

CEM(Coarsened Exact Matching)

방법의 이해와 활용

- 연구개발 보조금 효과를 중심으로 -

김 상 신*

논문초록

본 연구는 프로그램 평가에서 많이 활용되고 있는 매칭방법 중 CEM 분석방법을 소개하고 이를 활용하여 정부의 연구개발 보조금 지원이 기업의 연구개발 지출에 미치는 영향을 분석하였다. 또한 다른 매칭방법들과의 비교를 통해 CEM 분석방법이 가지는 장점과 단점을 제시함으로써 향후 프로그램 평가에서 CEM 방법이 다양하게 활용될 수 있는 기반을 제공하고자 하였다. CEM은 매칭방법 중 하나로 매칭절차에서 나타나는 불균형을 분석자가 사전에 통제할 수 있으며, 반복적 매칭절차 없이 공통실증영역으로 매칭된 자료를 제한함으로써 모형의존성을 줄일 수 있는 장점이 있다. 또한 PSM 등 타 매칭방법과의 비교 결과, 불균형을 가장 크게 축소하는 매칭결과를 산출해 내는 것으로 나타났다. 물론 CEM 분석방법의 경우 처치집단의 관측치를 일부 활용하지 못하는 단점이 존재하지만, 비교대상 관측치가 많을 경우 불균형을 축소시킬 수 있다는 측면에서 CEM은 향후 많은 활용이 가능한 방법론으로 생각된다.

핵심 주제어: 정책평가, CEM(Coarsened Exact Matching), PSM(Propensity Score Matching)

경제학문헌목록 주제분류: H2, C4

투고 일자: 2016. 4. 20. 심사 및 수정 일자: 2016. 7. 26. 게재 확정 일자: 2016. 8. 16.

* 한국개발연구원 전문연구원, e-mail: sskim@kdi.re.kr

I. 서론

정책을 평가하는 것은 경제학뿐만 아니라 다른 학문분야에서도 주요한 관심이다. 이러한 평가는 다른 대안적 정책으로부터 생성된 결과 값을 비교하는 것으로서 일반적으로는 수행되지 않는 정책들의 결과 값의 추정을 요구하게 된다(Heckman, 2007). 이는 반사실적 비교집단을 구성하는 것으로, Holland (1986)가 지적한 바와 같이 “인과추론의 근본적 문제”로 이상적 실험환경이 아닐 경우 비교대상을 구성할 수 없는 문제가 발생하게 된다.

정책평가의 근본적 문제를 해결하기 위해서 “처치효과” 접근법이 적용되어 왔다(Heckman, 2007). 처치효과는 처치를 받은 개체와 받지 않은 개체의 인과효과로서, 주로 정책수혜 여부, 처치 여부 등에 대한 평가를 위해 사회·경제 및 의학 분야에서 폭넓게 활용되고 있다. 처치효과의 추정은 계량경제학적으로 구조방정식 모형(Structural Equation Modeling), 도구변수(Instrumental Variable) 모형, 회귀단절(Regression Discontinuity) 모형, 매칭(Matching) 방법 등이 다양하게 활용되고 있다.

본 연구에서는 처치효과 추정방법으로 최근에 활발하게 활용되고 있는 매칭방법 중 Iacus, King, and Porro (2009)에 의해 개발된 CEM(Coarsened Exact Matching: 이하 CEM) 방법에 대해 자세히 소개하고 이를 활용하여 정부의 연구개발 보조금의 효과를 추정함으로써 CEM 방법의 활용 및 타 매칭방법론과의 차이를 비교 분석한다.

본 연구의 구성은 다음과 같다. 제Ⅱ장에서는 기본적인 처치효과의 개념과 매칭 방법론 및 CEM 방법에 대해서 살펴보고, 제Ⅲ장에서는 CEM 방법을 통해 정부의 연구개발 보조금의 효과를 추정해 본다. 제Ⅳ장에서는 정부의 연구개발 보조금의 효과에 대해서 다른 매칭방법들을 활용해서 추정한 후 이를 CEM 방법에서 추정된 결과와 비교할 것이다. 제Ⅴ장에서는 결론을 제시한다.

II. CEM 방법 설명

1. 처치효과(treatment effect)

처치효과 추정에서 결과 값의 단순 비교나 OLS 분석은 누락변수 편이 또는 선택 편의로 인하여 편이된 추정치를 제공한다. 정부보조금 지급 효과를 고려할 경우, 정부보조금 지급의 처치효과는 보조금을 받은 처치그룹의 성과와 동일한 그룹이 처치를 받지 않았을 때의 성과의 차이로서 추정할 수 있을 것이다.

Rubin (1974) 에 의해 처음으로 수식화된 인과효과는 다음과 같이 나타낼 수 있다.

$$\text{인과효과: } Y_i(1) - Y_i(0)$$

여기서 $Y_i(1) = Y_i(T_i = 1)$ 이고 T 는 처치변수이다. $Y_i(1)$ 과 $Y_i(0)$ 는 유닛 i 가 처치 또는 통제를 받은 것에 대한 잠재적 결과 값으로, “인과추론의 근본적 문제”(Holland, 1986)는 $Y_i(1)$ 과 $Y_i(0)$ 중 오직 하나만 관측된다는 것으로서 이는 손실자료(missing data) 문제로 생각될 수 있다(Rubin, 1976a).

비교대상의 부재로 인하여 개별적 인과효과를 추정할 수 없기 때문에 연구자들은 처치그룹 내의 처치효과인 처치에 대한 평균처치효과(average treatment effect on the treated: 이하 ATT)와 모든 개체에 대한 효과인 평균처치효과(average treatment effect: 이하 ATE)에 주목하였다.

$$\text{ATE: } E[Y_i(1) - Y_i(0)] = E[Y_i(1)] - E[Y_i(0)]$$

$$\text{ATT: } E[Y_i(1) - Y_i(0)|T_i = 1] = E[Y_i(1)|T_i = 1] - E[Y_i(0)|T_i = 1]$$

일반적으로 경제학 등 사회과학에서는 ATT에 보다 초점을 맞추고 있다. 예를 들어 본 분석에서 추정하고자 하는 인과효과는 보조금을 지급받은 기업들의 성과와 동일한 기업들이 보조금을 받지 않았을 때의 성과와의 차이를 추정하고자 하는 것으로서 이는 ATE보다 ATT를 추정하는 것이 더 합리적이다. 또한 Heckman (1997)은 ATE의 경우 해당 프로그램 참여를 의도하지 않은 이들에 대한 효과를 포

합하기 때문에, 실질적으로 프로그램에 참여를 의도한 이들의 효과를 추정하기 위해서 ATE 보다 ATT가 정책결정자들에게 더 유용할 것이라고 하였다.

ATT에서 $E[Y_i(0)|T_i = 1]$ 는 결측자료이다. 따라서 $E[Y_i(0)|T_i = 0]$ 을 $E[Y_i(0)|T_i = 1]$ 대신에 활용하는 방안이 고려되어 왔다. Heckman and Smith (1995)에 따르면, 이러한 비처치자들의 잠재적 결과 값을 활용하는 것은 처치된 그룹의 평균처치효과에 선택편의를 더하는 것과 같은 추정치를 산출하게 된다. 이러한 선택편의는 처치그룹과 비처치그룹의 관측가능하거나 관측가능하지 않은 이질성으로부터 나타난다.

$$\begin{aligned} & E[Y_i(1)|T_i = 1] - E[Y_i(0)|T_i = 1] \\ &= E[Y_i(1)|T_i = 1] - E[Y_i(0)|T_i = 1] + E[Y_i(0)|T_i = 1] - E[Y_i(0)|T_i = 0] \\ &= \tau + E[Y_i(0)|T_i = 1] - E[Y_i(0)|T_i = 0] \end{aligned}$$

여기서 $\tau = ATT$ 이다.

즉, 문제는 처치의 결과 값과 통제의 결과 값이 비처치 상황에서 동일하지 않다는 것이다.

모집단으로부터 두 임의추출 표본 C 와 T 를 고려하면, 이들 표본들은 전체 모집단에서 통계적으로 동일하기 때문에 식 (1)과 (2)가 성립하고, 이를 ATE에 적용하면 “인과추론의 근본적인 문제”를 해결할 수 있다. 즉, 임의화 할당은 ATE를 추정함에 있어, 통제집단 C 를 처치집단 T 의 대응적사실 상황의 값으로 활용하여 문제를 해결할 수 있는 것이다.¹⁾

$$E[Y_i(0)|i \in C] = E[Y_i(0)|i \in T] = E[Y_i(0)] \quad (1)$$

$$E[Y_i(1)|i \in C] = E[Y_i(1)|i \in T] = E[Y_i(1)] \quad (2)$$

$$\begin{aligned} ATE &\equiv E[Y_i(1)] - E[Y_i(0)] \\ &= E[Y_i(1)|i \in T] - E[Y_i(0)|i \in C] \end{aligned}$$

다음으로 ATT에서 처치할당이 임의화라고 가정하면 즉, $Y(1), Y(0) \perp T$ 라면,

1) ATT와 ATE에 대한 수식은 Ichino (2007)의 내용을 참고해서 작성하였다.

다음 식 (3), (4) 와 같고 이를 ATT에 넣으면, 다음과 같다.

$$E[Y_i(0)|T_i = 0] = E[Y_i(0)|T_i = 1] = E[Y_i(0)] \quad (3)$$

$$E[Y_i(1)|T_i = 0] = E[Y_i(1)|T_i = 1] = E[Y_i(1)] \quad (4)$$

$$\begin{aligned} ATT &\equiv E[Y_i(1)|T_i = 1] - E[Y_i(0)|T_i = 1] \\ &= E[Y_i(1)|T_i = 1] - E[Y_i(0)|T_i = 0] \\ &= E[Y_i|T_i = 1] - E[Y_i|T_i = 0] \end{aligned}$$

$$Sample\ Selection\ Bias = E[Y_i(0)|T_i = 1] - E[Y_i(0)|T_i = 0] = 0$$

이 경우, 표본선택편의가 0이 되기 때문에 임의할당은 ATT를 추정가능하게 해준다. 만일 연구개발 보조금 지급이 임의적으로 할당되었다면, 보조금 지급의 성과는 보조금을 수혜한 처치그룹과 보조금을 수혜하지 않은 통제그룹 간의 성과의 차이로서 측정될 수 있다. 그러나 이상적인 실험상황이 아닌 비실험상황 또는 관측자료 연구에서는 처치와 통제를 받은 유닛의 결정이 특정 할당 메커니즘을 통해 이루어질 수밖에 없기 때문에 추가적인 비혼란성 가정이 필요하게 된다.

비혼란성(unconfoundedness) 가정: 처치할당이 주어진 사전처치 변수들에 의해 혼란되지 않는다. $[Y(1), Y(0) \perp D|X]$

여기서 X 는 사전처치 특성들의 집합으로 통상 공변량으로 언급되어진다.

비혼란성이 성립한다면 ATT는 식 (5), (6) 과 같으며,

$$E[Y_i(0)|T=0, X] = E[Y_i(0)|T_i = 1, X] = E[Y_i(0)|X] \quad (5)$$

$$E[Y_i(1)|T_i = 0, X] = E[Y_i(1)|T_i = 1, X] = E[Y_i(1)|X] \quad (6)$$

위 두식과 반복기대의 법칙(law of iterated expectation)을 활용하여 ATT를 다시 정리하면 다음과 같다.

$$\begin{aligned}
ATT &\equiv E[Y_i(1) - Y_i(0) | T_i = 1] \\
&= E[E[Y_i(1) - Y_i(0) | T_i = 1, X] | T_i = 1] \\
&= E[E[Y_i(1) | T_i = 1, X] - E[Y_i(0) | T_i = 1, X] | T_i = 1] \\
&= E[E[Y_i(1) | T_i = 1, X] - E[Y_i(0) | T_i = 0, X] | T_i = 1] \\
&= E[E[Y_i | T_i = 1, X] - E[Y_i | T_i = 0, X] | T_i = 1]
\end{aligned}$$

이는 비혼란성의 가정이 성립할 경우 관측자료에서 ATT를 추정할 수 있다는 것을 보여준다. 매칭방법을 통해 ATT를 추정함에 있어 중요한 요소는 “균형 (balancing)”을 달성하는 것이다. 매칭방법에서의 균형 달성은 매칭된 자료가 임의화된 실험에서와 유사하게 임의적 할당을 통해, 처치와 통제그룹들의 관측된 공변량에서 오직 임의적 차이만이 발생하는 것을 보장함으로써 ATT의 추정을 가능하게 해 준다.

2. 매칭방법

매칭방법은 관측자료에서 사전처치 공변량들의 확률적 동질성이 성립하지 않는 문제, 즉 추정된 처치효과가 처치그룹과 통제그룹 사이의 혼란요인들에 의해 편의되기 때문에 발생하는 혼란편의를 제거하기 위한 방법이다. 단순 매칭은 통계적 처치효과와 관련된 Cochran and Cox (1957), Rubin (1972) 등의 연구에 의해 개발되었다 (Qin, 2011). 매칭은 관측자료에서 사전처치 통제변수의 혼란영향을 통제하는 비모수적 방법으로서, 매칭 이후 처치그룹과 통제그룹의 그룹 내 공변량(X)의 실증적 분포가 더 유사하게 되는 것을 의미한다 (Iacus, King, and Porro, 2012).

일반적으로 매칭은 추정방법이 아닌, 덜 모형 의존적(model-dependent) 매칭 자료의 생성을 통해 처치효과를 추정하도록 하는 사전처리 방법으로 볼 수 있다²⁾ (Iacus, King, and Porro, 2009). 매칭에 의한 사전처치는 매칭을 활용하지 않을 때보다, 모형의존성과 통계적 편의를 줄이는 것으로 알려져 있다 (Ho, Imai, King, and Stuart, 2007).

2) 물론 정확한 매칭 절차를 통해 생성된 매칭 자료는 잠재적 결과 값들의 단순 평균차이 비교를 통해 처치효과를 추정할 수 있지만, 일반적으로 매칭방법은 추정보다는 사전처리 절차에 가깝다고 볼 수 있다.

매칭방법은 20세기 중반부터 활용되어져 왔으나, 이론적 기초는 공변량에 기반한 ATT를 추정하는 것에 초점을 둔 Cochran and Rubin(1973)과 Rubin(1973)에 의해 제시되었다(Stuart, 2010).

가장 기본적으로, 직관적인 매칭방법은 일대일 정확한 매칭일 것이다. 이는 관측된 처치유닛 i (결과 값 Y_i 와 공변량 X_i 를 가짐)에 상응하는 $\tilde{X}_i = X_i$ 에 의해 선택된 통제유닛의 결과 값을 통해 관측되지 않은 $Y_i(0)$ 를 추정한다. 만일 일대일 정확한 매칭을 통해 처치집단과 통제집단이 완전하게 매칭이 되었다면, 처치와 통제집단 내 결과 값 Y 의 단순 평균 차이는 인과효과의 완전한 비모수적 추정치를 제공할 것이다(Iacus, King, and Porro, 2012). 즉, 정확한 매칭을 통한 균형의 달성은 평균의 차이만으로 처치효과의 분석이 충분하다는 것을 의미한다.

그러나, 일대일 정확한 매칭방법은 현실적으로 실현되기 어려운 문제가 발생한다. 예를 들어, 10개의 이진 공변량을 가지는 매칭을 고려할 경우, 이들의 가능한 조합은 2^{10} , 즉 1,024가지나 되기 때문에 처치유닛과 통제유닛이 정확하게 매칭되는 숫자는 굉장히 큰 자료에 기반하지 않는 한 소수에 그칠 것이다. 실제로, Stuart(2010)은 Chapin(1947)의 연구를 예로 들어 671개 처치유닛과 523개 통제유닛 중 23개의 짝만이 6개 범주형 공변량에서 정확하게 매칭된 것을 통해 정확한 매칭의 비현실성을 지적하고 있다. 이러한 문제는 ‘차원의 저주’로 언급된다. 따라서 분석자들은 정확한 매칭 대신에 근사적 매칭방법을 활용할 수밖에 없다. 근사적 매칭은 먼저 공통실증영역 바깥의 처치와 통제유닛들을 제거하고, 모든 공변량 X 를 직접적으로 활용하는 대신에 처치와 통제유닛 사이의 근사성 측정 방법을 활용하여 매칭을 수행한다.

일반적으로 EPBR(Equal Percent Bias Reduction: 이하 EPBR) 기반³⁾의 MDM(Mahalanobis Distance Matching: 이하 MDM)과 PSM(Propensity Score Matching: 이하 PSM)이 가장 폭넓게 활용되어 왔다. 그러나 EPBR 기반의 매칭방법들은 사후

3) EPBR은 공변량들의 모든 선형 결합에서 편의를 줄이는 특성이라고 할 수 있다(Rubin, 1976b; Rubin and Thomas, 1992). 세부적으로 매칭 변수들 각각의 편위의 축소비율이 같을 때 매칭방법을 EPBR이라고 하며, 공변량들의 선형함수에서 편위의 같은 비율축소를 얻는 매칭방법이 EPBR이라는 것과 필요충분조건이다. 게다가, 만일 매칭방법이 EPBR이 아니라면, 모든 단일공변량(univariate covariate) 평균들이 매칭하지 않은 것 보다 매칭된 자료에서 더 가깝다고 할지라도 X 의 선형함수에 대한 편의가 증가한다(Rubin, 1976b).

적인 균형의 검토가 필수적이며, 매칭 이후 균형의 검토와 균형된 결과를 얻기 위해 재매칭을 수행하는 반복적 절차가 수반되어야 한다. 이때 하나의 공변량에 대한 균형의 향상이 다른 공변량의 균형에 영향을 미친다는 문제가 있다. 또한 PSM의 경우 비모수적 추정방법이면서 성향점수를 추정하는 과정에서 로짓 또는 프로빗 모형의 모수적 가정이 필수적이어야 한다는 다소 모순적인 문제가 발생한다.

이러한 EBPR 계열의 매칭방법이 가지는 문제를 피하기 위해서 Iacus, King, and Porro (2009)은 MIB (Monotonic Imbalance Bounding: 이하 MIB) 계열의 매칭방법인 CEM을 소개하였다. CEM은 일반적 무시가능성에 대한 가정을 넘어서는 어떠한 가정도 자료생성절차에서 요구하지 않으며, 처치와 통제그룹들 사이의 불균형의 정도를 분석자가 사전적으로 선택한 것보다 커질 수 없게 하며, 하나의 공변량에 대한 균형의 향상이 다른 공변량의 최대 불균형에 영향을 미치지 않도록 설계되어 있다(Iacus, King, and Porro, 2012).

그러나 어떠한 방법도 사전에 불균형의 수준을 경계하는 것과 사전에 매칭된 관측치의 수를 정할 수는 없다. EBPR 방법들은 분석자에게 사전적으로 매칭된 샘플 사이즈를 선택할 수 있게 하고, 사후적으로 균형을 계산한다. 반면에 MIB는 분석자가 사전적으로 최대 불균형을 선택하고, 사후적으로 매칭된 샘플 사이즈를 생성한다. 주로 자료가 풍부하고 비효율보다는 추정결과의 편의를 줄이는 것에 초점을 맞춘다면, MIB가 선호될 수 있다(Iacus, King, and Porro, 2012).

3. CEM(coarsened exact matching)

CEM은 MIB 기반의 매칭방법으로, 최대 불균형을 사전적으로 선택함으로써 통상 매칭 이후 처치그룹과 통제그룹 사이의 균형을 검토하고 재매칭하는 어려운 반복절차가 필요 없고,⁴⁾ 한 공변량의 최대 불균형에 대한 조정이 다른 공변량들의 최대 불균형에 영향을 미치지 않기 때문에 공변량을 통한 불균형의 조정이 보다 용이한 방법이다(Iacus, King, and Porro, 2012).

CEM이 개발된 핵심적 동기는 정확한 매칭이 완벽한 균형을 제공하나 '차원의 저

4) 다만, 거칠어진 구간(bin)을 설정함에 따라 매칭된 관측치의 수와 불균형이 변한다. 따라서 보다 나은 매칭을 위해서 분석자에 의해 거칠어진 구간을 변경하면서 반복적 재분석이 수행될 수 있다. 이에 대해서는 뒤에 보다 자세히 언급한다.

주' 때문에 소수의 매칭된 자료만을 생성하는 것을 보완하기 위해서이다. CEM의 아이디어는 분석에 포함되는 공변량들을 일시적으로 거칠게(coarsen) 한 후 거칠어진 공변량에 대해서 정확한 매칭이 이루어지고, 매칭 이후 원자료를 통해 처치효과를 분석하는 것이다.

CEM 알고리즘을 요약하면 다음과 같다(Iacus, King, and Porro, 2009).

1. 처치할당과 잠재적 결과 값에 영향을 미칠 것으로 판단되는 모든 공변량(X)들을 일시적으로 거칠게 한다. 예를 들어, 고용원 수를 1~50명, 51~100명, 101~200명 등으로 범주화하는 것이다.
2. 거칠어진 공변량(X)의 조합에 상응하는 계층에 해당 거칠어진 공변량 값을 가지는 유닛들을 배치하고 정렬한다.

예를 들어, X_1 과 X_2 두 개의 공변량을 가지는 모형을 고려해 보자. X_1 과 X_2 가 각각 세 개의 분류로 거칠어진다고 가정하면 분석대상 자료는 <Table 1>과 같이 9개의 잠재적 계층(strata)으로 분리될 수 있다.

<Table 1> Sorting the Units into Strata

X_1/X_2	1	2	3
1	s_{11}	s_{12}	s_{13}
2	s_{21}	s_{22}	s_{23}
3	s_{31}	s_{32}	s_{33}

이러한 각각의 계층들은 다음의 4가지 상황을 가지게 된다.

- (1) 포함된 유닛들이 없는 계층
- (2) 처치유닛과 통제유닛 모두를 포함하는 계층
- (3) 처치유닛만을 포함하는 계층
- (4) 통제유닛만을 포함하는 계층

3. 최소한 하나 이상의 처치와 통제유닛을 포함하지 않는 계층의 유닛들은 자료로부터 제외한다. 즉, 앞서 (2)에 해당하는 자료만을 남긴다. 물론 처치유닛만을 포함한 계층을 외삽법을 통해 통제집단을 매칭해서 활용할 수 있으나, 이 경우에는 모형의존성이 높아지는 결과가 초래된다.

여기서, 거칠어진 정도가 세분화되면, 모형의존성이 제거된 더 정확한 매칭 결과를 제공할 것이나 적은 관측치만이 활용가능하게 될 것이고, 거칠어진 정도가 넓은 너비를 가지면, 적은 관측치가 버려질 것이나, 계층 내 차이가 발생할 것이고 이는 통계적 모형을 통해 추가적으로 통제되어야 할 것이다(Iacus, King, and Porro, 2009). 거칠어진 정도와 관련해서 분석자가 선택할 수도 있지만, R, STATA와 같은 소프트웨어에서는 히스토그램에 기반하여 구간을 자동적으로 선택하도록 프로그래밍되어 있다.

CEM을 통해 거칠어진 자료의 정확한 매칭 절차가 완료된 이후 처치효과 추정을 위해 다시 거칠어지기 전의 자료를 활용한다. 이 때 앞서 언급한 바와 같이 계층 내 최소 하나 이상의 처치와 통제유닛을 가진 자료만을 처치효과 추정을 위해서 활용한다.

처치효과 추정은 매칭된 처치와 통제유닛이 가지는 CEM 가중치를 포함한 회귀분석 등 모수적 추정방법을 활용해서 얻을 수 있다. 만일 거칠어진 것 없이 정확한 매칭이 이루어졌다면, 종속변수에 결과 값, 설명변수에는 처치여부의 더미변수를 놓고 단순 가중 회귀분석을 수행하는 것이 적합할 것이다. 그러나 거칠어진 수준으로 인하여 원자료상에 남아있는 불균형이 존재하며, 이러한 경우 합리적인 접근은 통계적 모형을 활용하여 남아 있는 불균형을 조정하는 것이다. 물론 상당수 매칭방법을 활용하는 연구들에서 매칭 작업 이후에 결과 값의 단순 평균 차이를 처치효과로 해석하나, 관측자료 연구에서 정확한 매칭이 달성되는 것은 거의 불가능하기 때문에, CEM 이후 통계적 모형(OLS)을 활용하여 남아있는 불균형을 조정하는 것이 필요하다. 이는 단순 회귀분석을 활용할 경우, CEM을 통해 얻어진 가중치를 활용해 가중 회귀분석을 수행 하는 것으로 이때, 설명변수들은 CEM 매칭을 수행할 때 활용된 모든 공변량들이 포함되어야 할 것이다.

CEM을 통한 가중치는 다음과 같이 배정되며, 매칭되지 않은 유닛들은 가중치 $w_i = 0$ 을 받는다.

$$w_i = \begin{cases} 1, & i \in T^s \\ \frac{m_C}{m_T} \frac{m_T^s}{m_C^s}, & i \in C^s \end{cases}$$

s : 계층

T^s : 계층 내 처치유닛

C^s : 계층 내 통제유닛

m_T : 매칭된 전체 처치유닛 수

m_C : 매칭된 전체 통제유닛 수

m_T^s : 계층 내 매칭된 처치유닛 수

m_C^s : 계층 내 매칭된 통제유닛 수

4. CEM에서의 불균형의 측정

매칭방법에서는 처치집단과 매칭된 통제집단 또는 매칭된 처치집단과 매칭된 통제집단, $p(X|P=1)$, $p(\tilde{X}|T=0)$ 에 대한 사전처치 공변량의 다변량 실증 분포 사이의 차이 즉, 불균형(imbalance)의 측정이 필수적이다. 그러나 대부분 매칭방법을 활용한 연구들에서는 이러한 불균형을 측정할 때 처치와 통제그룹 내 평균의 단변량 즉, 개별 공변량 각각의 절대적 차이만을 비교하고 있다.

특히 상당수 PSM을 활용하는 연구들에서는 균형의 평가를 매칭된 처치그룹과 통제그룹의 공변량들의 평균 차이를 t-test 등을 통해 진단한다. PSM 매칭에서는 각 계층별로 공변량들의 차이 검증을 통해 균형 달성여부를 검토해야 하지만, 많은 연구들에서는 매칭된 이후에 처치그룹과 통제그룹의 공변량들의 전체 평균 차이의 결과만을 제시하고 있는 것이다. 평균차이에 대한 t-test의 문제점은 모집단이 아닌 관측된 표본이기 때문에 가설검정이 부정확할 수 있다는 문제점뿐만 아니라, 어느 정도의 평균의 차이가 좋은 것인지에 대해서 정보를 제공하지 않기 때문에(Ho, Imai, King, and Stuart, 2007) 분석자들의 지속적인 매칭과 균형검토를 요구하며, 심지어 최종적으로 제시하는 균형이 가장 효율적인 균형이 아닐 수도 있다.

CEM 분석방법에서의 불균형은 비모수추정 방법인 다변량 밀도함수추정(multivariate density estimation) 방법의 L_1 통계치에 의해서 측정된다. L_1 은 처치

그룹과 통제그룹의 모든 사전처치 공변량들의 다차원 히스토그램 사이의 차이를 통해서 계산된다.⁵⁾

먼저, 거칠어진 공변량들의 구간(bin) 크기를 계산하고, 이후, 처치와 통제그룹 각각 공변량들의 카티션곱(cartesian product)에 의해 생성된 셀들의 집합 $H(X_1) \times \dots \times H(X_k) = H(X) = H$ 으로서 각 셀들의 교차표를 작성한다. 그리고 처치 $f_{l_1 \dots l_k}$ 와 통제 $g_{l_1 \dots l_k}$ 에 대한 k 차원 상대빈도(k -dimensional relative frequency)를 각 셀에 기록한다. 마지막으로 모든 셀 값들에 걸쳐 차이의 절대 값으로 불균형을 측정한다.

$$L_1(f, g; H) = \frac{1}{2} \sum_{l_1 \dots l_k \in H(X)} |f_{l_1 \dots l_k} - g_{l_1 \dots l_k}|$$

L_1 : 다변량 불균형 측정

$H(X) = \prod_{i=1}^k H(X_i)$: 거칠어진 공변량 X_i 내 구간(bin)들의 집합인 $H(X_i)$ 의 카티션곱에 의해 생성된 셀들의 집합으로부터 구성된 다차원 히스토그램이며, 이는 매칭에 대한 사전적 최대불균형의 수준임.

f : 처치유닛들의 실증빈도분포

g : 통제유닛들의 실증빈도분포

$f_{l_1 \dots l_k}$: 좌표 $l_1 \dots l_k$ 를 가지는 셀에 속한 처치유닛들에 대한 상대빈도

$g_{l_1 \dots l_k}$: 좌표 $l_1 \dots l_k$ 를 가지는 셀에 속한 통제유닛들에 대한 상대빈도

처치그룹과 통제그룹이 완전한 균형이면, 다변량 히스토그램의 차이는 $L_1 = 0$ 일 것이고, 이보다 큰 값은 불균형을 나타낸다. $L_1 = 1$ 이면, 두 그룹간의 완전한 구분을 나타낸다.

CEM 방법에서 L_1 값의 값은 불균형을 측정하는 지표이다. 먼저 매칭되지 않은 처치그룹과 비교그룹의 L_1 의 값을 L_{10} 라고 한다면 이 값은 매칭된 자료의 불균형 값과의 비교 기준이 되며, 이러한 의미에서 앞서 최대불균형을 사전적으로 결정한

5) CEM 분석에서의 불균형 측정과 관련해서는 Blackwell, Iacus, King, and Porro (2010) 연구를 참고하여 정리하였다.

다고 표현한 것이다. 즉, 매칭 이후의 매칭된 처치그룹과 비교그룹의 L_1 의 값⁶⁾이 L_{10} 보다 작은 경우에 불균형이 향상되었다는 것을 의미하며, 가능한 매칭된 그룹에서의 L_1 의 값이 작아지는 것을 희망한다. 거칠어지는 구간은 프로그램(히스토그램 방법)에 의해 자동으로 선택되어질 수 있지만, 또한 분석자가 이 구간을 설정할 수 있다. 만일 분석자가 구간을 설정한다면(프로그램과 다르게), 매칭된 개체 수와 L_1 값이 변하게 된다. 따라서 CEM 방법 역시 매칭된 수와 불균형 수준을 고려해서 최적의 거칠어진 구간을 선택해야 하는 과정을 반복적으로 수행할 필요가 있으나, 앞서 PSM 방법과는 달리 최대불균형이라는 비교 기준이 존재하기 때문에 반복적 수행의 과정이 보다 수월할 것이다.

5. 매칭방법별 비교

Iacus, King, and Porro(2011, 2012), King, Nielsen, Coberley, Pope, and Wells(2011)은 EPBR 계열의 매칭방법과 CEM을 포함한 시뮬레이션에서 CEM이 불균형, 모형의존성, 추정오차, 편의, 분산, 평균제곱오차와 같은 기준에서 다른 매칭방법들보다 우월하다는 것을 보였다.

Iacus, King, and Porro(2012)은 Lalonde(1986)의 미국 직업훈련 프로그램 자료를 활용, CEM, PSM, MDM, GM(Genetic Matching) 방법별 처치효과를 몬테카를로 시뮬레이션을 활용하여 추정한 후, 이들 결과의 편의, 표준편차, 평균제곱오차, 불균형(L_1)을 측정하였다. 5,000번의 몬테카를로 시뮬레이션을 통해 추정된 ATT 추정치⁷⁾의 편의, 표준편차, 평균제곱오차와 불균형에서 CEM이 다른 모든 매칭방법들보다 더 나은 결과를 제시하는 것을 보였다.

Iacus, King, and Porro(2011)은 Gu and Rosenbaum(1993) 실험을 복제하여, 처치집단 1,000개 통제집단 3,000개의 랜덤샘플을 활용하여 MDM, PSM, CEM의 분석결과를 비교하였다. 비교결과 CEM이 다른 매칭방법에 비해 공변량들의 평균차이와 불균형(L_1) 측면에서 더 나은 결과를 나타내는 것을 보였다.

6) 매칭 후 L_1 값은 매칭된 처치그룹과 비교그룹에 속한 관측치들의 매칭 가중치를 활용하여 L_1 값을 계산하게 된다.

7) 앞서 매칭방법에서 활용되는 처치효과의 추정치는 실질적으로 표본 ATT인 SATT이다. 그러나 논의과정에서의 혼돈을 방지하기 위해서 ATT로 일관되게 언급한다.

CEM이 다른 매칭방법들에 비해서 많은 장점을 가지나 CEM 역시 단점이 존재한다. 가장 큰 문제는 CEM 역시 여전히 ‘차원의 저주’가 나타날 수 있다는 점이다. 그러나 일반적으로 CEM 방법이 고려된 근본적인 이유가 연속공변량에서 나타나는 일대일 매칭 문제를 해결하기 위한 것임을 가정한다면, 연속 공변량을 포함하는 경우에 좀 더 효율적인 매칭방법으로 고려될 수 있을 것이다.

Iacus, King, and Porro(2011, 2012), King, Nielsen, Coberley, Pope, and Wells(2011)의 결과에서는 CEM 방법이 다른 매칭방법에 비해 우월하다는 것을 보이고 있으나, 매칭방법을 선택하는 것은 자료와 공변량의 형태를 감안하고, 다수의 매칭방법을 통해 가장 적합한 매칭방법을 선택하는 것이 필요하다.

CEM을 활용한 국내 연구의 경우 김기현·신인철(2012)에서 보육시설 미경험자와 경험자를 매칭하는 과정에서 CEM을 활용하였으나, 현재까지는 제한적으로 활용되고 있다.

III. CEM 분석의 활용

본 장에서는 CEM 방법을 활용하여 정부의 연구개발 보조금이 기업의 연구개발 지출에 미치는 효과를 평가함으로써, 실제 CEM 방법의 활용에 대해서 설명한다.

1. 연구개발 보조금의 효과 분석 개요

정부가 기업의 연구개발에 대해 보조금을 지급하는 것은 연구개발이 공공재의 특성을 가지기 때문이며, 인위적 개입이 없을 경우 사회적으로 바람직한 수준 이하에서 연구개발 활동이 이루어지는데, 이는 이로 인해 사회적 후생이 감소하기 때문이다. 즉, 정부의 직접보조금을 통해 기업의 연구개발 한계비용을 낮춰 연구개발을 촉진시키기 위함이다.

조가원(2009)에서는 정부의 연구개발 지원이 미치는 효과를 네 가지로 구분하였다. 첫째, 기업이 혁신활동을 수행하도록 하는 참여효과, 둘째, 기업의 추가적인 연구개발 지출을 유도하는 투자효과, 셋째, 기업 혁신활동의 기술적 성과에 미치는 혁신효과, 넷째, 기업 혁신활동의 경제적 성과에 미치는 효과인 경제적 효과로 각각 구분하였다. 정부의 연구개발 지원의 효과는 흔히 부가성(additionality)의 개념

으로 설명되고 있다. 즉, 정부 지원 규모를 넘어선 부가적 효과가 발생하였는지가 주요 관심인 것이다.

이 중 정부지원의 투자효과와 관련해서 정부의 직접보조금이 구축효과를 발생시켰는지를 CEM 분석을 통해 확인해 볼 것이다. 구축효과는 정부의 직접보조금을 받은 기업이 보조금을 받지 않았을 때 지출하고자 하는 연구개발 지출보다 그 규모를 줄이는 경우에 발생한다.

김호·김병근(2012)은 정부보조금의 민간연구개발투자에 대한 효과를 분석하면서, 관련된 선행연구들을 정리하였다. 김호·김병근(2012)에 기초해서 선행연구들을 살펴보면, 권남훈·고상훈(2004), 신태영(2004), 유민화(2006), 최석준·김상신(2007, 2009), 김기완(2008), 송종국·김혁준(2009), 이우성 외(2009), 주홍신 외(2011)⁸⁾ 등이 있으며, 이 중 권남훈·고상원(2004)의 경우 완전한 대체효과를 보였으나, 다른 연구들에서는 결과가 상이하게 나타나 명확한 결론을 도출하기는 어렵다고 밝히고 있다. 국내 연구의 경우 DID(Difference-in-Difference: 이하 DID) 모형과 PSM, 2단계 프로빗 모형, 패널고정효과 모형 등 대체로 표본선택 편의를 고려한 분석모형이 활용되었다.

해외연구와 관련해서는 연구개발 보조금의 효과성을 추정하기 위해 메타회귀분석을 수행한 Dimos and Pugh(2016) 연구를 통해 선행연구들을 확인할 수 있다. Dimos and Pugh(2016)의 연구는 2000년 이후 총 52개 미시수준 연구자료를 활용하여 메타분석을 수행하였다. 분석에 활용된 연구들 중 정부 보조금 수혜여부와 기업의 총 연구개발 지출의 자료를 활용한 분석결과의 90% 정도에서 계수추정치가 부가성, 효과 없음, 부분적 구축효과를 가지는 것으로 나타났으며, 10% 정도가 완전구축효과를 가지는 것으로 나타났다. Dimos and Pugh(2016)가 정리한 선행연구 중 투자효과 관련 49개 연구에서 효과 추정시 PSM 방법 25회,⁹⁾ 도구변수(IV) 방법이 13회, 패널분석이 7회, DID가 6회, OLS가 2회 활용되어 매칭방법이 가장 활발히 활용되고 있음을 알 수 있다.

8) 해당 연구들의 출처는 김호·김병근(2012)의 연구를 참고하기 바란다.

9) 연구별로 복수의 분석방법이 활용되는 경우가 포함되어 있다.

2. 분석자료 및 변수설정

분석자료는 과학기술정책연구원에서 기업들의 혁신활동을 조사한 ‘한국기업혁신조사’ 2014년 자료를 활용한다. 본 조사는 상시종사자 수 10인 이상의 제조기업을 대상으로 조사된 자료이며, 2011~2013년 기업의 현황 및 혁신활동 결과에 대해 조사하고 있다.

분석에 활용되는 결과변수는 기업의 연구개발 지출액으로 조사상 내부 연구개발 비용과 외부 연구개발 비용을 합한 값의 로그 값을 활용한다.¹⁰⁾

처치변수는 정부지원제도 중 기술개발 및 사업화지원(자금지원)을 받은 기업여부의 이산변수를 활용하고, 설명변수로는 기업 특성을 통제하기 위해 대기업, 중기업, 소기업, 벤처기업 여부와 매출액,¹¹⁾ 기업의 수출 여부, 전체 고용원 중 연구개발 전담인원 비율을 고려하였고, 산업 특성의 통제를 위해 OECD 기술수준 구분¹²⁾에 따른 고기술·중고기술·중저기술·저기술 산업 여부와 시장구조를 나타내는 지표로 상위 3개 기업의 누적시장점유율을 나타내는 누적시장집중도(CR3)를 활용하였다.¹³⁾ 설명변수 중 매출액과 수출기업 여부는 연구개발 지출과의 내생성을 줄이기 위해서 2011년 자료를 활용하였다. 분석대상 자료는 연구개발을 수행한 기업들 즉, 연구개발 지출이 0 이상인 기업들만을 대상으로 분석한다.

각 변수별 설명 및 기초통계량은 〈Table 2〉, 〈Table 3〉과 같다.

-
- 10) 물론, 정확한 연구개발 보조금의 투자효과를 분석하기 위해서는 기업의 연구개발 지출액에서 정부로부터 지원받은 보조금을 뺀 순 연구개발 지출액만을 결과변수로 활용하여야 한다. 그러나 본 분석에서 활용한 자료에서는 기업이 받은 보조금의 크기를 제공하고 있지 않아 부득이 기업의 전체 연구개발 지출액을 활용한다. 이에 따라 본 분석의 결과로는 완전 구축효과의 발생유무만을 확인할 수 있지, 부분구축효과의 발생여부는 확인할 수 없다.
 - 11) 설명변수 중 매출액과 수출여부의 경우 연속변수를 활용하는 것이 보다 적절하나, 본 조사자료에서 이들 변수들이 범주형 응답 값으로 조사되었기 때문에 부득이 매출액은 범주형 응답 값을 직접 활용하고 수출액의 경우, 수출여부의 이산변수로 활용하였다.
 - 12) OECD 기술수준 산업구분은 OECD (2011)에서 Hatzichronoglou (1997)이 연구개발 집중도에 따라 구분한 산업구분을 기반으로 보안을 통해 고기술산업, 중고기술산업, 중저기술산업, 저기술산업으로 산업을 재분류한 기준이다. 표준산업분류에 따른 기술수준별 산업분류 결과는 [Appendix] 〈Table A〉를 참고하기 바란다.
 - 13) 누적시장집중도는 이재형·양점삼·이상무(2013) ‘시장구조조사’에서 2digit 표준산업분류에 따른 값을 활용하였다.

3. 분석결과

분석은 CEM을 통해 모형의존성이 줄어든 매칭자료를 활용, 매칭 후 남아있는 불균형을 통제하기 위해서 매칭된 가중치를 이용하여 모수적 추정(OLS)을 수행한다. 분석대상 자료의 CEM 이전과 CEM 이후의 각 통제변수와 전체 L_1 의 변화(불균형) 및 최종 매칭 정보는 <Table 4>와 같다. CEM 이전과 이후의 L_1 의 변화를 보면 매칭 이후에 개별 통제변수들의 L_1 뿐만 아니라, 전체 L_1 의 값도 상당히 줄어든 것을 알 수 있다. 이는 매칭을 통해 통제변수들의 균형이 향상되었음을 나타낸다. CEM 이후 최종 매칭된 정보를 보면, 매칭되지 않은 관측치들은 가중치가 0이 되어 모수적 분석(OLS)시 활용되지 않는다.

<Table 4> Comparison of Imbalance before and after CEM

Variable	Imbalance (L_1)		Information of matched units		
	Before CEM	After CEM			
large	0.0298	≈ 0	Number of units	Treated	Untreated
middle	0.0087	≈ 0			
small	0.0385	≈ 0	Total units	760	399
venture	0.1016	≈ 0			
sale	0.0687	≈ 0	Matched units	321	262
export	0.1131	≈ 0			
rnd_employ_ratio	0.1355	0.0907	Non-matched units	439	137
high_tech	0.0302	≈ 0			
middle_high_tech	0.0949	≈ 0	Total strata	588	
middle_low_tech	0.0659	≈ 0			
low_tech	0.0593	≈ 0	Matched strata	149	
cr3	0.0910	0.0200			
Total	0.7118	0.3302			

Note: \approx is approximately equal signs.

CEM 이후 부여된 가중치를 기반으로 OLS 분석을 수행한 분석결과는 <Table 5>와 같다. 비교를 위해 매칭되지 않은 원자료의 OLS 분석결과와 매칭 이후 가중치를 활용한 분석결과를 같이 제시한다.

〈Table 5〉 OLS Estimates on R&D Expenditures before and after CEM

Variables	OLS before CEM		OLS after CEM	
	Coefficient	Std. Err.	Coefficient	Std. Err.
rnd_subsidy	0.189	0.133	0.412**	0.186
large	1.646***	0.322	0.373	0.596
middle	0.712***	0.195	0.427	0.385
venture	0.383***	0.145	0.284	0.238
sale	0.565***	0.073	0.676***	0.150
export	0.404***	0.145	0.298	0.268
rnd_employ_ratio	0.042***	0.005	0.028***	0.007
high_tech	1.210***	0.217	2.132***	0.440
middle_high_tech	0.588***	0.187	0.878**	0.388
middle_low_tech	0.124	0.207	-0.015	0.429
cr3	-0.024***	0.007	-0.059***	0.012
constant	1.789***	0.369	2.836***	0.668
Obs.	1,159		583	
Rrob>F	0.0000		0.0000	
Adj_R_squared	0.3523		0.3030	

Note: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

분석결과를 보면 매칭을 하지 않았을 경우 정부의 보조금을 수혜한 기업과 그렇지 않은 기업은 기업의 연구개발 지출에서 통계적으로 유의하게 차이가 나지 않았으나, CEM을 통해 매칭을 한 분석결과에서는 보조금을 받은 기업은 그렇지 않은 기업에 비해 다른 조건이 동일할 때 41%¹⁵⁾ 연구개발 지출을 늘리는 것으로 나타났다. 다만, 종속변수에서 연구개발 보조금의 양을 제외하지 않았기 때문에 이러한 결과를 통해 단지 정부의 연구개발 보조금의 지급이 완전구조효과를 나타내지는 않는다는 것을 보여준다는 점에서 한계가 있다. 다음으로 CEM 분석결과와 다른 매칭방법들의 분석결과를 비교한다.

15) 종속변수가 로그 값이고 설명변수가 더미변수인 경우 엄밀히 해석하자면, 두 가지 측면에서 결과 값을 해석할 수 있다. 즉, 보조금을 받지 않은 기업이 보조금을 받을 경우 $100(\exp(0.412) - 1) = 50.98\%$ 기업의 연구개발 지출이 증가하고, 보조금을 받은 기업이 보조금을 받지 않을 경우에는 $100(\exp(-0.412) - 1) = -33.77\%$ 기업의 연구개발 지출이 감소하는 것으로 해석할 수 있다.

VI. 매칭방법별 비교

앞서 언급한 바와 같이 CEM이 다른 매칭방법에 비해 장점이 존재하지만, 타 매칭방법 또는 모수적 추정방법에 비해서 절대적으로 우월하다는 것은 아니다. 따라서 다양한 방법론을 활용하여 분석하고 이들 결과를 비교함으로써 보다 강건한 분석결과를 얻을 수 있을 것이다.

본 장에서는 CEM 이외에 다른 매칭방법을 활용하여 연구개발 보조금의 투자효과를 분석하였고, 분석결과와 분석모형에 따른 불균형 정도를 비교하였다. 분석방법은 MDM과 PSM 중 one-to-one Nearest Neighbor 매칭(PSM_NN), Radius 매칭(PSM_R), Kernel 매칭방법(PSM_K)을 활용하였다. MDM은 마할라노비스 거리에 기반하여 처치집단과 통제집단을 매칭하는 방법이고, PSM_NN, PSM_R, PSM_K는 연구개발 보조금 수혜여부를 종속변수로 한 성향점수를 기반으로 매칭을 수행하는 방법들로, PSM_NN은 가장 성향점수가 유사한 관측치 간에 매칭을 수행하며, PSM_R은 성향점수 차이가 일정 수준 이내인 자료들을 가중치를 이용하여 활용하며, PSM_K는 성향점수 차이에 역비례하는 가중치를 통제변수에 부여하여 매칭을 수행하는 방법이다. PSM 분석 시 매칭과정에서 통제집단에 속한 기업들의 중복매칭을 허용하였다.

〈Table 6〉 Matching Results of Control and Treatment Groups by Matching Methods

Number of units	Control Group		Treatment Group	
	Total units	Matched units	Total units	Matched units
CEM	760	321	399	262
MM	760	259	399	399
PSM_NN	760	277	399	399
PSM_R	760	748	399	398
PSM_K	760	760	399	399

Note: PSM_R(caliper=0.005).

먼저 각 매칭방법별로 매칭된 관측치 정보를 보면 〈Table 6〉과 같다. PSM이나 MDM의 경우에는 사전에 모든 처치집단의 매칭을 전제하고 있기 때문에 매칭된 관측치의 손실이 발생하지 않으나, CEM의 경우 매칭되지 않은 처치집단의 관측치

가 제외되어 매칭된 관측치의 수가 적어지는 단점이 존재한다. 또한 PSM_R, PSM_K의 경우에는 가중치를 이용하여 가급적 많은 통제집단의 자료를 활용하기 때문에 매칭에 활용된 통제집단의 관측치 수가 많이 포함되어 있다.

다음으로는 각 매칭방법별로 매칭된 자료에 대한 불균형을 측정하였다. 불균형은 CEM 방법에서 활용하는 L_1 값과 PSM에서 주로 활용되고 있는 매칭된 설명변수들 간의 처치집단과 통제집단 차이의 통계적 유의성 검증을 활용하였다.

먼저 각 매칭방법별 L_1 값을 보면 매칭 전 값이 0.7118이었고 CEM 매칭 후에는 0.3302, MDM에서는 0.5514, PSM계열 방법들은 0.7040-0.7093의 값을 보이고 있다. L_1 값으로 측정된 불균형측면에서는 CEM 매칭방법의 불균형이 제일 작은 것으로 나타나고 있다. 다음으로 매칭된 각 설명변수별 차이를 보면, 모든 매칭방법에서 매칭된 처치집단과 통제집단의 설명변수의 평균차이가 통계적으로 유의하지 않는 것으로 나타났다. 즉, 매칭된 집단 간의 각 설명변수들의 차이가 0이 아니라는 귀무가설을 기각하는 결과로, 각 설명변수들의 집단 간 차이가 존재하지 않는 것으로 볼 수 있다.

(Table 7) Comparison of Imbalance(L_1) before and after Matching

Variables	Before Matching	After Matching				
		CEM	MDM	PSM_NN	PSM_R	PSM_K
large	0.0298	≈0	0	0.0025	0.0043	0.0028
middle	0.0087	≈0	0.0025	0.0100	0.0033	0.0082
small	0.0385	≈0	0.0025	0.0125	0.0010	0.0110
venture	0.1016	≈0	0.0075	0.0075	0.0068	0.0007
sale	0.0687	≈0	0.0150	0.0451	0.0554	0.0565
export	0.1131	≈0	0.0025	0.0426	0.0021	0.0076
rnd_employ_ratio	0.1355	0.0907	0.0602	0.0902	0.0765	0.0750
high_tech	0.0302	≈0	0	0.0251	0.0061	0.0047
middle_high_tech	0.0949	≈0	0	0.0251	0.0069	0.0022
middle_low_tech	0.0659	≈0	0	0.0025	0.0013	0.0043
low_tech	0.0593	≈0	0	0.0025	0.0022	0.0018
cr3	0.0910	0.0200	0.0426	0.0627	0.0619	0.0533
Total	0.7118	0.3302	0.5514	0.7093	0.7090	0.7040

Note: ≈ is approximately equal signs.

〈Table 8〉 Difference in Means before and after Matching, for Each Variable Listed

Variables	Before Matching	After Matching				
		CEM	MDM	PSM_NN	PSM_R	PSM_K
large	0.030	≈0	0	0.003	0.004	0.003
middle	0.009	≈0	0.003	0.010	-0.003	0.008
small	-0.039	≈0	-0.003	-0.013	-0.001	-0.011
venture	0.102***	≈0	0.008	0.008	-0.007	-0.001
sale	0.128	≈0	-0.030	0.030	-0.005	0.029
export	0.113***	≈0	0.003	-0.043	0.002	0.008
rnd_employ_ratio	3.577***	-1.193	1.464	1.273	0.180	0.668
high_tech	0.302	≈0	≈0	0.025	0.006	0.005
middle_high_tech	0.095***	≈0	≈0	-0.025	-0.007	-0.002
middle_low_tech	-0.066**	≈0	≈0	-0.003	-0.001	-0.004
low_tech	-0.059**	≈0	≈0	0.003	0.002	0.002
cr3	1.126*	0.109	0.134	0.681	0.207	0.220

Note: ≈ is approximately equal signs.

〈Table 9〉 Average Treatment Effect(ATT) of Subsidies on R&D Expenditure by Matching Methods

		Outcome of treatment group	Outcome of control group	Difference (ATT)	Std. Err.	t-value
Before Matching		5.291	4.681	0.610***	0.161	3.78
After Matching	CEM	5.381	5.009	0.372*	0.222	1.68
	MDM	5.291	5.056	0.235	0.223	1.06
	PSM_NN	5.291	4.875	0.416*	0.222	1.88
	PSM_R	5.280	5.030	0.250	0.169	1.48
	PSM_K	5.291	5.045	0.246	0.165	1.49

Note: *** p<0.01, ** p<0.05, * p<0.1.

〈Table 9〉는 각 매칭방법을 통한 처치에 대한 ATT 결과이다. 먼저 매칭 이후 단 순히 처치집단과 통제집단의 ATT를 계산한 결과를 보면, 매칭 이후 평균처치효과 가 매칭 전보다 줄었음을 볼 수 있다. 또한 CEM와 PSM_NN에서만 통계적 유의성 (10%) 이 나타나는 것을 볼 수 있다. 〈Table 10〉은 매칭된 가중치를 활용하여 앞서 CEM 분석과 동일하게 OLS 분석을 수행한 결과이다. 분석결과를 보면 CEM 분석

결과보다 계수추정치 크기는 줄어들었으나, PSM_NN과 PSM_R의 분석결과에서도 통계적으로 유의하게 연구개발 보조금이 기업 연구개발 지출에 양의 효과를 미치는 것을 보여주고 있다.

(Table 10) The Coefficients of R&D Grant Dummy in Weighted OLS Estimates on R&D Expenditures after Matching

	Before Matching	After Matching				
		CEM	MDM	PSM_NN	PSM_R	PSM_K
rnd_subsidy	0.189 (0.133)	0.412** (0.186)	0.188 (0.162)	0.368** (0.159)	0.245** (0.123)	0.192 (0.122)

Note: () standard error, *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

V. 결론

본 연구는 프로그램 평가에서 최근 많이 활용되고 있는 매칭방법 중 CEM 분석방법을 소개하고 이를 활용하여 정부의 연구개발 보조금지원이 기업의 연구개발 지출에 미치는 영향을 분석하였다. 또한 다른 매칭방법들과의 비교를 통해 CEM 분석방법이 가지는 장점과 단점을 제시함으로써 향후 프로그램 평가에서 CEM 방법이 다양하게 활용될 수 있는 기반을 제공하고자 하였다.

CEM을 통한 연구개발 보조금지원의 효과를 분석한 결과 정부의 연구개발 보조금지원이 기업의 연구개발 지출을 통계적으로 유의하게 높이는 것으로 나타났다. 비록 강건성 검증을 위해 다양한 매칭방법을 적용한 결과, 일부 매칭방법에서는 연구개발 보조금의 통계적 유의성이 나타나지는 않았지만, 가장 많이 활용되고 있는 PSM 방법(PSM_NN)을 활용한 분석에서는 CEM 결과와 같이 정부의 연구개발 보조금지원이 기업의 연구개발 지출을 통계적으로 유의하게 높이는 것으로 나타났다. 매칭방법에 따른 분석결과와의 비교는 CEM, MDM, PSM 방법 등 5가지의 매칭방법이 활용되었으며, 비교 결과 매칭된 집단의 공변량 간의 차이와 L_1 값으로 측정되는 불균형 측면에서 CEM 방법이 다른 매칭방법보다 더 나은 매칭결과를 생성하는 것으로 나타났다. 물론 CEM 방법의 경우 처치그룹을 모두 매칭에 활용하지 않는 단점 역시 존재하기 때문에 CEM이 다른 매칭방법에 비해 절대적으로 우월하다는 것은 아니다. 다만, 비교대상 관측치가 상대적으로 많고, 매칭대상 공변량에 연속

변수가 포함될 경우 CEM은 불균형이 최소화된 매칭결과를 생성한다는 점에서 향후 많은 활용이 가능한 방법론으로 생각된다.

기존의 많은 국내 정책평가 연구들에서 매칭방법은 가장 활발히 활용되는 분석방법 중 하나이나, 대체로 한 종류의 매칭방법을 통해 분석결과를 제시하고 있으며, 이 방법 또한 PSM에 집중되어 있는 경향이 있다. 그러나 Ho, Imai, King, and Stuart (2007) 이 충고하였듯이 매칭분석을 활용할 경우 “가능한 많은 매칭방법을 시도해야 하고 최적균형을 가지는 방법을 선택해야 한다.”는 말처럼 실제 분석에서는 분석자료의 특성에 따라 다양한 방법을 적용한 후 이들 결과를 비교함으로써 가장 적합한 매칭방법의 선택이 필요하며, 추정결과의 강건성을 확보하기 위해서 CEM을 포함한 다양한 분석결과가 같이 제시되어야 할 것이다.

■ 참고 문헌

1. 김기현 · 신인철, “유아교육 및 보육 경험의 장기 효과,” 『한국사회학』, 제46권 제5호, 2012, pp. 259-288.
(Translated in English) Kim, Ki-Hun and In-Chol Shin, “Kindergarten and Child Care Long-Term Effects: Peer and Teacher Relationships, and Academic Achievement,” *The Korean Journal of Sociology*, Vol. 46, No. 5, pp. 259-288.
2. 김 호 · 김병근, “정부보조금의 민간연구개발투자에 대한 효과분석,” 『기술혁신학회지』, 제15권 제3호, 2012, pp. 649-674.
(Translated in English) Kim, Ho and Byung-Kuen Kim, “Analyzing the Effectiveness of Public R&D Subsidies on Private R&D Expenditure,” *Journal of Korea Technology Innovation Society*, Vol. 15, No. 3, pp. 649-674.
3. 이재형 · 양정삼 · 이상무, 『시장구조조사』, 한국개발연구원, 2013.
(Translated in English) Lee, Jae-Hyung, Jeong-Sam, Yang and Sang-Moo Lee, *Market Structure Research*, Korea Development Institute, 2013.
4. 조가원, 『연구개발 정부보조금과 기업의 혁신성과』, 교육과학기술부, 2009.
(Translated in English) Jo, Ga-Won, *Innovation and Subsidies: Firm Level Analysis*, Ministry of Education, Science and Technology, 2009.
5. Blackwell, M., S. Iacus, G. King and G. Porro, “CEM: Coarsened Exact Matching in Stata,” *The Stata Journal*, Vol. 9, No. 4, 2010, pp. 524-546.

6. Dimos and Pugh, "The Effectiveness of R&D Subsidies: A Meta-regression Analysis of the Evaluation Literature," *Research Policy*, Vol. 45, No. 4, 2016, pp.797-815.
7. Heckman, J. J. and J. A. Smith, "Assessing the Case for Social Experiments," *The Journal of Economic Perspectives*, Vol. 9, No. 2, 1995, pp.85-110.
8. Heckman, J. J., "Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations," *The Journal of Human Resource*, Vol. 32, No. 3, 1997, pp.441-462.
9. _____, "Chapter 70 Econometric Evaluation of Social Programs, Part 1: Causal Models, Structural Models and Econometric Policy Evaluation," 2007.
10. Ho, D. E., K. Imai, G. King and E. A. Stuart, "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference," *Political Analysis*, Vol. 15, 2007, pp.199-236.
11. Holland, P. W., "Statistics and Causal Inference," *Journal of American Statistical Association*, Vol. 81, No. 396, 1986, pp.945-960.
12. Iacus, S. M., G. King and G. Porro, "CEM: Software for Coarsened Exact Matching," *Journal of Statistical Software*, Vol. 30, No. 9, 2009, pp.1-27.
13. _____, "Multivariate Matching Methods That Are Monotonic Imbalance Bounding," *Journal of American Statistical Association*, Vol. 106, No. 493, 2011.
14. _____, "Causal Inference without Balance Checking: Coarsened Exact Matching," *Political Analysis*, Vol. 20, No. 1, 2012, pp.1-24.
15. Ichino, A., "The problem of causality in microeconometrics," Available at http://www.cafed.sssup.it/~federico/etrix_allievi/Ichino_causality_new.pdf, 2007.
16. King, G., R. Nielsen, C. Coberley, J. E. Pope and A. Wells, "Comparative Effectiveness of Matching Methods for Causal Inference," Available at <http://gking.harvard.edu/files/psparadox.pdf>, 2011.
17. OECD, "ISIC REV.3 Technology Intensity Definition," 2011.
18. Qin, S., "Comparing the Matching Properties of Coarsened Exact Matching, Propensity Score Matching, and Generic Matching in a Nationwide Observational Data and a Monte Carlo Experiment," The University of Georgia, Master's paper, 2011.
19. Rubin, D. B., "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, Vol. 66, No. 5, 1974, pp.688-701.
20. _____, "Inference and Missing Data," *Biometrika*, Vol. 63, No. 3, 1976a, pp.581-592.
21. _____, "Multivariate Matching Methods That are Equal Percent Bias Reducing, I: Some Examples," *Biometrics*, Vol. 32, No. 1, 1976b, pp.109-120.
22. Rubin, D. B. and N. Thomas, "Affinely Invariant Matching Methods with Ellipsoidal Distributions," *The Annals of Statistics*, Vol. 20, No. 2, 1992, pp.1079-1093.
23. Stuart, E. A., "Matching Methods for Causal Inference: A Review and a Look Forward," *Statistical Science*, Vol. 25, No. 1, 2010, pp.1-21.

{ Appendix }

〈Table A〉 Classification of manufacturing industries into categories based on OECD

Classification	Industry (KSIC 9)
Low Technology Industry	10. Manufacture of Food Products 11. Manufacture of Beverage 13. Manufacture of Textiles 14. Manufacture of Wearing Apparel, Clothing Accessories and Fur Articles 15. Tanning and Dressing of Leather, Manufacture of Luggage and Footwear 16. Manufacture of Wood Products of Wood and Cork 17. Manufacture of Pulp, Paper and Paper Products 18. Printing and Reproduction of Recorded Media 32. Manufacture of Furniture 33. Other Manufacturing
Middle-Low Technology Industry	19. Manufacture of Coke, Hard-Coal and Lignite Fuel Briquettes and Refined Petroleum 22. Manufacture of Rubber and Plastic Products 23. Manufacture of Other Non-Metallic Mineral Products 24. Manufacture of Basic Metal Products 25. Manufacture of Fabricated Metal Products
Middle-High Technology Industry	20. Manufacture of Chemicals and Chemical Products 28. Manufacture of Electrical Equipment 29. Manufacture of Other Machinery and Equipment 30. Manufacture of Motor Vehicles, Trailers and Semitrailers 31. Manufacture of Other Transport Equipment
High Technology Industry	21. Manufacture of Pharmaceuticals, Medicinal Chemicals and Botanical Products 26. Manufacture of Electronic Components, Computer, Radio, Television and Communication Equipment and Apparatuses 27. Manufacture of Medical, Precision and Optical Instruments, Watches and Clocks

Understanding and Using the CEM Method

- Focusing on the Effect of Public R&D Subsidies -

Sangsin Kim*

Abstract

This study introduces CEM(Coarsened Exact Matching) method and evaluate the effects of public R&D subsidies on firm's R&D expenditure by using CEM. Also, it provides a basis on which CEM method can be variously utilized in future program evaluations by presenting the advantages and disadvantages of CEM through comparison with other matching methods such as PSM and MDM. As one of the matching method, CEM has the following advantages compared to other matching methods. First, the analyst can control the imbalance occurring in the matching process in advance. Second, by limiting the matched data with the common empirical support, CEM can reduce the modeling dependence without iterative matching procedures. In comparison with other matching methods, CEM generated the smallest imbalance. However, CEM can't use some of the observations in treatment group due to non-matching results. Although CEM has the above disadvantage, I think that CEM should be extensively used in terms of reducing imbalance in the matching process.

Key Words: program evaluation, CEM(Coarsened Exact Matching), PSM(Propensity Score Matching)

JEL Classification: H2, C4

Received: April 20, 2016. Revised: July 26, 2016. Accepted: Aug. 16, 2016.

* Research Fellow, Korea Development Institute, 263 Namsejong-ro, Sejong-si 30149, Korea, Phone: +82-44-550-4052, e-mail: sskim@kdi.re.kr