석사학위논문

다중 처리군에서
성향점수 가중치 방법의 비교

2022년 2월 22일

전 북 대 학 교 대 학 원

통계학과

김 은 진

다중 처리군에서 성향점수 가중치 방법의 비교

Comparing Weighting Methods in Propensity Score Analysis for Multiple Treatments

2022년 2월 22일

전 북 대 학 교 대 학 원

통계학과

김 은 진

다중 처리군에서
성향점수 가중치 방법의 비교

지도교수 최 혜 미

이 논문을 이학 석사 학위논문으로 제출함.

2021년 10월 20일

전 북 대 학 교 대 학 원

통계학과

김 은 진

김은진의 석사학위논문을 인준함.

위 원 장	전북대학교 조교수	최규빈	(인)
부위원장	전북대학교 조교수	양성준	(인)
위 원	전북대학교 교 수	최혜미	(인)

2022년 1월 4일

전 북 대 학 교 대 학 원

차 례

그림 차례

표 차례

ABSTRACT

Comparing Weighting Methods in Propensity Score Analysis for Multiple Treatments

Kim, Eun-jin

Department of Statistics

The Graduate School

Jeonbuk National University

In observational studies, the treatment effect estimates are often biased by the impact of covariate imbalances. The propensity score weighting methods such as inverse probability of treatment weights(IPW), overlap weights(OW) and matching weights(MW) can be used to reduce the bias. In the case of multiple treatments, the propensity scores are extended to generalized propensity scores. They have been estimated most often by multinomial logistic regression(MLR) and used for weighting methods. However, MLR, which is a parametric methods, may cause bias in the estimate of treatment effect when there are too many covariates for the sample size or

when normality cannot be assumed. The generalized boosted models(GBM) is robust to such problems as a nonparametric methods and also implemented well in very large size of data so that it is recently adopted for the estimation of GPS.

In this study, we compared small sample properties of generalized propensity scores estimated by MLR and GBM and three weighting methods(IPW, OW, MW) based on these scores through simulations. The result of simulation shows; The MLR had small bias and MSE, and CP was close to the specified value only when log-odds of the treatment or outcome was linear in covariates. The GBM, however, showed consistent results regardless of the models specification in bias, MSE and CP, and performs better than the MLR as both treatment and outcome models are non-linear. The weighting methods showed good performance in the order of MW, OW and IPW. MW and OW showed similar results according to the generalization propensity score estimation methods, and IPW performed better when estimated using the GBM.

Keywords : generalized propensity score, multiple treatments, weighting methods, multinomial logistic regression, generalized boosted models

제 1 장 서 론

임상연구에서 무작위 배정 시험(randomized controlled trial)은 처리군과 통제군간 효과의 차이를 편향 없이 추정하기 위한 시험설계이나 현실에서는 윤리적인 문제, 환자의 선호, 비용 등의 이유로 인해 시행하기 어렵다. 비무작위 배정 시험인 관찰연구(observational study)는 처리군과 통제군의 효과를 비교하는 연구를 설계 할 수 있으나 연구 설계 단계에서 공변량의 불균형으로 인한 편향(bias)이 발생하게 된다. Rosenbaum and Rubin [1983]은 관찰연구에서 공변량의 균형을 보정하여 편향을 줄이기 위해 성향점수(propensity score)를 제안하였다. 그러나 Rosenbaum and Rubin [1983]의 성향점수는 처리군이 이분형인 경우만 고려하였고, 실제 연구 목적에서는 다중 처리군의 비교가 필요한 경우도 많다. 예를 들어, 암, 우울증, 천식, 만성질환, 당뇨 등의 질병들은 복잡성과 합병증으로 인해 두 가지 이상의 치료요소가 포함되어 다중 처리의 비교가 필요하다. 질병의 치료를 위해 수술, 약물, 방사선 등 세 가지 이상의 치료방법을 취할 수 있고, 세 가지 이상의 다른 기간이나 약물 또는 용량 등을 적용할 수 있다 (Feng et al. [2012]).

Imbens [2000]는 기존의 성향점수 방법을 확장하여 처리군이 다범주인 경우에 적용할 수 있는 일반화 성향점수(generalized propensity score) 방법을

제안하였다. 각 처리 수준의 일반화 성향점수를 활용하여 해당 처리군의 평균 잠재적 결과를 개별적으로 추정할 수 있다. 성향점수 추정을 위해 주로 로지스틱 회귀 모형, 프로빗 모형 등과 같은 모수적 방법을 사용한다. 그러나 실제 자료에서는 표본의 크기에 비해 공변량의 수가 많거나 정규성을 가정 할 수 없으며, 이를 모수적 방법으로 추정할 경우 치료효과의 추정량에 편향이 발생할 수 있다. 이를 해결하기 위해, McCaffrey et al. [2004]는 비모수적 방법인 일반화 부스팅 모형(generalized boosted models)을 사용하여 이분형 처리에 대한 성향점수 추정을 제안하였으며, 이를 확장한 McCaffrey et al. [2013]에서는 다범주 처리군의 경우 일반화 부스팅 모형을 사용하여 성향점수를 추정하는 방법을 제안하였다. 일반화 부스팅 모형은 처리 할당과 공변량 사이의 복잡하고 비선형적인 관계를 추정하기 위해 다중 회귀 트리를 사용하는 절차를 반복하여 처리군 사이의 최상의 균형(balance)을 이끌어내는 성향점수 모형을 찾을 수 있다.

추정된 성향점수는 주로 매칭(matching), 층화(stratified), 가중치(weighting) 등의 방법을 사용하여 공변량의 균형을 보정하고 처리효과를 추정한다. 매칭 방법은 각 처리군간 추정된 성향점수가 비슷한 피험자를 매칭 시키는 방법으로 공변량의 균형을 맞추어 편향을 감소시키나 매칭되지 않은 표본은 제외되어 많은 수의 표본이 필요하다. 또한 처리군간 공변량의 불균형이 클수록 정보 손실이 많아 추정의 검정력이 떨어질 수 있다 (King and Nielsen [2019]). 층화 방법은 범위에 따라 몇 개의 간격으로 비슷한 성향점수를 가지는 피험자들을 같은 계층에 모아 처리효과를 추정한 후 층의 크기에 비례하게

가중치를 부여하여 전체 처리효과를 추정하는 방법이다. 같은 층에 있는 각 처리군의 피험자들은 공변량이 유사하기 때문에 바로 비교할 수 있으나 충분한 표본 확보가 필요하며 연구 설계가 복잡한 경우 적용이 힘들 수 있다 (Adelson et al. [2017]). 가중치 방법은 각 처리군 간의 공변량 분포의 균형을 맞추기 위해 성향점수를 적용한 가중치를 부여하여, 표본을 특정한 모집단을 나타내는 유사모집단으로 만드는 방법이다. 모든 표본을 그대로 사용하기 때문에 다른 방법들에 비해 추정의 검정력이 높아질 수 있으며, 복잡한 연구 상황과 분석 모형 설정에 적용하기 적합한 방법이다. 가중치 방법에서는 기본적으로 역확률 가중치(inverse probability of treatment weights) 방법이 가장 널리 알려져 있으며 McCaffrey et al. [2013]에서도 일반화 부스팅 모형으로 성향점수 추정 후 역확률 가중치를 사용하여 공변량을 보정하였다. 하지만, 역확률 가중치는 처리 수준에 관계없이 0 또는 1에 가까운 극단적인 성향점수를 가질 경우 과도한 편향과 오차가 생길 수 있다. 이를 보완하기 위해, Li and Greene [2013]와 Yoshida et al. [2017]는 과도하게 부여되는 가중치는 줄이고 수칙적 안정성을 부여하는 매칭 가중치(matching weights)를 제안하였으며, Li et al. [2018]와 Li [2019]에서는 각 가중치가 대조군 그룹에 할당될 확률에 비례하는 중복 가중치(overlap weights)을 제안하였다.

본 논문에서는 처리군이 다범주인 경우 두 가지의 일반화 성향점수 추정 및 세 가지의 가중치 방법의 성능을 비교하고자 한다. 일반화 성향점수 추정 방법으로는 다항 로지스틱 회귀 모형과 일반화 부스팅 모형, 가중치 방법으로는 역확률 가중치, 중복 가중치, 매칭 가중치 방법을 모의실험을 통해 비교하였다.

본 논문의 구성은 다음과 같다. 1장에서는 연구의 배경 및 목적을 소개하였고, 2장에서는 다중 처리군에서의 성향점수와 성향점수 추정 방법, 가중치 방법 대한 개념을 설명하였다. 3장에서는 2장에서 소개한 방법들을 비교하기 위한 모의실험 설계 및 모형 평가 방법과 분석 결과를 제시하였다. 마지막으로 4장은 모의실험을 통해서 본 연구에 대한 결론을 기술하였다.

제 2 장 연구방법

무작위 배정 연구에서 피험자는 무작위로 서로 다른 처리군에 할당되며 평균적으로 다른 그룹의 처리군 간 공변량은 평균적으로 차이가 없다. 그러나 관찰연구와 같은 비무작위 배정 연구에서는 다른 처리군의 피험자 간의 공변량 차이는 통제되기 어렵다. 결과적으로, 우리가 비교하고자 하는 집단들은 관찰된 공변량에서 상당한 차이를 보일 수 있으며, 이는 처리효과에 대한 편향된 추정으로 이어질 수 있다. 따라서 비무작위 배정 연구에서는 편향의 제거를 위해 공변량 보정 후 처리효과를 추정해야 한다.

제 1 절 다중 처리군의 성향점수

Rosenbaum and Rubin [1983]은 비무작위 배정 연구에서 공변량을 보정하기 위해 성향점수(propensity score; PS) 방법을 제안하였으며, 이를 해당 공변량 X에 의해 처리군 Z로 할당될 조건부확률이라 정의했다.

$$e(x) = Pr(Z = 1|X = x), \quad Z \in \{0, 1; 0 = \text{대조군}, 1 = \text{처리군}\}$$

성향점수를 통해 처리 전 공변량의 차이와 관련된 편향이 제거되지만 처리군이 이분형인 경우에만 적용할 수 있는 한계점이 있다. 이를 해결하기 위해, Imbens [2000]는 처리군이 다범주인 경우에 적용할 수 있는 일반화 성향점수(generalized propensity score; GPS) 방법을 제안하였다. $J(J \geq 3)$개의 처리군(treatments)을 나타내는 Z, 즉 $Z \in \{1, \ldots, J\}$, k개의 공변량 $\mathbf{X} = x$가 주어졌을 때, 일반화 성향점수를 다음과 같이 정의했다.

$$e_j(x) = Pr(Z = j|X = x), \quad j \in Z = \{1, \cdots, J\}$$

여기에서 $\sum_{j=1}^{J} e_j(x) = 1$임을 알 수 있다.

i번째 피험자가 처리 j를 받는 경우의 잠재 결과변수(potential outcome)는 $Y_i(j)$라하며, 받은 처리 Z_i에 해당하는 $Y_i = Y_i(Z_i)$만 관측된다. 공변량의 주변밀도함수를 $f(x)$라고 할 때, 목표 모집단에서의 공변량 밀도함수는 $g(x) = f(x)h(x)$로 나타낼 수 있는데, 여기에서 $h(x)$는 사전에 명시된 공변량의 함수인 기울기 함수(tilting function)이다. 처리군 j에서 공변량의 밀도함수는 성향점수와 다음과 같은 관계가 성립한다.

$$f_j(x) = f(x|Z = j) \propto f(x)e_j(x)$$

위와 같은 관계로부터 $h(x)$는 주로 성향점수 $e_j(x)$를 이용하여 결정한다. 처리 j를 받은 잠재적 결과 $Y(j)$의 목표 모집단에서의 기댓값은 조건부 기댓값을

$m_j(x) = E[Y(j)|X = x]$로 나타낼 때 다음과 같이 정의한다.

$$\mu_j^h = E_g[Y(j)] = \frac{E[h(X)m_j(X)]}{E[h(X)]}$$

처리군간 처리효과를 확인하기 위하여 μ_j^h들의 선형결합은 다음과 같이 정의될 수 있다.

$$\tau^h(\mathbf{a}) = \sum_{j=1}^{J} a_j \mu_j^h, \quad \mathbf{a} = (a_i, \cdots, a_J)$$

예를 들어 j와 j'의 평균처리효과의 차이는 $\mu_j^h - \mu_{j'}^h$로 표현할 수 있다. $h(x)$가 $e_j(x)$를 이용하여 결정이 되므로, μ_j^h는 성향점수를 이용하여 추정될 수 있다 (Li [2019]; Zhou et al. [2020b]). 성향점수를 이용하여 처리효과를 추정하기 위해 모든 $j \in Z$에 대하여 공변량 X가 주어졌을 때 잠재적 결과변수인 $Y(j)$와 $\mathbb{1}\{Z = j\}$는 독립이며, $0 < e_j(x) < 1$의 가정이 필요하다 (Rosenbaum and Rubin [1983]; Imbens [2000]). 다음 절에서는 이러한 가정을 만족하는 성향점수 추정방법에 대하여 살펴보았다.

제 2 절 성향점수 추정 모형

2.1 다항 로지스틱 회귀 모형

처리군이 3개 이상인 경우 Rubin [1997]은 두 개씩 쌍을 이루어 처리효과를 비교하기 위해 각 쌍별로 별도의 성향점수 모형을 만들 것을 제안하였는

데, 이 방법은 모든 처리군에서 선택할 확률이 1보다 커질 수 있으며, Imbens [2000]가 제안한 다항 로지스틱 회귀 모형 방법에 비해 덜 효율적이다. Imbens [2000]는 다중 처리군이 명목형인 경우 다항 로지스틱 회귀 모형(multinomial logistic regression; MLR)으로 추정할 수 있는 일반화 성향점수를 제안하였다. 일반화 성향점수는 관찰된 공변량이 주어졌을 때 다중 처리군 중에 하나의 처리를 받을 조건부 확률로 정의된다. 여기에서 각 피험자에 대해 관찰된 공변량을 기준으로 각 처리범주를 받을 확률이 추정된다. Imbens [2000]에서는 기존의 성향점수와 마찬가지로 일반화 성향점수가 균형점수이며 공변량 X의 전체에 대한 조건화 대신 일반화 성향점수에 대한 조건화로 충분하다는 것을 이론적으로 증명하였다. 따라서 다항 로지스틱 회귀 모형으로 추정한 일반화 성향점수를 사용하여 다중처리 비교에서 유효한 추정치를 얻을 수 있다 (Spreeuwenberg et al. [2010]).

$$P(Z = j|X = x) = \frac{e^{\beta_j' x}}{1 + \sum_{j=1}^{J-1} e^{\beta_j' x}}, \quad j = 1, \ldots, J-1$$
$$P(Z = J|X = x) = \frac{1}{1 + \sum_{j=1}^{J-1} e^{\beta_j' x}}$$

β_j를 추정하고 이를 통해 추정된 성향점수 $e_j(x)$를 적용하여 처리효과를 추정할 수 있다.

2.2 일반화 부스팅 모형

일반화 부스팅 모형(generalized boosted models; GBM)은 결정 트리 알고리즘(decision tree algorithm)과 부스팅(boosting) 방법의 결합으로, 결측치나 이상치에 강건하고 변수의 개수가 매우 많거나 빅데이터를 다룰 때 더욱 강력한 모형이다. 모형의 정확도를 향상시키기 위하여 여러 결정 트리를 반복적으로 적합 하는데, 단일 단순회귀트리로 시작하여 반복하면서 다른 트리가 추가되는 반복 적합 알고리즘이다. 새로운 트리는 이전 반복에서의 모형 잔차에 가장 잘 맞는 것으로 선택 된다. 즉 선택된 트리는 데이터에 대한 로그 가능도를 가장 크게 증가시킨다. 트리들을 결합할 때, 각 트리가 수축률(shrinkage rate)에 의해 조절되어 결과 모형의 매끄러움과 전반적인 적합도를 향상시키게 된다. 실제 알고리즘은 일반적으로 반복적인 트리 추가로 데이터에 과적합될 수 있는데, 이를 피하기 위해 일반화 부스팅 모형은 최종 모형에 대한 중간 반복(intermediate iteration)(또는 트리 수)을 선택하여 표본외 예측오류(out-of-sample prediction error) 또는 처리군간 공변량의 불균형과 같은 외적기준(external criteria)를 최소화 하게 된다.

Friedman [2001]과 Madigan and Ridgeway [2004]는 예측 오류 측면에서 부스팅이 다른 방법들보다 더 나은 성능을 보인다는 것을 보여주었으며, AdaBoost 알고리즘 (Freund and Schapire [1997]), 일반화 부스팅 모형 (Ridgeway [1999]), 로짓부스트 (Friedman et al. [2000]) 및 Gradient 부스팅 머신 (Friedman [2001])을 포함하여 많은 부스팅 모형이 연구되었다. 부스팅은 모형

이 공변량의 개수가 클 때 매우 효과적이며 (Bühlmann and Yu [2003]), 일반화 부스팅 모형은 다른 부스팅 방법과 달리 잘 보정된 확률 추정치를 산출하는 모형을 생성하도록 조정된다.

McCaffrey et al. [2004]는 일반화 부스팅 모형으로 이진 처리군의 성향점수를 추정하기 위해 베르누이 로그-우도 함수를 이용하였다.

$$l(u) = \sum_{i=1}^{N} Z_i u(X_i) - \log(1 + \exp(u(X_i)), \quad Z_i \in \{0, 1\}$$

여기에서 $u(x) = \log\left[\frac{p(x)}{1-p(x)}\right]$가 공변량의 선형함수이면 로지스틱 회귀 모형에 해당한다. $u(x)$가 음수일 때 $Z = 0$이거나 $u(x)$가 양수일 때 $Z = 1$이면 $l(u)$는 상대적으로 큰 값이며 잘 적합 된다고 할 수 있다. 일반화 부스팅 모형 알고리즘을 통한 이진 처리군의 성향점수를 추정하는 방법은 다음과 같다.

(i) 계산을 단순화하기 위해 일반화 부스팅 모형은 성향점수를 직접 모형화하는 대신 처리 할당의 로그-오즈로 모형화하여 알고리즘의 초기값으로 설정

$$u(x) = \log\left[\frac{\bar{Z}}{1-\bar{Z}}\right], \quad \bar{Z} = \text{전체 표본에 대한 처리 할당 평균}$$

(ii) 알고리즘이 데이터에 대한 성향점수 모형의 적합성을 향상시킬 수 있는

조정값 $s(x)$를 추정하여 $u(x)$에 추가

$$\widehat{u}(x) \leftarrow \widehat{u}(x) + \lambda s(x)$$

이 과정을 반복하여 최적의 $\widehat{u}(x)$로 적합

$s(x)$로는 주로 현재 적합에서 잔차 $r_i = Z_i - 1/(1+\exp(-\hat{u}(x))$를 공변량의 함수로 모형화하는 회귀 트리로 선택한다.

McCaffrey et al. [2013]에서는 일반화 부스팅 모형을 통해 이진 처리군의 성향점수 추정 방법을 확장하여 다중 처리군에서도 일반화 성향점수를 추정하는 방법을 제안하였다. McCaffrey et al. [2013]는 일반화 부스팅 모형을 반복적으로 적합하여 각 처리군 $Z_i = j$ 또는 $Z_i(j) = 1$에 대해 최적의 균형을 산출하는 성향점수 $\hat{p}_j(X_i)$를 추정하고, 추정된 성향점수 가중치를 적용하여 모집단 표준화 편향(population standardized bias; PSB)를 구하였다.

$$PSB_{j,k} = \frac{|\bar{X}_{kj} - \bar{X}_{kp}|}{\hat{\sigma}_{kp}}, \quad \bar{X}_{kj} = \frac{\sum_{i=1}^{n} Z_i(j) X_{ki} / \hat{p}_j(X_i)}{\sum_{i=1}^{n} Z_i(j) / \hat{p}_j(X_i)}, \quad j = 1, \ldots, J$$

$k = 1, \ldots, K$는 공변량의 수, $\bar{X}_{kj}$는 성향점수 가중치를 적용한 공변량의 평균, $\hat{p}_j(X_i)$는 일반화 부스팅 모형의 적합으로 추정된 성향점수, $\bar{X}_{kp}$와 $\hat{\sigma}_{kp}$는 원 자료의 모든 처리군의 합동표본에 대한 공변량의 평균과 표준편차를 나타낸다. 이러한 알고리즘을 반복하여 $PSB_{j,k}$의 최댓값 또는 평균값을 최소화하는 최적의 일반화 부스팅 모형을 적합하여 일반화 성향점수를 추정한다.

제 3 절 가중치 방법

가중치 방법은 각 처리군간의 공변량 분포의 균형을 맞추기 위해 일반화 성향점수를 이용하여 구한 가중치를 목표 모집단에 적용하여 처리효과를 추정하는 방법이다. 성향점수 가중치 방법(weighting methods)은 공변량의 균형(balancing covariates)뿐만 아니라 인과추론(causal inference)에서 공변량의 균형 처리효과 추정을 위한 일반적인 방법이다(Rosenbaum and Rubin, 1983). 두 범주 처리의 경우 성향점수 가중치 방법에 대한 폭 넓은 연구결과들이 있는데, 이에 대하여는 최근에 Ding and Li [2018]가 정리한 논문을 참고할 수 있다. 본 논문에서는 그 필요가 증가하고 있는 다범주 처리군의 비교를 위한 성향점수 가중치 방법을 정리해보았다.

Li [2019]는 다범주 처리의 경우 $Z = j$ 집단에서의 공변량 X의 밀도함수를 $f_j(x) = f(X|Z = j) \propto f(x)e_j(x)$인 관계와 주어진 임의의 공변량 기울기 함수 $h(x)$에 대하여 목표 모집단에서의 확률밀도함수 $f(x)h(x)$를 이용한 가중치를 적용하여 균형 가중치(balancing weights)를 정의하였다.

$$w_j(X) \propto \frac{f(X)h(X)}{f(X)e_j(X)} = \frac{h(X)}{e_j(X)}, \quad \forall \ j \in Z$$

즉, 균형가중치는 $h(x)$에 의존한다. $w_j(x)$를 적용한 μ_j^h에 대한 Hájek 추

정량은 다음과 같으며, $\widehat{\mu_j^h}$를 기반으로 처리효과를 추정 할 수 있다.

$$\widehat{\mu_j^h} = \frac{\sum_{i=1}^{N} w_j(x_i) D_{ij} Y_i}{\sum_{i=1}^{N} w_j(x_i) D_{ij}}$$

본 연구에서는 성향점수를 이용하는 여러 균형 가중치들 중 자주 사용되는 역확률 가중치, 매칭 가중치와 중복 가중치를 이용한 평균처리효과(average treatment effect; ATE) 추정을 비교하고자 한다.

역확률 가중치(inverse probability of treatment weights; IPW)

성향점수를 표본가중치로 활용하는 역확률 가중치는 Rosenbaum [1987]이 제안한 방법으로 각 처리군에 할당될 확률이 높은 피험자가 과대 표집되는 상황을 완화하기 위해 가중치로 공변량을 조정한 후 비편향된 평균처리효과를 추정하는 방법이다. 역확률 가중치를 통해 공변량의 균형이 맞는 유사집단을 형성할 수 있다. 각 처리군에 할당될 확률인 성향점수의 역수를 가중치로 가지며, $h(x)$는 1이다. 관측된 데이터를 그대로 사용할 수 있어 다른 성향점수 방법에 비해 정보 손실이 적으며, 가중치 방법 중 기본적으로 가장 많이 사용되는 방법이다. 역확률 가중치 방법은 제거되는 표본이 없어, 모든 피험자를 분석에 사용할 수 있는 장점이 있다 (Horvitz and Thompson [1952]; Robins

et al. [2000]; Hirano and Imbens [2001]).

$$w_j^{ipw}(x) = \frac{1}{e_j(x)}, \quad j = 1, \ldots, J$$

매칭 가중치(matching weights; MW)

역확률 가중치 방법은 범위가 $(0, \infty)$로 성향점수가 높거나 낮은 관측치에 비정상적으로 큰 가중치를 부여하며, 공변량의 중첩에 크게 의존한다는 문제가 있다. 이를 해결하기 위해, Li and Greene [2013]는 과도하게 부여되는 가중치는 줄이고 수칙적 안정성을 부여하는 매칭 가중치를 제안하였으며, 이를 확장하여 Yoshida et al. [2017]는 다중 처리군에서 사용할 수 있는 일반화 매칭 가중치를 제안하였다. 매칭 가중치는 역확률 가중치에서 확장되었으며, $h(x) = \min_t[e_t(x)]$이다. 매칭 가중치의 범위는 $(0, 1)$으로 비정상적으로 큰 가중치가 부여되는 문제를 해결하였다.

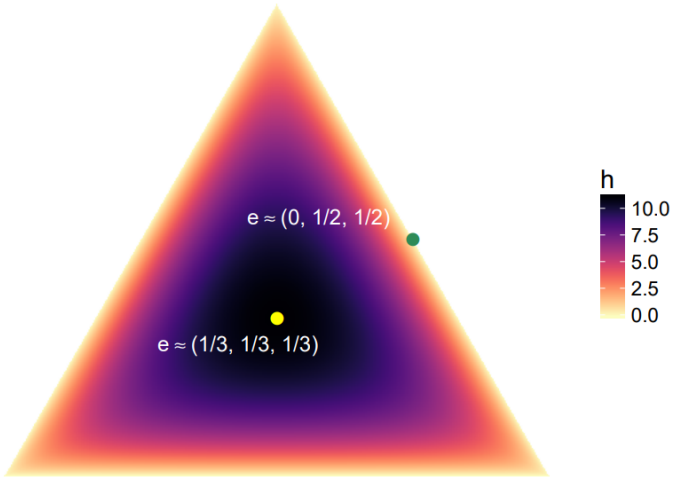
$$w_j^{mw}(x) = \frac{\min_t[e_t(x)]}{e_j(x)}, \quad j = 1, \ldots, J$$

중복 가중치(overlap weights; OW)

Li et al. [2018]는 처리군이 이분형일 때, 반대되는 처리군의 성향점수를 가중치로 부여하는 중복 가중치를 제안하였다. 즉, 처리군의 성향점수를 $e(x)$라 하였을 때, 처리군에 $1 - e(x)$, 대조군에는 $e(x)$를 가중치로 부여한다. Li

[2019]에서는 다중 처리군의 경우에 적용할 수 있는 역확률 가중치와 일반화 성향점수의 조화평균의 곱으로 구성된 일반화 중복 가중치를 추가로 제안하였다. $h(x) = \left[\sum_{t=1}^{J} 1/e_t(x)\right]^{-1}$이며, 그림 2.1은 $J = 3$인 $h(x)$의 삼원 플롯을 나타낸다.

그림 2.1: Ternary plot of optimal h(up to a proportionality constant) as a function of the generalized propensity score vector with $J = 3$ treatments.



Each point in the triangular plane represents a unit with certain values of the generalized propensity scores. The value of each generalized propensity score is proportional to the orthogonal distance from that point to each edge. It is evident that the new weighting scheme emphasizes the centroid region with good overlap, e.g., units with $e(X) \approx (1/3, 1/3, 1/3)$, and smoothly downweights the edges, e.g., units with $e(X) \approx (0, 1/2, 1/2)$.

최적의 기울기 함수 $h(x)$는 모든 처리그룹에서 성향점수가 0에 가깝지 않은 공변량의 영역에 상대적으로 가장 높은 가중치를 부여하는 함수이다. 적어도 한 차원에서 중첩되지 않는 영역에 대한 가중치는 줄여서, 해당하는 목표

모집단을 모든 처리군에서 공변량이 가장 많이 겹치는 부분 모집단으로 해석할 수 있다. 중복 가중치 범위가 $(0,1)$으로 제한되어있어 극단적인 성향점수의 문제를 해결할 수 있으며, 가중평균처리효과 추정시 점근적 분산을 최소화한다.

$$w_j^{ow}(x) = \frac{\left[\sum_{t=1}^{J} 1/e_t(x)\right]^{-1}}{e_j(x)}, \quad j = 1, \ldots, J$$

제 3 장 모의실험

제 1 절 모의실험 설계

일반화 성향점수 추정법으로는 다항 로지스틱 회귀 모형(MLR)과 일반화 부스팅 모형(GBM), 각 일반화 성향점수 추정법을 이용한 가중치 방법으로는 역확률 가중치(IPW), 중복 가중치(OW)와 매칭 가중치(MW) 방법, 즉 $2 \times 3 = 6$개의 일반화 성향점수 추정 및 가중치 방법들의 성능을 다양한 표본의 크기와 모형 가정 하에서 모의실험을 통해 비교하였다. 자료 생성 모형은 Brown et al. [2020]의 모의실험의 설정을 적용하였다.

세 개의 범주인 처리군 변수(Z), 이분형 처리결과 변수(Y)와 9개의 공변량(X)을 가진 모형에 대하여 표본의 크기 200, 500과 1000인 경우를 고려하였다. 이 때 처리군 모형과 처리결과 모형은 각각 로그오즈에 대한 선형 및 비선형의 경우를 가정하고 표본을 추출하였다. 처리결과는 9개의 공변량 중 6개의 공변량 $X_1, X_2, X_3, X_4, X_5, X_6$과 연관성이 있으며, 처리군은 6개의 공변량 $X_1, X_2, X_4, X_5, X_7, X_8$과 연관성이 있는 모형을 설정하였다. 9개의 공변량$(X_1, \ldots, X_9)$은 $\mu = 0,\ \sigma^2 = 1,\ Cov[X_i, X_j] = 0.2$인 다변량 정규분포를 따르도록 하였다.

표 3.1: True association between baseline covariates with treatment and outcomes

		treatment covariates		
		Strongly associated	Moderately associated	Independent
Outcome covariates	Strongly associated	X_1	X_2	X_3
	Moderately associated	X_4	X_5	X_6
	Independent	X_7	X_8	X_9

Note: X_1, X_2, X_4 and X_5 are simulated to be pretreatment confounders.

세 범주 처리변수와 이분형 처리결과 변수는 다음의 각 모형에서 계산된 확률을 가지는 다항 분포 및 베르누이 분포로부터 그 값을 추출하였다. 처리와 결과 각각 로그오즈 척도에 대한 선형성 여부, 즉 두 종류씩, 총 네 종류의 모형 M1, M2, M3와 M4 하에서 일반화 성향점수 추정 및 가중치 방법들의 성능을 비교하였다. 수식을 간결하게 표현하기 위해 먼저 (A)와 (B)를 다음과 같이 정의하였다.

$$
\begin{aligned}
(A) &= \beta_{1,j}x_{i,1} + \beta_{2,j}x_{i,2} + \beta_{4,j}x_{i,4} + \beta_{5,j}x_{i,5} + \beta_7 x_{i,7} + \beta_8 x_{i,8} \\
&\quad (\beta_{1,1}, \beta_{1,2}) = (\beta_{4,1}, \beta_{4,2}) = (0.7, 0.4), \\
&\quad (\beta_{2,1}, \beta_{2,2}) = (\beta_{5,1}, \beta_{5,2}) = (0.2, 0.3),\ \beta_7 = 0.6, \beta_8 = 0.2 \\
(B) &= \alpha_{z_i} + \alpha_1 x_{i,1} + \alpha_2 x_{i,2} + \alpha_3 x_{i,3} + \alpha_4 x_{i,4} + \alpha_5 x_{i,5} + \alpha_6 x_{i,6} \\
&\quad \alpha_{z_i} = (-0.1, 0.6, 0.3),\ \alpha_1 = \alpha_2 = \alpha_3 = 0.6,\ \alpha_4 = \alpha_5 = \alpha_6 = 0.4
\end{aligned}
$$

또한 간결한 표현을 위해 로그오즈에서 공변량 조건부 부분을 생략하였다. 즉, $\log\left[\frac{Pr(Z_i = j|\mathbf{X}_i = \mathbf{x}_i)}{Pr(Z_i = 3|\mathbf{X}_i = \mathbf{x}_i)}\right]$과 $\log\left[\frac{Pr(Y_i = 1|\mathbf{X}_i = \mathbf{x}_i, Z_i = z_i)}{1 - Pr(Y_i = 1|\mathbf{X}_i = \mathbf{x}_i, Z_i = z_i)}\right]$을

$\log\left[\frac{Pr(Z_i=j)}{Pr(Z_i=3)}\right]$과 $\log\left[\frac{Pr(Y_i=1)}{1-Pr(Y_i=1)}\right]$으로 표현하였다.

(M1) 선형 처리 모형과 선형 결과 모형

$$\log\left[\frac{Pr(Z_i=j)}{Pr(Z_i=3)}\right] = \theta_j + (A), \quad j=1,2,$$
$$\log\left[\frac{Pr(Y_i=1)}{1-Pr(Y_i=1)}\right] = \alpha_0 + (B),$$
$$(\theta_1,\theta_2) = (0.25, 0.3), \quad \alpha_0 = -0.2$$

(M2) 비선형 처리 모형과 선형 결과 모형

$$\log\left[\frac{Pr(Z_i=j)}{Pr(Z_i=3)}\right] = \theta_j + (A) + \beta_{1,j}(x_{i,1}+0.5)^2, \quad j=1,2,$$
$$\log\left[\frac{Pr(Y_i=1)}{1-Pr(Y_i=1)}\right] = \alpha_0 + (B),$$
$$(\theta_1,\theta_2) = (-0.5, 0), \quad \alpha_0 = -0.2$$

(M3) 선형 처리 모형과 비선형 결과 모형

$$\log\left[\frac{Pr(Z_i=j)}{Pr(Z_i=3)}\right] = \theta_j + (A), \quad j=1,2,$$
$$\log\left[\frac{Pr(Y_i=1)}{1-Pr(Y_i=1)}\right] = \alpha_0 + (B) + 0.5(x_{i,1}+0.5)^2,$$
$$(\theta_1,\theta_2) = (0.25, 0.3), \quad \alpha_0 = -0.8$$

(M4) 비선형 처리 모형과 비선형 결과 모형

$$\log\left[\frac{Pr(Z_i=j)}{Pr(Z_i=3)}\right] = \theta_j + (A) + \beta_{1,j}(x_{i,1}+0.5)^2, \quad j=1,2,$$
$$\log\left[\frac{Pr(Y_i=1)}{1-Pr(Y_i=1)}\right] = \alpha_0 + (B) + 0.5(x_{i,1}+0.5)^2,$$
$$(\theta_1,\theta_2) = (-0.5, 0), \quad \alpha_0 = -0.8$$

모의실험을 위하여 위의 각 모형 하에서의 평균처리효과(ATE)의 참값은

$$ATE_{2,1} = 0.7,\ ATE_{3,1} = 0.4,\ ATE_{3,2} = -0.3$$

으로 모두 동일하도록 α값들을 설정하였다.

모형설계 및 분석은 R 4.1.1을 사용했으며 일반화 부스팅 모형의 일반화 성향점수 추정을 위해서 `twang` (Ridgeway et al. [2017]) 패키지를, 다항 로지스틱 회귀 모형의 일반화 성향점수와 평균처리효과 추정을 위해서는 `PSweight` (Zhou et al. [2020a]) 패키지를 사용하였다.

제 2 절 평가방법

본 연구에서는 추정된 일반화 성향점수의 가중치 방법을 이용하여 공변량의 균형 보정 및 평균처리효과(ATE)를 추정하여 평가지표인 편향(bias), 평

균제곱오차(mean squared error; MSE), 포함확률(coverage probability; CP)과 표준화평균차이(standardized mean difference; SMD)로 비교하였다.

편향은 추정된 처리효과와 참값 간의 차이이며, 평균제곱오차는 편향의 제곱으로 정확도를, 포함확률은 추정된 처리효과의 신뢰구간이 참값을 포함하게 되는 비율로 안정성을 평가하기 위한 지표이다. 편향의 절댓값과 평균제곱오차는 작고, 포함확률은 신뢰도에 근접할수록 좋은 모형이라 할 수 있다. 본 연구에서는 포함확률의 신뢰도를 95%로 하였다.

$$\text{편향(bias)} = \frac{1}{500}\sum_{t=1}^{500}(\widehat{ATE}_{jj't} - ATE_{jj'})$$

$$\text{평균제곱오차(MSE)} = \frac{1}{500}\sum_{t=1}^{500}(\widehat{ATE}_{jj't} - ATE_{jj'})^2$$

$$\text{포함확률(CP)} = \frac{1}{500}\sum_{t=1}^{500} I(|\widehat{ATE}_{jj't} - ATE_{jj'}| \leq Z_{\alpha/2} SE(\widehat{ATE}_{jj't}))$$

$$j, j' \in \{1, 2, 3\}, \quad j \neq j'$$

표준화평균차이(SMD)는 공변량이 잘 균형화 되었는지 평가하기 위한 지표로 비교집단의 평균차이의 절댓값을 표준편차로 나눈 값이다. 표준화평균차이가 0.2 미만이면 균형화가 잘 된 것으로 보며, 0.4 는 보통, 0.6 이상은 차이가

큰 것으로 판단한다 (Cohen [2013]).

$$SMD_{j,j'} = \frac{\left|\bar{X}_j - \bar{X}_{j'}\right|}{\sqrt{\frac{s_j^2 + s_{j'}^2}{2}}}, \quad j, j' \in \{1, 2, 3\}, \quad j \neq j'$$

제 3 절 모의실험 분석 결과

다중 처리군의 기본형인 3가지의 처리군 $Z_i = \{1, 2, 3\}$, 2가지의 처리결과 $Y_i = \{0, 1; 0 = \text{효과 없음}, 1 = \text{효과 있음}\}$와 9개의 공변량($X$)을 가지는 모형으로 표본의 크기(N)가 200, 500과 1000인 표본을 추출하여 네 가지의 모형 가정 하에서 모의실험을 각각 500번 반복하였다. 다항 로지스틱 회귀 모형(MLR)과 일반화 부스팅 모형(GBM)으로 일반화 성향점수를 추정 한 후, 역확률 가중치(IPW), 중복 가중치(OW)와 매칭 가중치(MW) 방법을 사용하여 공변량 균형을 보정하고 처리효과를 추정하였다. 즉 $2 \times 3 = 6$개의 성향점수 방법의 성능을 평가지표를 통해 비교하였다.

다항 로지스틱 회귀 모형 또는 일반화 부스팅 모형으로 일반화 성향점수 추정시 9개의 공변량을 모두 적용하여 추정하였다. 일반화 부스팅 모형으로 일반화 성향점수 추정시 반복횟수(n.tree)는 10,000으로 하였고, 변수간의 균형을 측정하는 방법(stop.method)은 es.mean(standard effect mean)으로 지정하였다.

평균처리효과(ATE), 편향(Bias), 평균제곱오차(MSE)와 포함확률(CP)으로 표본의 크기와 모형 가정 하에서 일반화 성향점수 추정 및 가중치 방법들을 비교하였으며, 세 개의 처리군을 각각 짝을 지어 추정량을 비교하였다. 각각 짝을 지은 처리군간의 결과는 비슷하여 본문에서는 처리군 1과 3의 추정량을 다루었으며, 처리군 1과 2, 2와 3의 추정량은 부록에 정리하였다. 앞서 언급한 바와 같이, 처리군 1과 3의 평균처리효과(ATE)의 참값은 0.4이다. 공변량을 고려하지 않고 추정한 방법(unadjusted; Unadj)과 비교하여 6개의 성향점수 방법 모두 편향과 평균제곱오차는 작았으며, 포함확률은 0.95에 근접하게 나타났다.

평균처리효과의 분포는 그림 3.1에 나타내었으며, 참값과 추정량의 차이인 편향은 표 3.2에 정리하였다. M1, M2, M3에서는 6개의 성향점수 방법 모두 비슷하였으며, 그 중 편향은 다항 로지스틱 회귀 방법의 매칭 가중치 가장 작았다. 반면에 M4인 경우에는 일반화 부스팅 방법이 다항 로지스틱 회귀 방법보다 모든 편향이 작게 나타났으며, 평균처리효과의 분포도 0.4에 근접하게 나타났다. 표본의 크기가 작을수록 평균처리효과의 분포가 커지는 것을 볼 수 있었다. 가중치 방법은 실험 설계의 모든 조건에서 매칭 가중치, 중복 가중치, 역확률 가중치 순으로 편향이 작게 나타났다. 다항 로지스틱 회귀 방법의 역확률 가중치는 M4일 때와 표본의 크기가 200일 때 참값에서 크게 벗어난 이상치들이 존재하였다.

모형의 정확도를 평가하기 위한 평균제곱오차는 표 3.3에 정리하였다. M1,

M2, M3에서는 표본의 크기가 500 이상인 경우에는 6개의 성향점수 방법 모두 비슷한 결과를 보였으나, 표본의 크기가 200 인 경우에는 일반화 부스팅 방법의 역확률 가중치가 각 0.155(M1), 0.177(M2), 0.168(M3)로 6개의 성향점수 방법 중 가장 작았다. M4의 경우에는 일반화 부스팅 방법이 모두 0.3 이하로 다항 로지스틱 회귀 방법보다 작게 나타났으며, 가중치 방법 중에서는 매칭 가중치가 가장 작았다.

모형의 안정성을 평가하기 위한 포함확률은 표 3.4에 정리하였다. 포함확률은 모든 조건에서 표본의 크기가 500일 때 0.95에 가장 근접 하였으며, 1000일 때 가장 작았다. M1, M2, M3에서는 6개의 성향점수 방법 모두 비슷한 결과를 보였으며, 그 중 다항 로지스틱 회귀 방법의 매칭 가중치가 0.95에 가장 근접하였다. M4인 경우에는 일반화 부스팅 방법이 다항 로지스틱 회귀 방법보다 0.95에 근접하게 나타났으며, 일반화 부스팅 방법에 비해 다항 로지스틱 회귀 방법은 표본의 크기가 클수록 포함 확률이 현저히 줄어들었다. 가중치 방법 중에서는 대부분 매칭 가중치, 중복 가중치, 역확률 가중치 순으로 0.95에 근접하게 나타났다. 다른 성향점수 방법들에 비해 일반화 부스팅 방법의 중복 가중치와 매칭가중치는 표본의 크기와 모형 따른 포함확률의 변동이 작았다.

공변량의 균형을 평가를 위한 표준화평균차이의 평균 분포는 그림 3.2에 나타내었다. 성향점수 추정 및 가중치 방법 모두 공변량을 고려하지 않고 추정한 방법에 비해 균형화가 잘 되었음을 볼 수 있었다. 일반화 부스팅 방법이 다항 로지스틱 회귀 방법보다 평균이 크게 나타났지만, 중복 가중치와 매칭

가중치의 평균은 모두 0.2보다 작아 공변량의 균형화를 만족하였다. 반면에 일반화 부스팅 방법의 역확률 가중치는 표본의 크기가 500, 200 일 때 평균이 0.2보다 크게 나타났으나, 0.4 보다는 작게 나타나 균형이 나쁘지는 않았다. 다항 로지스틱 회귀 방법의 역확률 가중치의 평균은 0.2보다 작게 나타났으나, 이상치가 1과 근접하게 분포하였다. 역확률 가중치의 경우 표본의 크기가 작거나 처리 모형이 비선형 일 때, 균형화의 변동이 매우 커진 것을 볼 수 있다.

표 3.2: The bias for each method under each scenario between treatments 1 and 3.

	N	Unadj	MLR			GBM		
			IPW	OW	MW	IPW	OW	MW
M1	200	-1.267	-0.139	-0.100	-0.086	-0.143	-0.125	-0.116
	500	-1.278	-0.141	-0.109	-0.096	-0.144	-0.124	-0.113
	1,000	-1.282	-0.143	-0.118	-0.110	-0.143	-0.125	-0.116
M2	200	-1.413	-0.142	-0.085	-0.070	-0.137	-0.119	-0.109
	500	-1.407	-0.136	-0.095	-0.083	-0.135	-0.118	-0.110
	1,000	-1.421	-0.142	-0.117	-0.109	-0.146	-0.125	-0.118
M3	200	-1.394	-0.171	-0.127	-0.110	-0.192	-0.170	-0.156
	500	-1.415	-0.178	-0.141	-0.127	-0.188	-0.163	-0.150
	1,000	-1.403	-0.163	-0.138	-0.130	-0.172	-0.150	-0.140
M4	200	-1.767	-0.436	-0.323	-0.283	-0.328	-0.270	-0.245
	500	-1.776	-0.543	-0.354	-0.315	-0.309	-0.241	-0.218
	1,000	-1.761	-0.551	-0.337	-0.298	-0.267	-0.193	-0.168

M1: Linear $E[Z|X]$, Linear $E[Y|X]$, M2: Non-Linear $E[Z|X]$, Linear $E[Y|X]$, M3: Linear $E[Z|X]$, Non-Linear $E[Y|X]$, M4: Non-Linear $E[Z|X]$, Non-Linear $E[Y|X]$; N: sample size; Unadj: Unadjusted, IPW: Inverse Probability of treatment Weights, OW: Overlap Weights, MW: Matching Weights

표 3.3: The MSE for each method under each scenario between treatments 1 and 3.

	N	Unadj	MLR			GBM		
			IPW	OW	MW	IPW	OW	MW
M1	200	1.723	0.182	0.169	0.177	0.155	0.163	0.168
	500	1.684	0.079	0.065	0.066	0.067	0.067	0.068
	1,000	1.672	0.048	0.041	0.043	0.045	0.043	0.043
M2	200	2.128	0.262	0.202	0.214	0.177	0.197	0.214
	500	2.026	0.099	0.063	0.066	0.064	0.068	0.073
	1,000	2.046	0.054	0.046	0.048	0.048	0.047	0.049
M3	200	2.081	0.195	0.173	0.177	0.168	0.173	0.180
	500	2.060	0.086	0.073	0.075	0.080	0.077	0.076
	1,000	1.994	0.050	0.045	0.045	0.051	0.047	0.046
M4	200	3.262	0.549	0.303	0.286	0.267	0.248	0.249
	500	3.205	0.529	0.186	0.164	0.146	0.117	0.113
	1,000	3.128	0.540	0.146	0.122	0.096	0.066	0.060

표 3.4: The CP for each method under each scenario between treatments 1 and 3.

	N	Unadj	MLR			GBM		
			IPW	OW	MW	IPW	OW	MW
M1	200	0.044	0.880	0.896	0.928	0.888	0.890	0.896
	500	0.000	0.898	0.920	0.940	0.906	0.922	0.928
	1,000	0.000	0.864	0.874	0.878	0.846	0.872	0.864
M2	200	0.022	0.878	0.904	0.936	0.890	0.902	0.896
	500	0.000	0.932	0.946	0.950	0.924	0.932	0.934
	1,000	0.000	0.884	0.894	0.916	0.858	0.902	0.912
M3	200	0.028	0.872	0.894	0.906	0.888	0.898	0.902
	500	0.000	0.872	0.896	0.916	0.860	0.882	0.906
	1,000	0.000	0.826	0.860	0.874	0.786	0.846	0.872
M4	200	0.000	0.796	0.852	0.896	0.834	0.876	0.880
	500	0.000	0.666	0.726	0.802	0.754	0.846	0.872
	1,000	0.000	0.440	0.534	0.638	0.622	0.804	0.864

그림 3.1: Distribution of the ATE for each method under each scenario between treatments 1 and 3. The true ATE value of 0.4 is included as the dotted line.

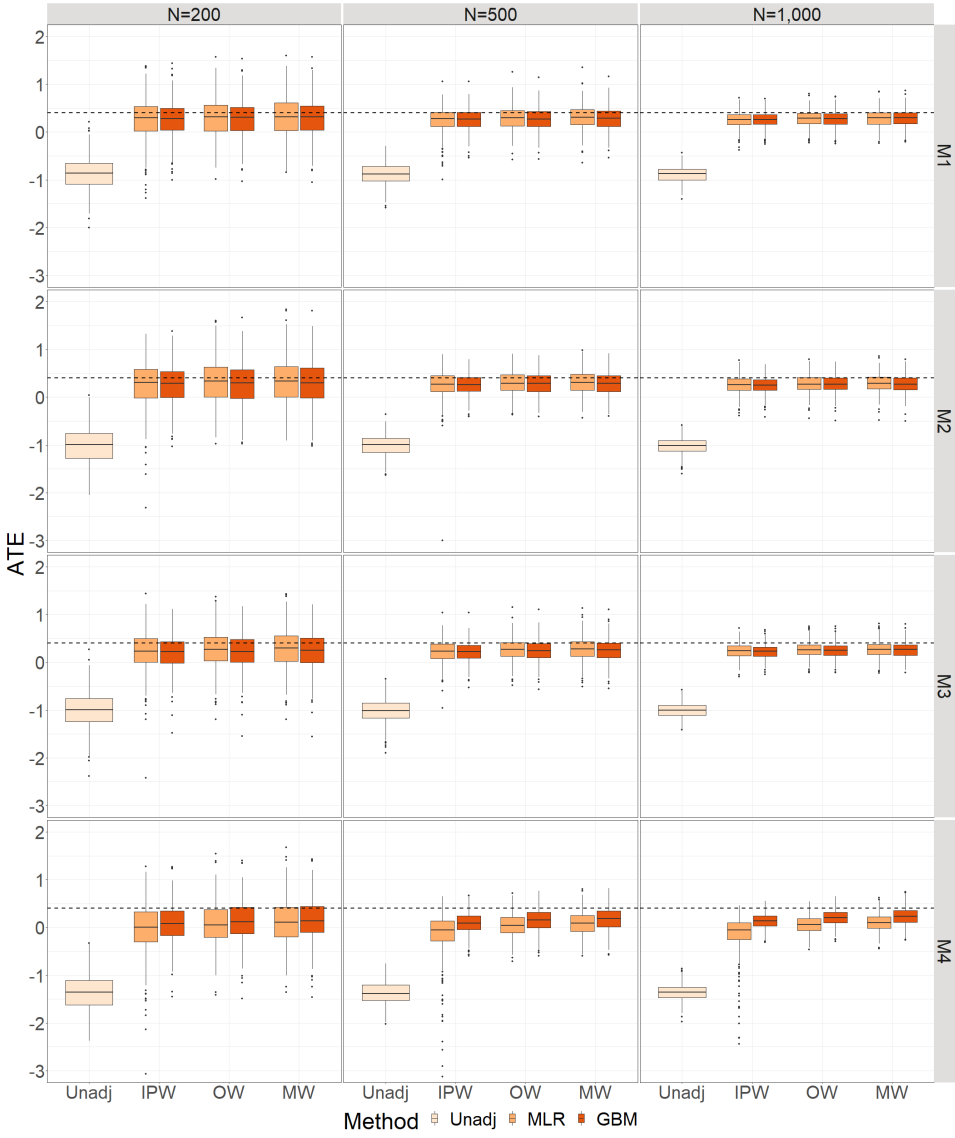
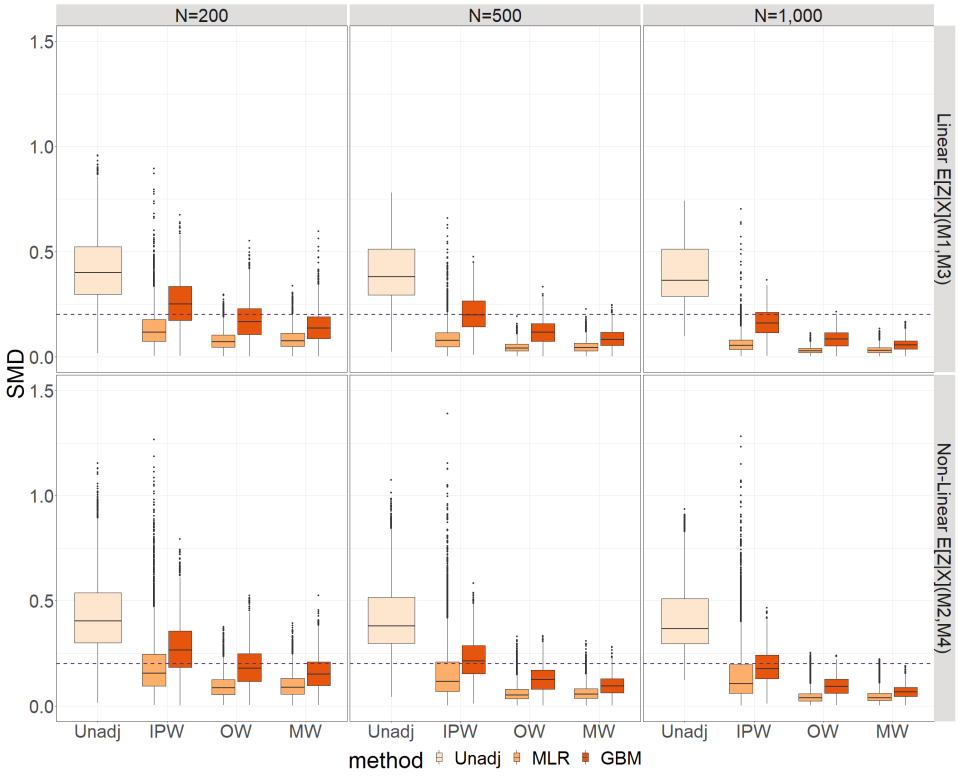


그림 3.2: Graphical representation of the covariate balance achieved by each method. SMD was calculated for all baseline covariates within each treatment pair.



제 4 장 결론

본 논문에서는 다중 처리군의 경우 표본의 크기와 네 가지 모형의 가정에서 일반화 성향점수 추정 방법인 다항 로지스틱 회귀 모형과 일반화 부스팅 모형, 가중치 방법인 역확률 가중치, 중복 가중치와 매칭 가중치를 모의실험을 통해 비교하여 보았다.

일반화 성향점수 추정에 가장 많이 사용되는 다항 로지스틱 회귀 방법은 처리 또는 결과 모형이 선형(M1,M2,M3)인 경우에는 편향과 평균제곱오차는 작고 포함확률은 0.95에 근접하였으나, 처리 및 결과 모형 모두 비선형(M4)인 경우에는 편향과 평균제곱오차는 커지고 포함확률은 크게 줄어들었다. 또한 처리 또는 결과 모형이 선형인 경우와 모두 비선형인 경우의 추정값은 큰 차이를 보였다. 반면에 비모수 방법인 일반화 부스팅 방법은 처리 또는 결과 모형의 선형, 비선형과 관계없이 편향, 평균제곱오차, 포함확률에서 큰 차이가 없었다. 특히, 비선형인 경우와 표본의 크기가 달라진 경우에도 다항 로지스틱 회귀 방법에 비해서 편향과 평균제곱오차는 작고 포함확률은 0.95에 근접하였다. 표준화평균차이에서는 일반화 부스팅 방법이 다항 로지스틱 회귀 방법보다는 크게 나타났으나, 0.2와 근접하여 공변량의 균형화를 만족하였다.

성향점수의 가중치 방법의 측면에서는 매칭 가중치, 중복 가중치, 역확률

가중치의 순으로 좋은 성능을 보였다. 특히, 모든 가정에서의 편향 및 포함확률, 처리 및 결과 모형 모두 비선형인 가정(M4)에서의 평균제곱오차와 일반화 부스팅 방법을 사용할 때의 표준화평균차이에서는 매칭 가중치가 가장 좋은 성능을 보였다. 가중치 방법 중 가장 널리 사용되고 있는 역확률 가중치는 편향과 평균제곱오차는 다항 로지스틱 회귀 방법보다 일반화 부스팅 방법을 사용할 경우 더 작았으며, 모형이 비선형이거나 표본이 작을 경우에는 다른 가중치 방법에 비해 평균처리효과의 변동이 크고 편향, 평균제곱오차와 포함확률의 결과가 좋지 않았다. 또한 역확률 가중치의 표준화평균차이는 일반화 부스팅 방법은 0.2보다 컸으며, 다항 로지스틱 회귀 방법은 처리 모형이 비선형(M2,M4)인 경우에 공변량을 고려하지 않고 추정한 방법보다 더 크고 1에 근접한 이상치들이 존재하였다.

성향점수 추정 및 가중치 방법들의 소표본 성질을 모의실험을 통하여 비교해 보았다. 처리 및 결과 모형으로 로그-오즈 선형 모형과 비선형 모형을 모두 고려해 보면, 다항 로지스틱 회귀 방법에 비해 일반화 부스팅 방법의 성능이 더 좋은 결과를 보였으며, 가중치 방법에서는 매칭 가중치 방법이 가장 효율적인 결과를 보였다. 본 연구에서 사용한 McCaffrey et al. [2013]의 일반화 부스팅 모형의 일반화 성향점수 추정법은 반복절차를 통해 최적의 성향점수 모형으로 추정하지만, 기준을 역확률 가중치 적용하였다. 기준을 매칭 가중치, 중복 가중치 등을 적용한다면, 공변량의 균형 및 처리효과의 추정이 더 좋을 것으로 보인다.

실제 관측 자료는 정확한 분포를 파악하기 어려우며, 처리수준, 처리할당 비율, 표본 수, 변수들 간의 관계 및 중복 등의 요소들을 고려해야 한다. 그러므로 다양한 가정을 반영하였을 때의 성향점수 추정 및 적용 방법들을 비교해 볼 필요도 있다. 또한 어떠한 환경에서도 공변량의 균형화를 만족하고 처리효과를 정확하게 추정 할 수 있도록 효율적인 성향점수 추정 및 적용 방법에 대해 추가적인 연구가 필요하다.

참고 문헌

Adelson, J. L., McCoach, D., Rogers, H., Adelson, J. A., and Sauer, T. M. (2017). Developing and applying the propensity score to make causal inferences: variable selection and stratification. *Frontiers in psychology*, 8:1413.

Brown, D. W., DeSantis, S. M., Greene, T. J., Maroufy, V., Yaseen, A., Wu, H., Williams, G., and Swartz, M. D. (2020). A novel approach for propensity score matching and stratification for multiple treatments: Application to an electronic health record–derived study. *Statistics in medicine*, 39(17):2308–2323.

Bühlmann, P. and Yu, B. (2003). Boosting with the l 2 loss: regression and classification. *Journal of the American Statistical Association*, 98(462):324–339.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Academic press.

Ding, P. and Li, F. (2018). Causal inference: A missing data perspective. *Statistical Science*, 33(2):214–237.

Feng, P., Zhou, X.-H., Zou, Q.-M., Fan, M.-Y., and Li, X.-S. (2012). Generalized propensity score for estimating the average treatment effect of multiple treatments. *Statistics in medicine*, 31(7):681–697.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139.

Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2):337–407.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Hirano, K. and 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):259–278.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling

without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710.

King, G. and Nielsen, R. (2019). Why propensity scores should not be used for matching. *Political Analysis*, 27(4):435–454.

Li, F. (2019). Propensity score weighting for causal inference with multiple treatments. *The Annals of Applied Statistics*, 13(4):2389–2415.

Li, F., Morgan, K. L., and Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400.

Li, L. and Greene, T. (2013). A weighting analogue to pair matching in propensity score analysis. *The international journal of biostatistics*, 9(2):215–234.

Madigan, D. and Ridgeway, G. (2004). Discussion of" least angle regression" by efron et al. *arXiv preprint math/0406469*.

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., and Burgette, L. F. (2013). A tutorial on propensity score estimation

for multiple treatments using generalized boosted models. *Statistics in medicine*, 32(19):3388–3414.

McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, 9(4):403.

Ridgeway, G. (1999). The state of boosting. *Computing science and statistics*, pages 172–181.

Ridgeway, G., McCaffrey, D., Morral, A., Burgette, L., and Griffin, B. A. (2017). Toolkit for weighting and analysis of nonequivalent groups: A tutorial for the twang package. *Santa Monica, CA: RAND Corporation.*

Robins, J. M., Hernan, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, 82(398):387–394.

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

Rubin, D. B. (1997). Estimation from nonrandomized treatment comparisons using subclassification on propensity scores.

Spreeuwenberg, M. D., Bartak, A., Croon, M. A., Hagenaars, J. A., Busschbach, J. J., Andrea, H., Twisk, J., and Stijnen, T. (2010). The multiple propensity score as control for bias in the comparison of more than two treatment arms: an introduction from a case study in mental health. *Medical care*, pages 166–174.

Yoshida, K., Hernández-Díaz, S., Solomon, D. H., Jackson, J. W., Gagne, J. J., Glynn, R. J., and Franklin, J. M. (2017). Matching weights to simultaneously compare three treatment groups: comparison to three-way matching. *Epidemiology (Cambridge, Mass.)*, 28(3):387.

Zhou, T., Tong, G., Li, F., and Thomas, L. E. (2020a). Psweight: An r package for propensity score weighting analysis. *arXiv preprint arXiv:2010.08893*.

Zhou, Y., Matsouaka, R. A., and Thomas, L. (2020b). Propensity score weighting under limited overlap and model misspecification. *Statistical Methods in Medical Research*, 29(12):3721–3756.

부록

표 A.1: The bias for each method under each scenario between treatments 1 and 2.

	N	Unadj	MLR			GBM		
			IPW	OW	MW	IPW	OW	MW
M1	200	-0.399	-0.244	-0.192	-0.174	-0.241	-0.218	-0.205
	500	-0.406	-0.238	-0.191	-0.178	-0.239	-0.212	-0.200
	1000	-0.421	-0.240	-0.199	-0.186	-0.240	-0.211	-0.197
M2	200	-0.562	-0.235	-0.171	-0.159	-0.237	-0.211	-0.196
	500	-0.574	-0.237	-0.178	-0.162	-0.234	-0.201	-0.189
	1000	-0.589	-0.241	-0.197	-0.184	-0.245	-0.211	-0.198
M3	200	-0.515	-0.289	-0.230	-0.211	-0.280	-0.257	-0.245
	500	-0.517	-0.276	-0.226	-0.212	-0.268	-0.239	-0.228
	1000	-0.533	-0.276	-0.235	-0.220	-0.270	-0.240	-0.228
M4	200	-0.834	-0.446	-0.331	-0.297	-0.339	-0.295	-0.280
	500	-0.836	-0.525	-0.335	-0.296	-0.312	-0.259	-0.245
	1000	-0.838	-0.549	-0.339	-0.299	-0.302	-0.242	-0.225

M1: Linear $E[Z|X]$, Linear $E[Y|X]$, M2: Non-Linear $E[Z|X]$, Linear $E[Y|X]$, M3: Linear $E[Z|X]$, Non-Linear $E[Y|X]$, M4: Non-Linear $E[Z|X]$, Non-Linear $E[Y|X]$; N: sample size; Unadj: Unadjusted, IPW: Inverse Probability of treatment Weights, OW: Overlap Weights, MW: Matching Weights

표 A.2: The MSE for each method under each scenario between treatments 1 and 2.

	N	Unadj	MLR			GBM		
			IPW	OW	MW	IPW	OW	MW
M1	200	0.298	0.169	0.175	0.188	0.163	0.174	0.188
	500	0.216	0.101	0.089	0.092	0.101	0.095	0.098
	1000	0.201	0.077	0.063	0.063	0.077	0.067	0.066
M2	200	0.448	0.208	0.181	0.201	0.169	0.181	0.196
	500	0.375	0.126	0.083	0.087	0.097	0.091	0.093
	1000	0.371	0.081	0.062	0.061	0.079	0.069	0.068
M3	200	0.408	0.210	0.186	0.193	0.178	0.183	0.193
	500	0.321	0.125	0.105	0.106	0.113	0.105	0.108
	1000	0.307	0.095	0.078	0.075	0.090	0.079	0.077
M4	200	0.831	0.486	0.271	0.269	0.226	0.222	0.237
	500	0.744	0.498	0.164	0.145	0.135	0.114	0.114
	1000	0.730	0.531	0.143	0.119	0.112	0.085	0.082

표 A.3: The CP for each method under each scenario between treatments 1 and 2.

	N	Unadj	MLR			GBM		
			IPW	OW	MW	IPW	OW	MW
M1	200	0.774	0.828	0.886	0.916	0.834	0.854	0.866
	500	0.558	0.746	0.820	0.858	0.752	0.804	0.832
	1,000	0.238	0.582	0.740	0.806	0.554	0.684	0.746
M2	200	0.654	0.846	0.884	0.906	0.846	0.872	0.884
	500	0.272	0.810	0.866	0.894	0.786	0.846	0.872
	1,000	0.036	0.642	0.772	0.830	0.572	0.734	0.792
M3	200	0.682	0.772	0.852	0.900	0.790	0.822	0.840
	500	0.336	0.636	0.798	0.838	0.658	0.750	0.796
	1,000	0.074	0.450	0.638	0.700	0.450	0.604	0.682
M4	200	0.366	0.714	0.794	0.856	0.766	0.800	0.818
	500	0.038	0.580	0.692	0.786	0.634	0.794	0.828
	1,000	0.004	0.336	0.462	0.596	0.412	0.650	0.712

표 A.4: The bias for each method under each scenario between treatments 2 and 3.

	N	Unadj	MLR			GBM		
			IPW	OW	MW	IPW	OW	MW
M1	200	-0.869	0.104	0.092	0.088	0.098	0.093	0.089
	500	-0.872	0.097	0.082	0.082	0.095	0.087	0.086
	1,000	-0.862	0.097	0.081	0.076	0.098	0.086	0.082
M2	200	-0.851	0.093	0.086	0.088	0.100	0.092	0.087
	500	-0.833	0.101	0.082	0.079	0.099	0.084	0.078
	1,000	-0.831	0.099	0.080	0.075	0.099	0.086	0.080
M3	200	-0.879	0.118	0.103	0.101	0.089	0.087	0.089
	500	-0.898	0.099	0.085	0.085	0.080	0.076	0.078
	1,000	-0.870	0.113	0.096	0.090	0.098	0.090	0.088
M4	200	-0.932	0.010	0.008	0.014	0.012	0.025	0.035
	500	-0.940	-0.018	-0.019	-0.019	0.003	0.018	0.027
	1,000	-0.923	-0.002	0.002	0.001	0.035	0.049	0.057

표 A.5: The MSE for each method under each scenario between treatments 2 and 3.

	N	Unadj	MLR			GBM		
			IPW	OW	MW	IPW	OW	MW
M1	200	0.886	0.180	0.168	0.178	0.153	0.164	0.175
	500	0.810	0.063	0.054	0.058	0.052	0.055	0.059
	1,000	0.766	0.033	0.030	0.032	0.031	0.031	0.033
M2	200	0.852	0.183	0.159	0.174	0.136	0.147	0.158
	500	0.741	0.064	0.064	0.069	0.058	0.060	0.064
	1,000	0.717	0.034	0.033	0.035	0.032	0.032	0.034
M3	200	0.910	0.153	0.148	0.155	0.128	0.139	0.148
	500	0.859	0.065	0.059	0.063	0.053	0.056	0.062
	1,000	0.781	0.034	0.031	0.033	0.029	0.030	0.032
M4	200	1.002	0.143	0.155	0.169	0.131	0.137	0.148
	500	0.934	0.054	0.054	0.059	0.047	0.053	0.059
	1,000	0.878	0.024	0.026	0.028	0.024	0.028	0.031

표 A.6: The CP for each method under each scenario between treatments 2 and 3.

	N	Unadj	MLR			GBM		
			IPW	OW	MW	IPW	OW	MW
M1	200	0.330	0.854	0.880	0.908	0.882	0.882	0.888
	500	0.022	0.882	0.944	0.948	0.918	0.934	0.938
	1,000	0.000	0.898	0.916	0.930	0.900	0.908	0.916
M2	200	0.344	0.870	0.902	0.932	0.890	0.898	0.900
	500	0.042	0.898	0.918	0.924	0.908	0.918	0.920
	1,000	0.000	0.866	0.920	0.932	0.882	0.910	0.932
M3	200	0.340	0.876	0.914	0.948	0.906	0.910	0.904
	500	0.018	0.888	0.928	0.932	0.920	0.932	0.930
	1,000	0.000	0.866	0.918	0.932	0.898	0.920	0.928
M4	200	0.274	0.884	0.922	0.934	0.910	0.924	0.924
	500	0.014	0.922	0.940	0.948	0.936	0.934	0.916
	1,000	0.000	0.948	0.968	0.966	0.946	0.938	0.942

그림 A.1: Distribution of the ATE for each method under each scenario between treatments 1 and 2. The true ATE value of 0.7 is included as the dotted line.

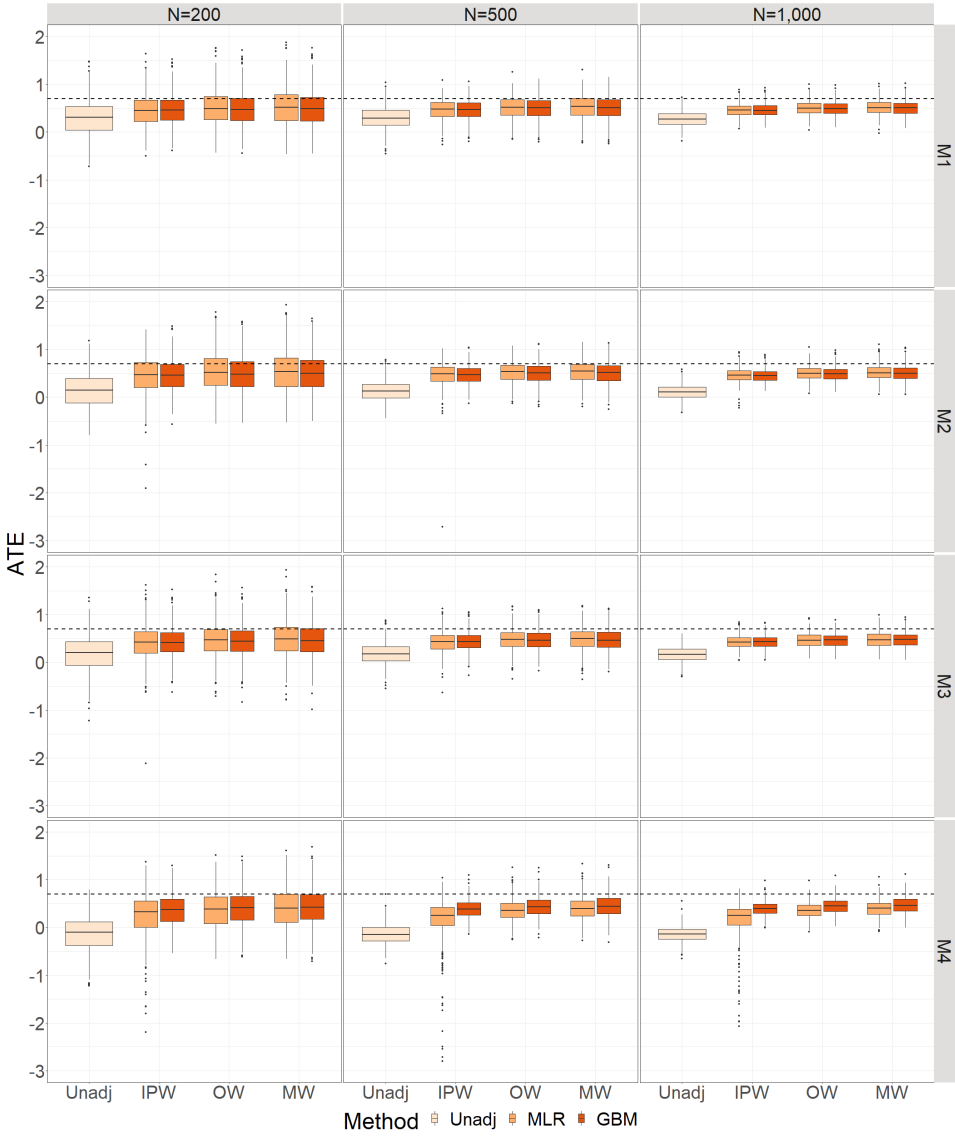
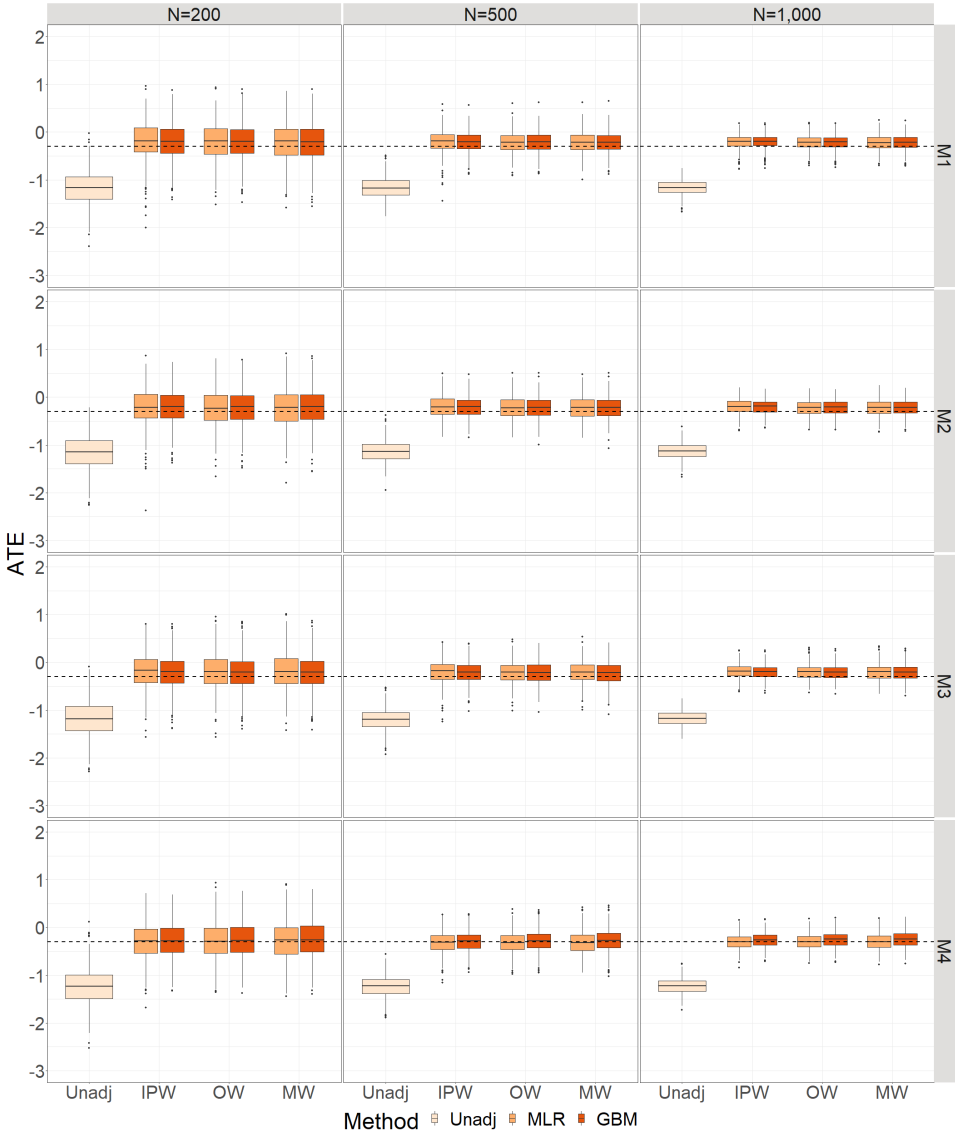


그림 A.2: Distribution of the ATE for each method under each scenario between treatments 2 and 3. The true ATE value of -0.3 is included as the dotted line.



국문초록

다중 처리군에서 성향점수 가중치 방법의 비교

관찰연구에서 처리효과 추정치는 공변량 불균형의 영향으로 인한 편향이 발생한다. 역확률 가중치, 중복 가중치 및 매칭 가중치와 같은 성향점수의 가중치 방법은 추정치의 편향을 줄일 수 있다. 다중 처리군의 경우에는 성향점수를 확장한 일반화 성향점수 방법이 제안되었다. 주로 일반화 성향점수 추정시 다항 로지스틱 회귀 모형이 사용되며, 가중치 방법에도 적용되었다. 그러나 표본 크기에 비해 공변량이 너무 많거나 정규성을 가정할 수 없는 경우, 모수적 방법인 다항 로지스틱 회귀의 치료효과 추정치는 편향이 발생 할 수 있다. 반면 일반화 부스팅 모형은 비모수적 방법과 같은 문제에 적합하다. 또한 대용량 데이터에서도 잘 구현되어 최근에는 일반화 성향점수 추정에도 사용되고 있다.

본 연구는 모의실험을 통해 다항 로지스틱 회귀 모형과 일반화 부스팅 모형에 의해 추정된 일반화 성향점수와 이를 기반으로 하는 역확률 가중치, 중복 가중치와 매칭 가중치의 소표본 속성을 비교하였다. 모의실험 결과, 처리 또는 결과 모형이 선형인 경우에만 다항 로지스틱 회귀 방법은 편향과 MSE가 작았으며 CP는 높게 나타났다. 그러나 일반화 부스팅 방법은 모형의 선형 여부에 관계없이 편향, MSE와 CP는 일관된 결과를 보였으며, 처리 또는 결과 모형이 비선형인 경우에는 다항 로지스틱 회귀 방법보다 우수한 성능을 보였다. 가중

치 방법은 전반적으로 매칭 가중치, 중복 가중치, 역확률 가중치 순으로 우수한 성능 보였다. 매칭 가중치와 중복 가중치는 일반화 성향점수 추정 방법에 따른 결과와 비슷하게 나타났으며, 역확률 가중치는 일반화 부스팅 모형으로 추정할 경우 성능이 더 좋게 나타났다.

주제어 : 다중 처리군, 일반화 성향점수, 다항 로지스틱 회귀, 일반화 부스팅 모형, 가중치 방법