50	1	2
	수학수업이해정도	3,244	2.92	1.45	1	5
	수학수업태도	3,244	3.09	1.09	1	5
	수학교과흥미	3,244	2.89	1.01	1	5
	2차년도 수학성취도	3,244	36.68	21.88	0	100

2. 경향점수 추정 모형

경향점수 추정을 위한 로지스틱 회귀모형은 식(6)과 같다. 이는 처치변수와 성과변수에 영향을 주는 7개의 예측변수를 투입하여 설정한 최종 모형이다.

$$\eta_i = \beta_0 + \beta_1(\text{월평균 가구소득})_i + \beta_2(\text{부모학력})_i + \beta_3(\text{성별})_i + \beta_4(\text{수학 수업이해정도})_i + \beta_5(\text{수학수업태도})_i + \beta_6(\text{수학교과흥미})_i + \beta_7(\text{2차년도 수학성취도})_i \quad (6)$$

식(6)의 연구모형에 의해 추정된 경향점수는 식(7)과 같다.

$$\text{경향점수} = \Pr(T_i = 1|X_p) = \frac{e^{\eta_i}}{1+e^{\eta_i}} \quad (7)$$

T_i: i번째 학생의 수학 사교육 참여 여부(참여=1, 미참여=0)

X_p: 총 7개의 공변인$(p = 1, 2, \cdots, 7)$

2) 변수 중 수학 사교육 참여 여부, 성별은 이분변수에 해당함

3. 분석절차

경향점수 활용방법에 따른 인과효과 추정치를 비교·분석하는 구체적인 절차는 다음과 같다.

1단계: 경기교육종단연구 1~3차년도 자료를 활용하여 연구에서 사용하는 변수에 대한 모든 값이 주어진(경향점수 추정이 가능한) 사례들로 분석 자료를 구성한다(전체자료: 3,244명).

2단계: 전체자료에 로지스틱 회귀모형을 적용하여 경향점수 추정치를 산출한다.

3단계: 추정된 경향점수에 대해 네 가지 방법(대응, 가중, 층화, 이중경향점수 보정)을 활용하여 인과효과를 추정한다.

3-1단계(대응): 추정된 경향점수에 대해 최근린법을 적용하여 대응(일대일, 비복원, caliper =0.1)한 후, 처치 및 비교집단 간 동등성 정도를 파악하고, 사교육효과 추정을 위한 독립표본 t검정을 실시한다.

3-2단계(가중): 추정된 경향점수에 대해 처치 및 비교 집단 각각의 역확률 가중치를 적용한 후, 처치 및 비교집단 간 동등성을 확인하고 사교육효과 추정을 위한 독립표본 t검정을 실시한다(ATE 추정).

3-3단계(층화): 추정된 경향점수에 따라 5개의 하위집단을 구성한 후, 하위집단 내 처치 및 비교집단 간 동등성을 확인하고 각 층별 처치집단의 사례수 비율에 따른 가중평균을 적용하여 사교육효과 추정을 위한 독립표본 t검정을 실시한다(ATT 추정).

3-4단계(이중경향점수 보정): 추정된 경향점수에 대해 이중경향점수 보정을 가한다. 즉, 최근린 전체대응을 실시한 후, 비교집단의 경향점수를 공변인으로 성과변수를 예측하는 단순 선형회귀모형을 적용하여 대응된 처치집단만을 가지고 처치집단에 경향점수를 대입하여 처치를 받지 않았다는 가정된 상황에서의 추정치, 즉 잠재적 비교집단의 추정치를 얻는다. 이후, 사교육효과 추정을 위한 독립표본 t검정을 실시한다(ATT 추정).

4단계: 각 활용방법에서 도출된 사교육 및 비사교육집단의 평균, 표준편차에 근거하여 효과크기와 효과크기의 신뢰구간을 산출한다.

5단계: 4단계의 결과값을 경향점수 활용방법(대응, 가중, 층화, 이중경향점수 보정)에 따라 비교한다.

동등성 진단을 위해서 경향점수 및 공변인 기술통계치, 표준화 차이계수 등의 통계치를 확인하며, 보다 직관적인 확인을 위하여 상자도표(box plots)를 제시한다. 또한, 처치효과 추정을 위해 두 집단의 평균을 비교하기 위한 t검정을 실시하고 효과크기를 제시한다. 이 연구의 목적이 사교육효과 추정이 아닌 경향점수 활용방법에 대한 방법론적 비교에 있으므로, 통계적 유의성보다는 효과크기(effect size, Cohen's d)[3]에 주목하여 그 차이를 파악한다. 효과크

기는 평균치의 차이를 통합된 표준편차(pooled variance)로 나누어서 얻는 값(식(8) 참조)으로, 서로 다른 방법으로 측정된 유사한 관측값에 대한 통계 결과를 비교하는 데 용이하다.

$$d = \frac{m_1 - m_2}{\sigma} \tag{8}$$

효과크기는 통계적 유의성 이상의 정보를 제공하고, 다양한 방법으로 얻어진 결과들을 비교할 수 있도록 공통의 단위로 변화시킨다는 이점이 있으나, 표본크기에 따라 정확성이 달라질 수 있다는 한계를 갖는다(임시혁, 2016). 따라서 이 연구에서는 효과크기에 포함될 수 있는 오차를 정량화하고자 효과크기의 신뢰구간을 함께 제시하며(Lee, 2016), 이 때 신뢰구간의 범위는 95%으로 설정한다. 효과크기의 신뢰구간은 식(9), 식(10)으로 산출한다(Hedge & Olkin, 2014).

$$\sigma(d) = \sqrt{\frac{N_1 + N_2}{N_1 \times N_2} + \frac{d^2}{2(N_1 + N_2)}} \tag{9}$$

N_i: i집단의 사례수

$$95\% \text{ 신뢰구간: } [d - 1.96 \times \sigma(d), d + 1.96 \times \sigma(d)] \tag{10}$$

분석을 위한 도구로써, 기술통계치 및 상관분석, 경향점수 추정, 역확률 가중치 부여, 인과효과 추정을 위한 독립표본 t검정에 범용 통계 소프트웨어(SPSS 22)를 활용하였고, 경향점수 대응·층화·이중경향점수 보정 방법의 적용 및 동등성 진단에 R 3.0.3의 MatchIt(King, G., Ho, D., Stuart, E. A., & Imai, K, 2011), CaTools(Tuszynski, J., & ORPHANED, M., 2013) 등의 패키지를 활용하였다.

Ⅳ. 연구결과

1. 사교육 참여 여부에 따른 집단별 특성 비교

분석자료에 대한 기본적인 이해를 위해 사교육 여부에 따른 집단간 차이검증 결과를 제시한다. 총 3,244명의 사례가 분석에 활용되었으며, 그 중 사교육에 참여한 학생의 비율은 약 44.05%(1,

3) 표준화 차이계수와 동일한 수리식을 사용하나, 해석 면에서 차이를 가지므로 용어를 구분하여 사용함

429명)로, 사교육 미참여 학생보다 조금 적은 수준이다(<표 Ⅲ-1> 참조). 사교육 참여 여부에 따른 학생 특성을 확인하기 위해 대한 집단 간 차이 검증을 실시하였다(<표 Ⅳ-1> 참조).

<표 Ⅳ-1> 사교육 참여 여부에 따른 집단별 특성 비교

변수	사교육 (n=1,429)	비사교육 (n=1,815)	평균차이 (사교육-비사교육)	t
월평균 가구소득	5.94(2.43)	4.98(2.26)	0.96	11.57**
부모학력	3.06(0.95)	2.72(0.94)	0.34	10.06**
성별(남학생=1)	1.51(0.50)	1.54(0.50)	-0.03	-1.72
수학수업 이해정도	3.53(1.31)	2.44(1.38)	1.09	22.92**
수학수업태도	3.48(0.96)	2.78(1.08)	0.70	19.36**
수학교과흥미	3.24(0.94)	2.61(0.98)	0.63	18.51**
2차년도 수학성취도	44.43(23.81)	30.59(18.04)	13.84	18.24**
3차년도 수학성취도	51.05(25.24)	32.16(21.19)	18.89	22.69**

*p<.05, **p<.01

분석 결과, 두 집단 간에는 성별을 제외한 모든 변수들에 대해 통계적으로 유의미한 차이가 있는 것으로 나타났다. 이는 처치효과를 검증함에 있어 처치변수 외의 변수들이 성과변수에 유의미한 영향을 미칠 수 있음을 의미한다. 이러한 요소들은 인과효과 추정에서 편의를 발생시키며, 이는 연구 상황에 적절한 통계적 통제가 필요함을 시사한다.

2. 경향점수 추정

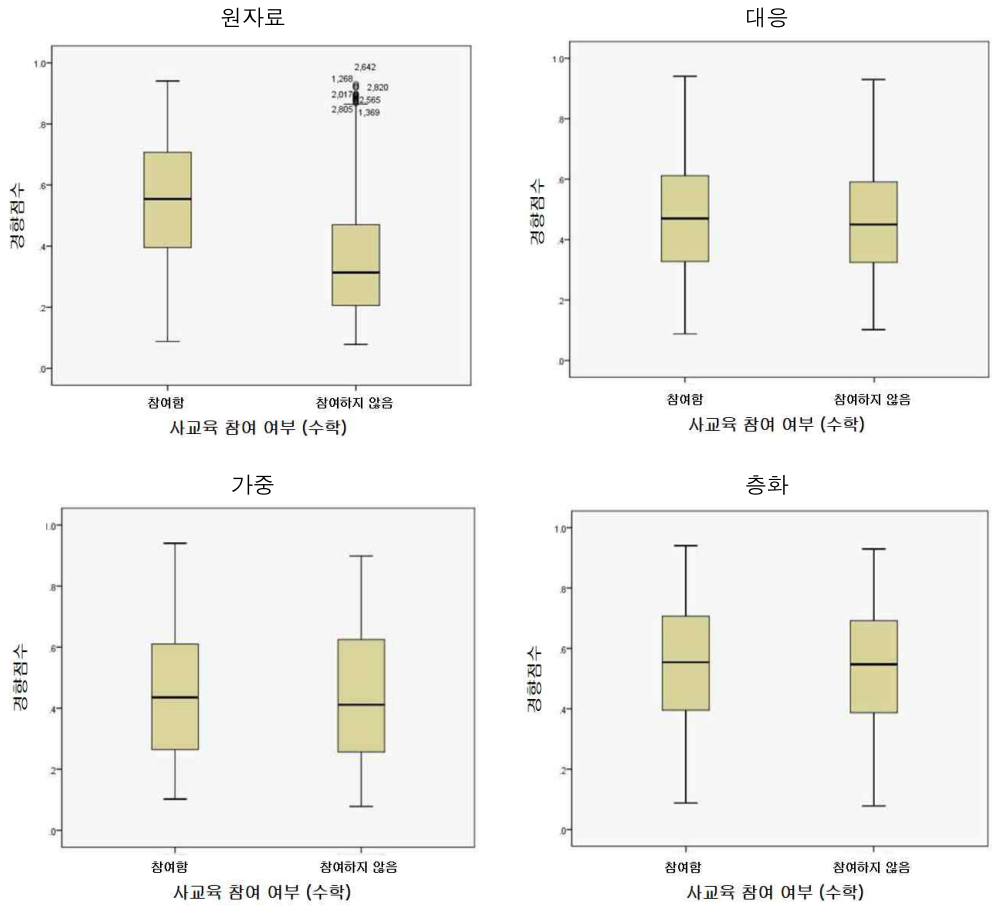
사교육 참여에 영향을 미칠 것으로 예상되는 다양한 변수를 포함하여 로지스틱 회귀분석으로 경향점수를 추정하였다(<표 Ⅳ-2> 참조).

<표 Ⅳ-2> 로지스틱 회귀모형 추정 결과

변수	회귀계수	표준오차	Wald	EXP(β)
상수	-3.64	0.20	343.62	0.03
월평균 가구소득	0.12	0.02	48.00	1.13
부모학력	0.15	0.04	11.49	1.16
성별(남학생=1)	-0.10	0.08	1.42	0.91
수학수업이해정도	0.29	0.04	47.93	1.34
수학수업태도	0.15	0.06	6.55	1.16
수학교과흥미	0.18	0.06	10.60	1.20
2차년도 수학성취도	0.01	0.00	32.76	1.01

로지스틱 회귀분석 결과, 성별 이외의 모든 변수들이 학생의 사교육 참여 여부를 설명하는데 유의미하게 기여하는 것으로 나타났다. <표 Ⅳ-2>에 제시된 바와 같이 월평균 가구소득이 많을수록, 부모학력과 전년도 수학성취도점수가 높을수록, 수학수업태도와 수학교과흥미가 긍정적일수록 학생이 사교육에 참여할 확률이 높았다. 그 중 가장 높은 승산비(EXP(β): 사교육에 참여할 확률을 참여하지 않을 확률로 나눈 값)를 갖는 변수는 수학수업이해정도(1.34)로 나타난다. 한편, 로지스틱 회귀모형을 통한 사교육 참여 여부의 추정정확성은 69.9%이며, 일반 회귀모형에서 설명력을 나타내는 중회귀계수와 유사한 Nagelkerke R^2(Nagelkerke, 1991)는 .24로 나타난다.

3. 경향점수 활용방법에 따른 결과 비교



[그림 Ⅳ-1] 원자료 대비 경향점수 활용 방법에 따른 경향점수 분포 비교

원자료의 상자 도표를 보면, 사교육에 참여한 학생들과 사교육에 참여하지 않은 학생들의 경향점수 분포가 상이하게 나타나며, 사교육 미참여 학생들이 참여 학생들보다 상대적으로 낮은 경향점수 값을 가짐을 알 수 있다([그림 Ⅳ-1][4] 참조). 이렇듯 사교육 참여 여부에 따라 구분한 집단이 서로 다른 분포를 갖는 것은 사교육에 참여한 학생과 사교육에 참여하지 않은 학생이 갖는 사교육 참여여부 외의 특성이 다름을 내포한다.

이후 제시된 상자 도표(대응, 가중, 층화)에서는, 통계적 통제 이후 사교육을 받은 학생들과 받지 않은 학생들의 경향점수 분포가 유사해졌음을 알 수 있다. 구체적으로, 범위, 사분위간 범위상자의 분포, 중앙값 등이 유사하게 나타난다. 이 때, 원자료에 나타났던 사교육 미참여 집단 내 극단치도 제외되었다. 이 때 극단치는 각 활용방법에 따라 대응에서는 대응되지 않은 사례들의 누락, 가중에서는 10 이상의 극단적인 가중치 제외 등의 방법으로 제외되었다. 한편, 표준화 차이계수 산출이 가능한 대응·층화 방법의 경우, 통계적 통제 이후 경향점수의 표준화 차이계수가 각각 0.072, 0.022로 0.1 이하의 값을 보이는데, 이 역시 경향점수 방법이 처치 및 비교집단의 동등성에 긍정적인 영향을 주어 두 집단 간 균형이 맞춰진 것으로 해석할 수 있다.

모든 방법에 대해 두 집단의 동등성이 확인되었으므로, 인과효과 추정을 위한 사교육/비사교육 집단 간 평균 비교(t검정)를 실시하였다(<표 Ⅳ-3> 참조).

<표 Ⅳ-3> 경향점수 활용방법별 사교육 및 비사교육 수학 성취도 평균 비교

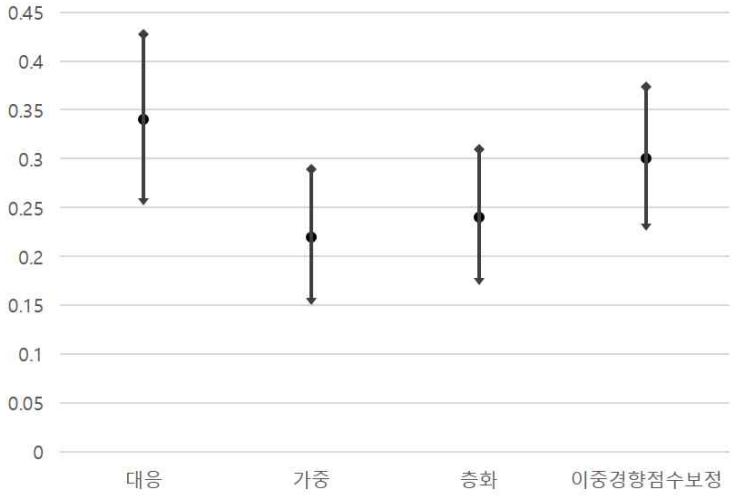
방법	사교육			비사교육			평균 차이	t	d	df
	n	평균	표준 편차	n	평균	표준 편차				
대응 (n=2,020)	1,010	46.15	23.83	1,010	38.07	24.18	8.09	7.57**	0.34	2,018
가중 (n=3,240)	1,428	44.19	23.81	1,812	38.77	25.63	5.42	8.82**	0.22	3,238
층화 (n=3,244)	1,429	51.05	25.24	1,815	44.57	26.83	6.48	7.06**	0.24	3,141
이중경향점수[5] (n=2,858)	1,429	51.05	25.24	1,429	45.13	11.84	5.92	8.03**	0.30	2,027
원자료 (n=3,244)	1,429	51.05	25.24	1,815	32.18	21.19	18.89	22.69**	0.81	2,778

*p<.05, **p<.01

4) 이중경향점수 보정방법은 동일한 처치집단의 관찰치와 추정치가 활용됨에 따라 [그림 Ⅳ-1]에 이중경향점수 보정방법의 경향점수 비교 그림이 제시되지 않음

5) 비교집단의 경향점수를 공변인으로 하여 학업성취도 점수를 예측한 회귀모형 $\widehat{Y^c} = 7.73 + 61.88ps(x)_t$을 산출하고 이를 처치집단에 적용하여 처치집단의 통제 조건 하에서의 성과변수(counterfactual) 추정치를 산출하였음

이상의 결과를 바탕으로 효과크기의 신뢰구간을 산출하여 도식화하였다.



[그림 Ⅳ-2] 효과크기의 신뢰구간 비교

정리된 자료를 통해 경향점수의 네 가지 활용방법은 사례수, 평균차이, 효과크기, 나아가 효과크기의 신뢰구간 측면에서 비교될 수 있다. 먼저, 분석에 활용된 사례수의 관점에서 각 방법의 특징을 살펴보았다. 대응 방법은 분석에 활용된 사례수가 2,020명으로, 가장 적은 사례수를 포함한다(전체 3,244명). 이는 방법론적 측면에서 경향점수 대응이라는 통계적 방법이 처치할당 확률의 균등화를 위한 매우 엄격한 절차임을 보여주는 것이다(민경석, 2008). 동일한 자료를 사용했음에도 적은 사례수를 포함하는 것은 보다 유사한(더 극단적으로는 동일한) 경향점수를 갖는 사례끼리 짝지어졌기 때문이므로 처치효과 추정치의 편의(bias)가 가장 작을 것으로 기대할 수 있다. 다만, 대응된 자료가 원자료의 특성과 어느 정도의 유사성을 갖는가에 대한 추가적인 연구 절차를 거쳤을 때 보다 의미있는 연구의 일반화가 가능하다.

특히 본 연구 분석자료와 같이 처치집단에 비해 대응할 비교집단이 크지 않고(44.1% vs 55.9%), 두 집단의 이질성이 명확히 드러나는 경우 상당수의 처치 집단이 성과분석에서 삭제될 가능성이 있으며, 본 연구에서도 약 30%(처치집단 1429명 중 419명)의 처치 사례가 최종 분석에서 배제되었음에 주목할 필요가 있다. 앞의 기술통계에서도 알 수 있듯이 원자료 처치집단의 수학 평균이 51.05였으나 대응 후의 처치 집단 평균이 46.15로 낮아져 대응 후 분석에 포함된 처치 집단의 모집단 특성을 어떻게 규명할지에 대한 연구자의 주의가 필요하며, 나아가 대응에 기반한 인과효과 추정치가 ATE, ATT의 추정값인지에 대한 논의는 집단 크기, 자료의 구조, 대응 이후의 두 집단의 분포 등이 고려되어야 한다.

한편, 가중, 층화, 이중경향점수 보정 방법은 처치집단의 전체 사례를 분석에 활용하며, 그 중 층화 방법은 비교집단의 사례도 모두 활용한다는 특징을 갖는다. 다만 본 연구에서는 경향점수 산출 이후 대응 방식과의 비교를 위해 가중 방식의 경우 식(3)의 가중치 공식을 활용하여 ATE를 추정하였으며, 반면 층화 방식은 이중경향점수 보정 방식과의 비교를 위해 층별로 산출된 인과효과 추정값에 층별 처치집단 사례수에 근거하여 가중치를 부여함으로써 전체 처치집단에 대한 ATT 인과효과를 추정하였다. 위의 기술통계값에서도 이를 확인할 수 있는데, <표 Ⅳ-3>에서 대응과 가중에서의 처치 및 비교집단의 평균값이 유사하였고, 층화와 이중경향점수 보정 방식에서 비교집단의 평균이 유사하였으며 처치집단의 경우 원자료 처치집단과 동일하여 처치집단의 사례수를 유지하였음을 확인할 수 있다.

다음으로 성과변수에 대한 독립표본 t검정 결과, 모든 방법에서 처치 및 비교집단의 평균차이가 통계적으로 유의한 것으로 확인되었다. 즉, 사교육에 참여한 학생과 참여하지 않은 학생의 수학 학업성취도 평균이 통계적으로 유의미한 차이를 보임이 확인되었다. 평균의 차이는 '대응(8.09) > 층화(6.48) > 이중경향점수 보정(5.92) > 가중(5.42)'의 순서대로 크게 나타났다. 주목할 것은 원자료와 경향점수 방법 적용 자료들의 평균차이로, 경향점수 방법 적용 시 평균차이가 원자료의 평균차이(18.89)에 비해 큰 폭으로 감소됨이 확인되는데 이는 원자료 평균에는 편의(bias)가 포함된 것임을 다시 한 번 확인하는 기회가 되었다.

유사한 맥락에서 효과크기를 살펴보면, 모든 활용방법이 낮은 효과크기에 속하는 0.22에서 0.34사이 값을 보임을 알 수 있다. 그 중 대응이 상대적으로 큰 값(0.34)으로 나타났고, 가중방법이 가장 낮은 값(0.22)으로 나타났다. 층화와 이중경향점수 보정은 각각 0.24, 0.30으로 확인된다. 네 가지 활용방법 모두 원자료의 효과크기(0.81, 큰 효과크기에 해당)에 비해 큰 폭 줄어들었다는 점에 주목해보면, 이는 경향점수 방법이 강한 조정이 이루어지는 통계적 통제 방법임을 보여주는 결과라고 해석할 수 있다.

효과크기의 신뢰구간은 가중[0.14, 0.30], 층화[0.17, 0.31], 이중경향점수 보정[0.23, 0.37]이 비교적 유사한 구간 폭을 보이는 반면, 대응[0.25, 0.43] 방법이 상대적으로 큰 구간 폭을 보임이 확인된다([그림 Ⅳ-2] 참조). 이는 가중치 방법이 대응 방법에 비해 좁은 신뢰구간을 제공했다는 김준엽, 정혜경, Seltzer(2008)의 연구와도 일치하는 결과이다. 이러한 차이는 각 방법의 분석에서 활용된 사례수에 의한 차이로 해석할 수 있다. 즉, 강한 통계적 통제로 인해 제외되는 사례수가 많이 발생하는 대응에서 보다 넓은 신뢰구간이 추정되는 것이다. 신뢰구간이 넓다는 것은 통계적 추정의 정밀성이 다소 낮아진다는 의미를 갖는다.

결과적으로, 네 가지 경향점수 활용방법은 인과효과 추정 시 통계적 유의성 면에서는 동일한 결과를 보이나, 평균차이, 효과크기, 효과크기의 신뢰구간에 있어 다소 다른 패턴을 보였다. 즉, 동일한 자료로 경향점수를 추정하여 분석한다 하더라도 활용방법에 따라 평균차이, 효과크기 등의 통계치가 달라질 수 있음을 알 수 있다. 물론 통계적 유의도 면에서 동일한 결과

가 확인되지만, 보다 엄격한 관점에서 본다면, 경향점수 활용방법에 따라 인과효과의 정도가 달리 추정된다고 볼 수 있다. 본 연구 분석을 통해 ATT를 측정하는 층화와 이중경향점수 보정 방법은 근접한 결과를 보였으며, 가중 방식을 적용한 ATE 추정치가 ATT에 비해 다소 작게 추정된 것으로 확인되었으나 그 신뢰구간이 겹쳐 유의한 차이가 있다고 볼 수는 없다. 대응 방식의 경우 또한 다른 적용 방식과 인과효과의 신뢰구간이 겹쳐 통계적으로 동일한 결론에 도달하나 나머지 세 방식과 다르게 처치집단의 상당수를 분석에서 제외함으로써 추론 결과에 대한 명확한 해석 측면에서 제한점이 있음을 확인하였다.

한편, 이 연구에서 소개한 이중경향점수 보정 방법은 처치집단의 관찰치(Y_i^t)와 추정치($\widehat{Y_i^c}$)가 활용된다는 측면에서, 교육적으로 특수한 집단을 연구대상으로 설정하여 처치집단의 사례수가 현저하게 작을 때, 두 집단이 극단적으로 다른 분포를 보일 때 유용할 것으로 기대된다. 즉, 층화 방법이나 가중 방법을 활용했을 때 무리한 통계적 통제가 이루어질 것으로 판단되는 분포, 사례수라고 판단될 경우 고려해볼만 하다(Austin, 2017).

Ⅴ. 결론 및 제언

경향점수는 비실험 연구에서 처치 및 비교집단 간 동등성을 확보하기 위한 통계적 통제방법 중 하나이다. 경향점수는 처치변수와 성과변수에 영향을 미치는 공변인들에 의해 추정되며, 이를 활용하여 실험연구와 유사한 상황을 통계적으로 구현한다. 이 연구는 경향점수 활용방법에 주목하여 경향점수 대응, 가중, 층화, 이중경향점수 보정에 따른 인과효과 추정 결과를 비교·분석하였으며, 다음과 같은 결론을 도출하였다.

대응 방법의 경우, 경험적 자료를 활용하는 연구 상황에서 내적타당도를 높여 편의(bias)를 줄이는 이점이 있음에도 불구하고 이질성이 높은 집단일수록 대응의 엄밀성이 떨어지거나 다수의 사례가 탈락하는 등의 방법적 한계를 갖는다. 다수의 사례가 탈락한다는 것은 연구자가 처치하고자 하는 집단의 특성을 얼마나 대변하는가에 대한 문제로 연결되며, 이는 일반화가능성과 직결된다. 이 경우 처치 효과 추정의 결과 해석에 있어서 ATE와 ATT에 해당하는지에 대한 결론이 모호할 수 있으며, 따라서 보다 정교한 연구를 위해서는 대응된 자료가 원자료의 특성과 유사한지에 대한 추가 분석을 실시하는 것이 필요하다. 결과적으로 대응 방식은 전체 처치집단의 분포를 전반적으로 포괄하면서도 대응할 수 있는 충분한 비교집단의 사례수를 가진 원자료를 확보한 경우, 캘리퍼나 최적화 방식 등을 접목하여 편의표집을 줄이는 효과적 방식이라 할 수 있겠다.

반면, 가중, 층화, 이중경향점수 보정 방법의 장점은 처치집단의 모든 사례수를 분석에 활용한다는 것이다. 일반적으로 추정하는 모수에 따라 평균 처치효과(ATE)를 볼 수도, 처치집단에 대한 평균 처치효과(ATT)를 볼 수도 있는데 이 때 처치집단에 대한 평균 처치효과(ATT)는 처치집단에 초점을 맞추는 것으로, 처치를 받을만한 집단에 대한 인과효과를 집중적으로 보고자 할 때 유용하다(Min, Jung & Kim, 2017). 본 연구의 실증분석 결과 가중 방법의 경우 평균 처치 효과(ATE)를, 층화와 이중경향점수 보정 방법의 경우 처치집단에 대한 평균 처치효과(ATT)를 추정하였기 때문에 처치집단에 대한 평균 처치효과를 추정하는 층화와 이중경향점수 보정방법은 상대적으로 유사한 결과를 보였으며, ATE를 추정한 가중 방식은 미세하나마 나머지 두 방식보다 작은 효과크기를 보였다.

결론적으로, 경향점수를 활용한 인과효과 분석의 경우, 자료의 구조, 활용 자료의 사례수, 연구자의 관심 등에 따른 각 활용방법의 특성을 고려하여 보다 적절한 방법을 선택하여 적용하는 것이 바람직하다. 조사연구에서 처치집단과 비교집단의 사전 동등성의 차이가 크거나 비교집단의 사례수가 작은 경우 처치집단과 비교집단의 범위를 제한하여 동등성을 설정하는 대응 방법보다는 가중이나 층화 방식이 권장될 수 있으며, 특히 ATT 추정에 있어 원자료의 처치집단을 그대로 유지하면서도 비교집단의 사례수에도 균형을 이루는 이중경향점수 보정방법이 보다 적절한 활용방법으로 선택될 수 있을 것이다.

이 연구에 대한 후속 연구를 위해 제언할 사항은 다음과 같다. 첫째, 이 연구는 경기교육종단연구의 자료를 활용한 연구로, 사례들의 개별 특성이 반영된 경험적 자료에 의한 분석이 실시되었다. 이에 본 연구의 결과는 활용 자료에서의 제한적인 결과라는 한계를 갖는다. 따라서, 방법론적 접근에 있어 보다 신뢰로운 결과를 도출하기 위하여 모의실험을 활용한 추가 연구를 진행해 볼 필요가 있다. 모의실험 설계 시 처치집단과 비교집단의 사례수를 다양하게 구성해보는 것도 고려해볼 수 있다. 둘째, 이 연구에서는 인과효과 검정을 위한 방법으로 독립표본 t검정만을 사용하였다. 따라서 인과효과 검정방법을 확장하여 회귀모형, 위계적 선형모형 등의 방법을 적용해 볼 필요가 있다. 마지막으로, 이중경향점수 보정 방법에 대한 추가연구가 필요하다. 특히, 이 방법이 의학통계 연구 분야에서 처음 제안된 방법인 만큼, 교육연구를 비롯한 사회과학 연구에서의 적용 가능성을 제고하는 탐색이 진행되어야 할 것이다.

참 고 문 헌

경기도교육연구원(2017). 경기도교육연구원 홈페이지 http://www.gie.re.kr (검색일: 2017. 07.10.)

김준엽, 정혜경, Seltzer, M. H. (2008). Drawing causal inferences using propensity score methods in educational research. **교육평가연구, 21**(3), 19-242.

김지현(2016). 인과연구에서 중첩편향을 제거하기 위한 공변량선택기준. **응용통계연구, 29**(5), 849-858.

민경석(2008). 자아 존중감에 대한 경향점수를 이용한 평준화 효과 분석. **교육평가연구, 21**(3), 1-21.

성기선, 김준엽, 박소영, 민병철(2013). **경기교육종단연구 2차년도 문항개발 연구.** 경기도교육연구원. GEPS-6-2013(4).

이진실(2016). 교육연구에서 경향점수를 활용한 순차적 처치 효과 분석. 서울대학교 대학원 박사학위논문.

임소정, 정인경(2017). 이중 성향점수 보정 방법을 이용한 처리효과 추정치의 표준오차 추정: 붓스트랩의 적용. **응용통계연구, 30**(3), 453 - 462.

임시혁(2016). 검증력, 효과의 크기, 신뢰구간 사용 실태 분석. **교육논총, 53**(1), 1-13.

장은진, 안정훈, 정선영, 황진섭, 이자연, 심정임, 이선희(2013). 측정된 교란요인을 고려한 성과분석 방법. **NECA 연구방법 시리즈,** 1-272.

정혜경(2012). 비실험 연구에서 인과효과 추정을 위한 방법론적 고찰. **교육학연구, 50**(3), 20-50.

최강식(2007). 고용영향 분석평가 방법론 연구. **직업능력개발연구,** 10(3), 181-202.

Abadie, A., & Imbens, G. W. (2011). Bias-corrected matching estimators for average treatment effects. *J Bus Econ Stat, 29,* 1 - 11.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research, 46*(3), 399-424.

Austin, P. C. (2017). Double propensity-score adjustment: a solution to design bias or bias due to incomplete matching. *Statical methods in medical research, 26*(1), 201-222.

Becker, S. O., & Ichino, A. (2002). Estimation of average treatment effects based on

propensity scores. *The stata journal, 2*(4), 358-377.

Guo, S., & Fraser, M. W. (2010). *Propensity score analysis: Statistical methods and applications.* Thousand Oaks, CA: Sage.

Hedge, L., & Olkin, I. (2014). *Statistical Methods for Meta-Analysis (1st ed.).* Academic press.

Hirano, K., & Imbens, G, W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology, 2*(3-4), 259-278.

Holland, P. W. (1986). Statistics and causal inference. *Journal of American Statistical Association, 81*(396), 945-960.

King, G., Ho, D., Stuart, E. A., & Imai, K. (2011). Package 'MatchIt'. Retrieved from https://cran.r-project.org/web/packages/MatchIt/MatchIt.pdf

Lee, D. K. (2016). Alternatives to P value: confidence interval and effect size. *Korean journal of anesthesiology, 69*(6), 555-562.

Min, K. S., Jung, H., & Kim, C. M. (2017). Examining a causal effect of Gyeonggi innovation schools in Korea. *KEDI Journal of Educational Policy, 14*(2), 3-20.

Morgan, S. L., & Todd, J. J. (2008). A diagnostic routine for the detection of consequential heterogeneity of causal effects. *Sociological Methodology, 38*(1), 231-281.

Morgan, S., & Winship, C. (2007). Counterfactuals and causal inference: Methods and principles for social research. New York, NY: Cambridge University Press.

Nagelkerke, N. J. (1991). A note on a general definition of the coefficient of determination. *Biometrika, 78*(3), 691-692.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41-45.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association, 79*(387), 516-524.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician, 39*(1), 33-38.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66,* 688-701.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology, 2*(3), 169-188.

Rubin, D. B. & Thomas, N. (2000) Combining propensity score matching with additional adjustments for prognostic covariates. *J Am Stat Assoc, 95*, 573 - 585.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton-Mifflin.

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods, 13*(4), 279 - 313. doi:10.1037/a0014268

Tuszynski, J., & ORPHANED, M. (2013). Package 'caTools'. Retrieved from https://cran.r-project.org/web/packages/caTools/caTools.pdf

· 논문접수 : 2019.01.04. / 수정본접수 : 2019.05.03. / 게재승인 : 2019.05.17.

ABSTRACT

Comparison of propensity score methods for causal inference : matching, weighting, subclassification, and double propensity score adjustment

Hyun-Jin Jang
Researcher, Korea Institute for Curriculum and Evaluation
Hyekyung Jung
Associate Research Fellow, Korea Institute for Curriculum and Evaluation
Kyung-Seok Min
Professor, Sejong University

This study aimed to introduce the double propensity score adjustment method and compare it with matching, weighting, and subclassification methods using estimated propensity scores. Based on empirical data analysis, the study emphasized the recognition of the estmand (ATT vs ATE) a researcher wants to investigate, considering the data structure, the sample sizes of the treatment and comparison groups, and the distributions of covariates between the groups to draw a valid and meaningful inference with respect to the effect of a treatment on the population of interest. To do so, we compared the results of four different propensity score utilization methods by analyzing the effects of private education on mathematic achievement with Gyeonggi educational longitudinal research data(GEPS2012). The results showed that matching was relatively limited to use when the sample size of the comparison group was not large enough or the distributions of propensity score between the treament and comparison groups were not sufficiently overlapped. On the other hand, the double propensity score adjustment method could be useful to estimate the average treatment effect for the treated when research targeted a special population and the treatment and comparison groups were different in various aspects.

Key Words: causal inference, propensity score, double propensity score adjustment, ATT