The Korean Journal of Applied Statistics (2017) DOI: https://doi.org/10.5351/KJAS.2017.30.5.615
30(5), 615–632

Latent causal inference using the propensity score from latent class regression model

Misol Lee[a] · Hwan Chung[a,1]

[a]Department of Statistics, Korea University

(Received July 5, 2017; Revised August 22, 2017; Accepted August 23, 2017)

Abstract

Unlike randomized trial, statistical strategies for inferring the unbiased causal relationship are required in the observational studies. The matching with the propensity score is one of the most popular methods to control the confounders in order to evaluate the effect of the treatment on the outcome variable. Recently, new methods for the causal inference in latent class analysis (LCA) have been proposed to estimate the average causal effect (ACE) of the treatment on the latent discrete variable. They have focused on the application study for the real dataset to estimate the ACE in LCA. In practice, however, the true values of the ACE are not known, and it is difficult to evaluate the performance of the estimated the ACE. In this study, we propose a method to generate a synthetic data using the propensity score in the framework of LCA, where treatment and outcome variables are latent. We then propose a new method for estimating the ACE in LCA and evaluate its performance via simulation studies. Furthermore we present an empirical analysis based on data form the 'National Longitudinal Study of Adolescents Health,' where *puberty* as a latent treatment and *substance use* as a latent outcome variable.

Keywords: average causal effect, latent class analysis, observational study, propensity score

1. 서론

인과효과(causal inference)를 추정할 경우 무작위 통제시험(randomized controlled trial)과는 달리 관찰연구(observational study)에서는 교란변수에 의한 편향이 발생하여 이에 대한 통계적 전략이 요구된다. Rosenbaum과 Rubin (1983)은 교란변수에 의한 편향을 줄이기 위해 성향점수(propensity score)를 제안하였으며, 이러한 성향점수를 기반으로 역확률 가중치 방법과 성향점수 짝짓기 방법 등이 고안되었다 (Robins 등, 2000; Rosenbaum, 2002). 관찰연구에서 불편향 인과효과를 추정하기 위한 이러한 방법은 처치변수와 결과변수 모두 직접 관측이 가능한 관측변수(manifest variable)를 대상으로 한다. 따라서 능력, 태도, 가치와 같이 직접 관측되지 않는 잠재변수(latent variable)를 포함한 모형에서 기존의 방법을 사용하는데 제한이 있다. 이러한 한계를 극복하기 위해 최근 Lanza 등 (2013)은 결과변수가 잠재변수인 상황에서 인과효과를 추론하는 방법을 제안하였고, Park과 Chung (2014)은 처치변수

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2015R1D1A1A01056846 to Hwan Chung).
[1]Corresponding author: Department of Statistics, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul 02841, Korea. E-mail: hwanch@korea.ac.kr

와 결과변수가 모두 잠재변수인 상황에서 인과효과를 추론하는 방법을 제안하였다. 그러나 이러한 연구들은 예제를 기반으로 하여 참값을 알 수 없고, 따라서 그들이 제안한 방법의 성능을 파악할 수 없다는 단점을 가지고 있다고 할 수 있다.

본 연구에서는 Park과 Chung (2014)이 제안한 방법을 개선하여, 다항범주형 처치변수가 잠재변수인 상황에서 다항범주형 결과변수에 미치는 인과효과 추정방법을 제안하고 처치변수와 결과변수가 잠재변수 또는 관측변수를 포함하는 여러 상황에서 본 연구가 제안한 인과효과 추정방법의 성능을 모의실험연구를 통하여 평가하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 공변량을 고려한 잠재범주분석(latent class analysis; LCA)에 대하여 소개하고 다항범주형 처치변수가 잠재변수인 상황에서 평균인과효과를 추정하는 방법을 제시하고자 한다. 3장에서는 모의자료를 생성한 후 모의실험을 통해 제안된 방법의 성능을 보일 것이며, 4장에서는 'National Longitudinal Study of Adolescents Health' (Udry, 2003) 자료를 이용하여 미국 여성 청소년 성장과 약물사용 간의 인과효과를 추론하고자 한다. 마지막으로 5장에서는 결론과 향후 연구에 대하여 논의하고자 한다.

2. 잠재범주분석을 기반으로 한 상황별 인과효과 추론

2.1. 공변량을 고려한 잠재범주분석

LCA에서는 관측 가능한 여러 변수를 이용하여 능력, 태도, 가치와 같은 직접 관측하기 어려운 변수를 측정하고 문항들의 응답 패턴을 통해 개체들을 몇 개의 잠재범주(latent class)로 분류하는 통계적 기법이다 (Clogg와 Goodman, 1984; Goodman 1974). 잠재범주를 측정하는 M개의 문항변수 $\mathbf{Y} = (Y_1, \ldots, Y_M)$이 있다고 가정하자. M개의 문항변수에 대한 i번째 개체의 관측치 $\mathbf{y}_i = (y_{i1}, \ldots, y_{iM})$의 m번째 관측치 y_{im}은 r_m개의 응답범주를 가진다. 또한, 잠재범주변수 U는 L개의 잠재범주를 가지고 있다고 하자. 이때 LCA의 i번째 개체에 대한 우도함수는 다음과 같이 주어진다.

$$\begin{aligned} P(\mathbf{Y} = \mathbf{y}_i) &= \sum_{l=1}^{L} P(U = l) P(\mathbf{Y} = \mathbf{y}_i \mid U = l) \\ &= \sum_{l=1}^{L} P(U = l) \prod_{m=1}^{M} P(Y_m = y_{im} \mid U = l) \\ &= \sum_{l=1}^{L} \gamma_l \prod_{m=1}^{M} \prod_{k=1}^{r_m} \rho_{mk|l}^{I(y_{im}=k)}. \end{aligned} \tag{2.1}$$

식 (2.1)에서 $I(y_{im} = k)$는 i번째 개체가 관측변수 Y_m에 대해 k라고 응답한 경우는 1이고, 그 외의 경우는 0인 지시함수를 나타내며 LCA의 우도함수는 다음의 두 가지 모수로 구성된다. $\gamma_l = P(U = l)$은 개체가 잠재범주 l에 포함될 확률이며, $\rho_{mk|l} = P(Y_m = k \mid U = l)$는 어떠한 개체가 잠재범주 l에 속할 때 m번째 문항에 대해 k번째 범주에 응답할 확률이다. 식 (2.1)과 같이 LCA는 잠재범주가 주어진 경우 각 문항변수가 독립이라는 조건부 독립의 가정(local independence assumption)이 요구된다 (Lazarsfeld와 Henry, 1968).

사후확률(posterior probability)은 개체가 특정 관측값을 가졌을때 각 잠재범주에 속할 확률을 의미하며, 이를 통해 모든 개체의 잠재범주를 할당할 수 있다. 식 (2.1)에서 i번째 개체의 관측변수가 $\mathbf{y}_i$일때

Table 2.1. Data composed of binary treatment and outcome variables

개체 (i)	교란변수 (X_i)	T_i	$Y_i(0)$	$Y_i(1)$
1	x_1	0	$Y_1(0)$	$Y_1(1)$
2	x_2	0	$Y_2(0)$	$Y_2(1)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
m	x_m	0	$Y_m(0)$	$Y_m(1)$
$m+1$	x_{m+1}	1	$Y_{m+1}(0)$	$Y_{m+1}(1)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_n	1	$Y_n(0)$	$Y_n(1)$

잠재범주 l에 포함될 확률은 다음과 같이 나타낼 수 있다.

$$P(U = l \mid \mathbf{Y} = \mathbf{y}_i) = \frac{P(L = l)P(\mathbf{Y} = \mathbf{y}_i | L = l)}{P(\mathbf{Y} = \mathbf{y}_i)} = \frac{\gamma_l \prod_{m=1}^{M} \prod_{k=1}^{r_m} \rho_{mk|l}^{I(y_{im}=k)}}{\sum_{c=1}^{C} \gamma_c \prod_{m=1}^{M} \prod_{k=1}^{r_m} \rho_{mk|c}^{I(y_{im}=k)}}. \tag{2.2}$$

LCA에서 공변량(covariate)은 다항로지스틱 회귀분석의 형태로 특정 잠재범주 l에 속할 확률인 γ_l을 예측하는 데 사용된다. $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$를 i번째 개체의 공변량에 대한 관측치라고 하면 식 (2.1)에서 주어진 LCA의 우도함수는 다음과 같이 주어진다 (Dayton과 Macready, 1988).

$$P(\mathbf{Y} = \mathbf{y}_i) = \sum_{l=1}^{L} \gamma_l(\mathbf{x}_i) \prod_{m=1}^{M} \prod_{k=1}^{r_m} \rho_{mk|l}^{I(y_{im}=k)}. \tag{2.3}$$

식 (2.3)의 i번째 개체가 잠재범주 l에 포함될 확률은 다음과 같이 나타낼 수 있다.

$$\gamma_l(\mathbf{x}_i) = \frac{\exp\left(\mathbf{x}_i^T \boldsymbol{\beta}_l\right)}{1 + \sum_{c=1}^{L-1} \exp\left(\mathbf{x}_i^T \boldsymbol{\beta}_c\right)}. \tag{2.4}$$

또한, 공변량을 고려한 LCA의 사후확률은 다음과 같이 나타낼 수 있다.

$$P(U = l \mid \mathbf{Y} = \mathbf{y}_i, \mathbf{x}_i) = \frac{\gamma_l(\mathbf{x}_i) \prod_{m=1}^{M} \prod_{k=1}^{r_m} \rho_{mk|l}^{I(y_{im}=k)}}{\sum_{c=1}^{C} \gamma_c(\mathbf{x}_i) \prod_{m=1}^{M} \prod_{k=1}^{r_m} \rho_{mk|c}^{I(y_{im}=k)}}. \tag{2.5}$$

2.2. 다항범주자료에서 Rubin의 인과모형 및 성향점수

Rubin의 인과모형(Rubin's causal model; RCM)에서는 한 개체가 여러 처치를 동시에 받았을 때, 각 처치에 대한 결과들의 차이를 인과효과라고 보았다 (Rubin, 1974, 1977, 1978). 그러나 현실에서 하나의 개체는 하나의 처치만 받을 수 있어 인과효과를 직접 추정할 수 없으므로, 가실험설계(pseudo experimental design)에 근거하여 추정되는 원인의 영향력을 밝히는 것이 RCM의 목적이라고 할 수 있다.

RCM의 이해를 돕기위해 이항형 처치변수와 결과변수로 이루어진 자료를 Table 2.1과 같이 나타내었다. Table 2.1은 처치변수와 결과변수가 이항형 변수인 자료의 형태를 나타내고 있다. T_i는 처치를 받

았는지에 대한 정보를 나타내며 i번째 개체가 처치를 받았을 경우 1, 받지 않았을 경우 0을 의미하는 지시변수이다. $Y_i(0)$은 i번째 개체가 처치를 받지 않았을 때 개체가 가지는 결과범주이고 $Y_i(1)$은 처치를 받았을 때 개체가 가지는 결과범주를 나타내며 음영의 셀은 결측된 결과범주를 의미한다. 즉, 첫 번째 행의 개체부터 m번째 행의 개체까지는 처치를 받지 않았으므로 $Y_i(0)$은 관측되지만 $Y_i(1)$은 관측되지 않는다. 반대로 $m+1$번째 행의 개체에서 n번째 행의 개체까지는 처치를 받았으므로 $Y_i(1)$은 관측되지만 $Y_i(0)$은 관측되지 않는다. 따라서 모든 개체에서 결측자료가 발생하게 되어 개별적인 인과효과를 추정할 수 없으며 인과효과추론의 문제는 결측자료의 문제가 되는 것이다. 이러한 문제를 해결하기 위하여 평균인과효과(average causal effect; ACE)를 추정하게 된다 (Rosenbaum과 Rubin, 1983). 연속형 결과변수에 대한 ACE는 처치 간 결과의 평균 차이로 나타낼 수 있으나 범주형 결과변수에 대한 ACE는 다음과 같은 오즈비로 정의할 수 있다. 예를 들어, 이항처치범주 $(0,1)$에 대한 이항결과범주 $(1,2)$의 ACE는 식 (2.6)과 같이 나타낼 수 있다.

$$\begin{aligned}\text{ACE} &= \frac{P(Y(1)=2)/P(Y(1)=1)}{P(Y(0)=2)/P(Y(0)=1)} \\ &= \frac{P(Y=2 \mid T=1)/P(Y=1 \mid T=1)}{P(Y=2 \mid T=0)/P(Y=1 \mid T=0)}.\end{aligned} \tag{2.6}$$

결국 ACE의 추정문제는 처치범주가 주어졌을 때 특정한 결과범주에 속할 조건부확률을 추정하는 문제가 되는 것이다. 또한, 인과효과를 추론할 때 전체모집단을 대상으로 인과효과를 추론하는 ACE 외에 특정 처치를 받은 집단을 대상으로 인과효과를 추론하는 평균처리인과효과(average causal effect on the treated; ACE_t)도 주된 관심 모수가 된다. ACE_t의 추정문제는 처치범주 j에 속한 개체들이 처치범주 t가 주어질 경우 특정한 결과범주 k에 속할 조건부확률을 추정하는 문제가 되는 것이다.

관찰연구에서는 처치의 할당이 무작위로 배정될 수 없으므로 식 (2.6)를 이용하여 ACE 및 ACE_t 추정하면 편향이 발생하게 된다. 이러한 편향은 처치와 결과 모두에 영향을 미치는 제3의 변수인 교란변수 $\mathbf{X}_i$에 의해 발생한다. 즉, 교란변수가 주어진 상황에서 처치의 할당과 결과가 독립이라는 조건부 독립의 가정이 만족하면 결측된 결과변수는 무작위 결측(missing at random, MAR)의 조건을 충족하게 된다 (Rubin, 1976). 따라서, 처치와도 관련이 있고 결과변수와도 관련이 있는 교란변수를 이용하여 ACE 및 ACE_t의 불편향 추정치를 구할 수 있다 (Rosenbaum과 Rubin, 1983). Rosenbaum과 Rubin (1983)은 여러 개의 교란변수, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$를 이용하여 성향점수를 계산하고 이를 통해 ACE를 추정하는 방법을 제시하였다. 성향점수는 교란변수가 주어진 조건에서 특정 처치에 속할 확률로 정의된다. 즉, 성향점수가 주어진 상황에서 처치의 할당과 결과는 서로 독립이므로 성향점수로 보정한 ACE 추정량은 불편추정량이 된다 (Rosenbaum과 Rubin, 1983). 처치가 L개의 범주를 갖는 다항범주인 경우 i번째 개체가 처치범주 t를 받았을 때 성향점수 $\pi_{it}(\mathbf{x}_i)$는 다항로지스틱 회귀분석을 이용하여 다음과 같이 추정할 수 있다.

$$\hat{\pi}_{it}(\mathbf{x}_i) = \frac{\exp\left(\mathbf{x}_i\hat{\boldsymbol{\xi}}_t\right)}{1+\sum_{l=1}^{L-1}\exp\left(\mathbf{x}_i\hat{\boldsymbol{\xi}}_l\right)}. \tag{2.7}$$

본 논문에서는 ACE 추정량의 편향을 제어하는 방법으로 성향점수의 역확률 가중치(inverse probability weighting; IPW) 방법을 적용할 것이다 (Robins 등, 2000; Rosenbaum, 2002). IPW 방법은 가중치를 주어 성향점수가 높은 개체는 낮게 보정하고, 낮은 개체는 높게 보정하여 처리범주 간의 균형을 맞추어 주는 방법이다. 이 방법은 개체들의 처치가 무작위로 할당되는 상황과 유사하게 만들어준다. ACE 및 ACE_t의 불편추정량을 구하기 위한 가중치는 각각 아래와 같다 (McCaffrey 등, 2013; Frölich,

2004).

$$w_i(t) = \frac{1}{\hat{\pi}_{it}(\mathbf{x}_i)}, \qquad w_i(t,j) = \frac{\hat{\pi}_{ij}(\mathbf{x}_i)}{\hat{\pi}_{it}(\mathbf{x}_i)}. \tag{2.8}$$

ACE의 불편추정량 계산을 위한 가중치 $w_i(t)$는 i번째 개체가 처치범주 t에 속할 때 주어지는 가중치이며, ACE_t의 불편추정량 계산을 위한 가중치 $w_i(t,j)$는 처치범주 j를 받은 i번째 개체를 대상으로 그 개체가 처치범주 t에 속할 때 주어지는 가중치이다.

2.3. 상황 별 평균인과효과 추정

본 절에서는 Park과 Chung (2014)에 의해 제안된 방법(propensity for logit model; PLogit)과 본 연구에서 제안하는 방법(propensity for LCA; PLCA)을 소개하고자 한다. 두 방법론의 성능 비교를 위한 모의실험을 위해 총 네 가지의 경우를 고려하고자 한다. 첫 번째 경우는 처치변수와 결과변수 둘 다 관측변수인 상황이며, 두 번째 경우는 처치변수가 잠재변수이고 결과변수가 관측변수인 상황이다. 세 번째 경우는 처치변수가 관측변수이고 결과변수가 잠재변수인 상황으로 Lanza 등 (2013)이 ACE 추정 방법을 제안했던 상황이다. 마지막으로 네 번째 경우는 처치변수와 결과변수가 모두 잠재변수인 상황으로 Park과 Chung (2014)이 방법을 제안했던 상황이다. 네 가지의 경우 가운데 처치변수가 관측변수인 상황, 즉 첫 번째와 세 번째의 경우는 식 (2.7)를 이용하여 성향점수를 계산한 후, 식 (2.8)을 통해 가중치를 계산할 수 있다. 따라서 결과변수가 잠재변수인 경우에는 가중치를 사용한 LCA를 통해 모형의 모수를 추정하고 이에 따라 ACE 및 ACE_t를 계산할 수 있다 (Lanza 등, 2013). 본 논문에서는 처치변수가 잠재변수인 상황, 즉 두 번째와 네 번째의 경우에 Park과 Chung (2014)이 제안한 PLogit을 개선한 PLCA를 제안하며 모의실험을 통해 PLCA의 성능을 기존 방법인 PLogit의 성능과 비교하고자 한다. ACE와 ACE_t의 추정방법은 성향점수를 이용한 가중치 계산만 다를 뿐 모든 절차는 동일하므로 본 논문에서는 ACE의 추정방법만 제시하고자 한다. 처치변수가 잠재변수인 경우 PLogit과 PLCA를 ACE 추정방법은 다음과 같다.

- 성향점수 추정

  PLogit: 처치범주에 관련된 관측문항으로 식 (2.1)에 주어진 LCA를 실시한다. i번째 개체의 사후확률을 식 (2.2)를 통해 계산하고 가장 높은 사후확률을 가지는 잠재처치범주를 T_i에 할당하며 이를 모든 n개체에 대해 시행한다. 잠재처치범주 T_i를 반응변수로 하고 교란변수 $\mathbf{x}_i$를 설명변수로 하는 다항로지스틱 회귀분석을 적합하여 식 (2.7)의 $(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{L-1})$의 추정치를 이용하여 성향점수를 추정한다.

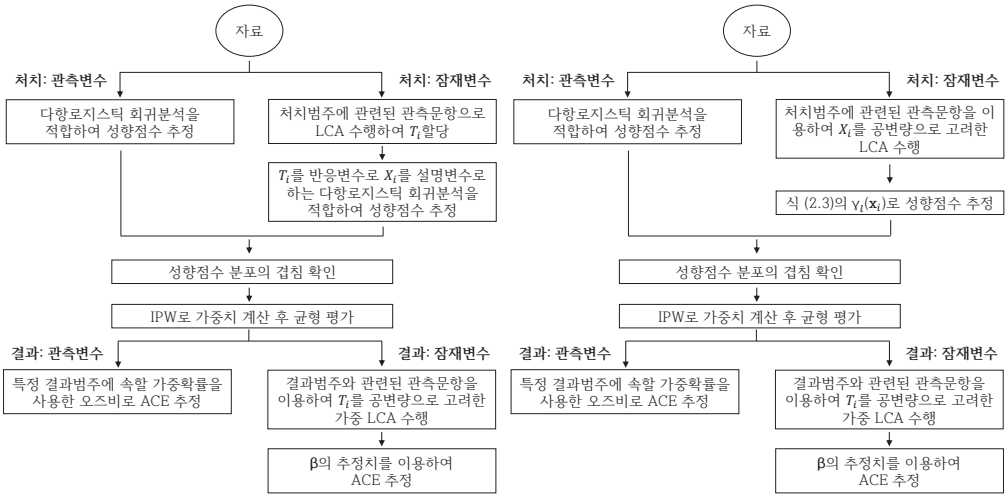
  PLCA: 처치범주에 관련된 관측문항을 사용하여 식 (2.3)와 같이 교란변수 $\mathbf{x}_i$를 공변량으로 고려한 LCA를 실시한다. 이때 얻어지는 식 (2.5)의 사후확률을 이용하여 가장 높은 사후확률을 가지는 잠재처치범주를 T_i에 할당하며 식 (2.4)의 $(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{L-1})$의 추정치를 이용하여 성향점수를 계산한다.

- 성향점수 분포의 겹침(overlap) 평가

  성향점수를 추정한 후 앞의 단계에서 할당된 처치범주 간 성향점수 분포의 겹침이 이루어졌는지 확인해야 한다. 겹침이 이루어지지 않으면 두 처치 간 비교 가능한 개체가 없다는 것을 의미하기 때문이다. 성향점수 분포의 겹침은 처치를 받은 그룹과 받지 않은 그룹 간 성향점수에 대한 박스플롯을 통해 확인할 수 있다.

- 가중치 계산 및 균형 평가

  교란변수로 인한 편향을 조정하기 위해 식 (2.8)과 같이 각 ACE에 대한 가중치를 계산한다. 가중치가 계산되면 가중치를 적용하였을 때 처치 간 교란변수의 균형이 이루어지는지 평가해야 한다. 균



(a) PLogit (b) PLCA

Figure 2.1. The calculation flows of PLogit and PLCA methods for ACE estimation. PLogit = propensity for logit model; PLCA = propensity for latent class analysis; ACE= average causal effect.

형 평가는 가중치를 적용하기 전과 후의 교란변수의 처치범주 간 표준평균차이(standardized mean difference; SMD)를 구하여 평가할 수 있다. 가중치를 적용한 후 SMD의 절댓값이 0.2보다 작으면 교란변수의 처치 간 균형이 이루어졌다고 볼 수 있다 (Cohen, 1988). 만약 균형이 이루어지지 않으면 교호작용항이나 고차항을 추가하여 균형을 맞추어야 한다.

- ACE 추정

  관측결과변수: 균형이 이루어지면 식 (2.6)에 의해 ACE를 추정한다. 이때 위의 단계에서 계산된 가중치를 적용하여 계산된 처치 내에서 결과변수의 특정 범주에 속할 가중확률을 사용한 오즈비로 ACE를 추정한다.

  잠재결과변수: 결과변수에 관련된 관측문항을 가지고 LCA를 수행한다. 이때, 처치범주를 공변량으로 고려하며 위의 단계에서 계산된 가중치를 적용한 LCA를 수행한다. 잠재처치변수의 경우 앞에서 할당된 처치범주 T_i를 공변량으로 사용한 LCA에 가중치를 적용해야 한다. 그 결과로 부터 특정한 처치범주가 주어졌을 때 결과변수의 특정 범주에 속할 조건부확률을 사용하여 ACE를 추정한다.

Figure 2.1은 네 가지 경우에 대한 PLogit 및 PLCA 방법을 순서흐름도로 표현한 것이다. 처치변수가 관측변수일 때 두 방법 모두 처치변수를 반응변수로, 교란변수를 설명변수로 하는 다항로지스틱 회귀분석을 시행하여 성향점수를 추정한다. 처치변수가 잠재변수일 경우 PLogit은 처치범주에 관련된 관측문항을 이용하여 LCA를 수행한 후 개체마다 잠재처치범주를 할당한다. 할당된 범주를 반응변수로, 교란변수를 설명변수로 다항로지스틱 회귀분석을 통해 성향점수를 추정한다. 반면 PLCA의 경우 교란변수를 공변량으로 사용한 LCA를 수행하여 잠재처치범주를 할당하기 전에 성향점수를 추정한다. 성향점수가 추정되면, 성향점수 분포의 겹침을 확인하며 IPW를 통해 가중치를 계산한 후 교란변수의 균형을 평가하게 된다. 균형이 이루어 졌으면 ACE를 추정하게 된다. 결과변수가 관측변수인 경우에는 처치가 주어진 경우 특정 결과범주에 속할 가중확률을 이용하여 ACE를 추정하며 결과변수가 잠재변수인 경우에는 결과범주에 관련된 관측문항을 이용하여 처치를 공변량으로 고려한 가중 LCA를 수행한 후, 식 (2.4)에서 주어진 $\boldsymbol{\beta}$의 추정치를 이용하여 ACE를 계산하게 된다.

3. 모의실험연구

3.1. 자료 생성

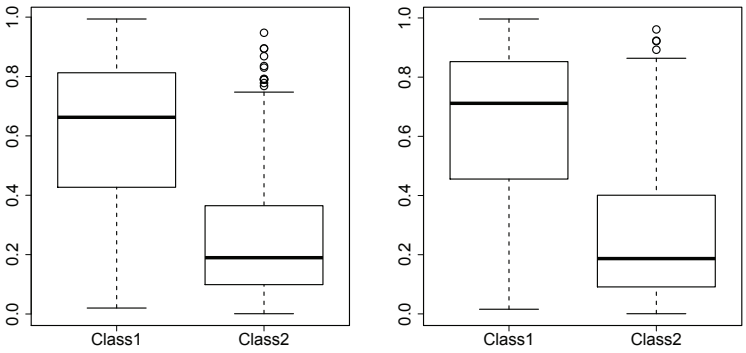
ACE 추정방법의 성능을 파악하기 위해 다음과 같이 자료를 생성하였다. 표본 수는 500이며, 먼저 5개의 교란변수, $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_5)^T$를 평균 $\boldsymbol{\mu} = (0, 1, -1, 1, -1)^T$과 분산 1을 가지는 정규분포에서 각각 독립적으로 생성하였다. 자료는 각각 4개의 처치 문항과 결과 문항을 가지고 있으며 각 문항은 이항범주를 가진다. 또한, 각각 2개의 처치범주와 결과범주를 가지고 있다. 자료는 추정에 필요한 $\boldsymbol{\beta}$, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, $\boldsymbol{\rho}_1$, $\boldsymbol{\rho}_2$, $\boldsymbol{\eta}_1$, $\boldsymbol{\eta}_2$와 같은 모수의 참값을 지정한 후 생성하였다. 자료 생성 절차는 다음과 같다.

1) 성향점수 생성 및 처치범주의 할당
   처치범주를 할당하기 위해 참값 $\boldsymbol{\beta} = (-0.3, 0.5, 0.8, 1.2, 0.8, -1)^T$를 지정한 후 식 (2.4)를 계산하여 개체가 처치범주에 속할 확률을 계산한다. 이와 같은 $\boldsymbol{\beta}$의 참값을 사용하면 개체의 63%정도가 처치범주 1에 속하게 되며, 37%정도가 처치범주 2에 속하게 되어 특정한 처치범주에 속할 확률이 급격히 치우치지 않게된다.
2) 처치문항 생성
   처치문항을 생성하기 위해 단계 1)에서 할당된 처치범주의 문항확률 $\boldsymbol{\rho}_1 = (0.9, 0.9, 0.9, 0.9)^T$과 $\boldsymbol{\rho}_2 = (0.1, 0.1, 0.1, 0.1)^T$를 이용하여 처치문항을 생성한다. $\boldsymbol{\rho}_1$는 처치범주가 1일 경우 4개의 처치문항에 대해 첫 번째 범주로 응답할 확률이며, $\boldsymbol{\rho}_2$는 처치범주가 2일 경우 4개의 처치문항에 대해 첫 번째 범주로 응답할 확률이다. 각 확률이 0.9와 0.1과 같이 1과 0에 가까운 것은 처치문항이 각 처치범주를 측정하기 위한 좋은 관측변수임을 나타내며 $\boldsymbol{\rho}_1$과 $\boldsymbol{\rho}_2$가 크게 다른 것은 두 처치범주의 특성이 매우 상이하다는 것을 나타낸다.
3) 결과범주 할당
   개체가 처치범주 1에 속한다면 $\boldsymbol{\beta}_1 = (-0.8, 0.5, 0.8, -1, 0.8, 0.5)^T$를 적용하고, 처치범주 2에 속한다면 $\boldsymbol{\beta}_2 = (-0.3, 0.5, 0.8, 1.2, 0.8, -1)^T$를 적용하여 개체가 결과범주에 속할 확률을 식 (2.4)를 통해 계산하며 계산된 확률로 결과범주 1과 2를 할당한다. 이는 처치범주에 따라 결과범주에 속할 확률을 다르게 주기 위함이다. 처치범주 1의 경우, $\boldsymbol{\beta}_1$을 적용하여 결과범주 1에 속할 확률은 53%정도이며, 결과범주 2에 속할 확률은 47%정도이다. 처치범주 2의 경우, $\boldsymbol{\beta}_2$를 적용하여 결과범주 1에 속할 확률은 49%정도이며, 결과범주 2에 속할 확률은 51%정도가 된다.
4) 결과문항 생성
   단계 2)와 같이 할당된 결과범주의 문항확률 $\boldsymbol{\eta}_1 = (0.9, 0.9, 0.9, 0.9)^T$과 $\boldsymbol{\eta}_2 = (0.1, 0.1, 0.1, 0.1)^T$를 이용하여 결과문항을 생성한다. $\boldsymbol{\eta}_1$는 결과범주가 1일 경우 4개의 결과문항에 대해 첫 번째 범주로 응답할 확률이며, $\boldsymbol{\eta}_2$는 결과범주가 2일 경우 4개의 결과문항에 대해 첫 번째 범주로 응답할 확률이다. 0.9와 0.1과 같이 1과 0에 가까운 것은 결과문항이 각 결과범주를 측정하기 위한 좋은 관측변수임을 나타내며 $\boldsymbol{\eta}_1$과 $\boldsymbol{\eta}_2$가 크게 다른 것은 두 결과범주의 특성이 매우 상이하다는 것을 나타낸다.

이렇게 생성된 하나의 자료로 네 가지 상황에서 ACE를 추정할 수 있다. 이 자료에서 ACE의 참값은 식 (2.6)과 같이 계산된 가중치를 적용하여 처치범주가 주어진 상황에서 결과범주에 속할 가중확률을 이용한 오즈비가 된다.

3.2. 결과

본 절에서는 처치변수가 잠재변수인 상황에서 PLogit과 PLCA의 모의실험 결과을 제시하고자 한다. ACE 추정 방법의 성능을 평가하기 위해 100개의 독립적인 데이터를 생성하여 모의실험을 진행하였다.



(a) PLogit (b) PLCA

Figure 3.1. Boxplots for distribution of propensity score when treatment is latent. PLogit = propensity for logit model; PLCA = propensity for latent class analysis.

Table 3.1. ACE and ACE_t estimates from PLogit and PLCA (average of relative biases*) when treatment is latent

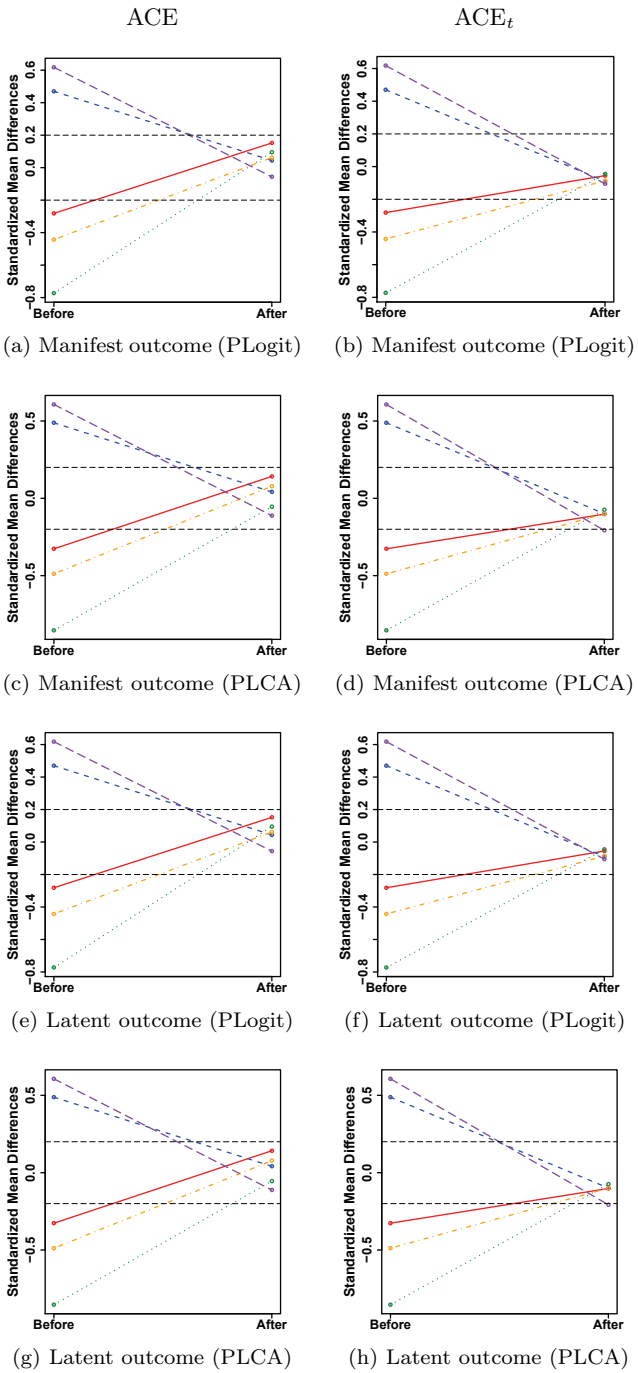
결과변수	참값		보정 전		PLogit		PLCA	
	ACE	ACE_t	ACE	ACE_t	ACE	ACE_t	ACE	ACE_t
관측변수	7.209	3.715	4.720	4.720	4.856	3.303	6.540	3.574
			(0.337)	(0.544)	(0.332)	(0.242)	(0.154)	(0.187)
잠재변수	7.209	3.715	5.298	5.298	6.867	3.488	6.534	3.606
			(0.279)	(0.690)	(0.218)	(0.216)	(0.177)	(0.219)

*: relative bias = $(\widehat{\text{ACE}} - \text{ACE})/\text{ACE}$ or $(\widehat{\text{ACE}}_t - \text{ACE}_t)/\text{ACE}_t$.
ACE = average causal effect; ACE_t = ACE on the treated; PLogit = propensity for logit model; PLCA = propensity for latent class analysis.

Figure 3.1은 PLogit과 PLCA를 이용하여 추정한 성향점수 분포의 겹침을 평가하기 위한 박스플롯이며 이를 통해 두 방법 모두 전반적으로 결과변수의 종류에 상관없이 처치 간 분포의 겹침이 이루어졌음을 확인할 수 있다. Figure 3.1은 100개의 데이터 중 하나의 데이터에 대한 박스플롯이며 나머지 99개의 데이터에서도 유사한 결과를 보여 처치 간 성향점수 분포 겹침의 수준에 문제가 없음을 알 수 있다.

Figure 3.2는 ACE 및 ACE_t의 가중치를 주기 전과 후의 처치범주 간 교란변수의 SMD를 계산하여 비교한 그림으로써 왼쪽 다섯 개 점은 가중치를 적용하기 전 처치범주 1과 처치범주 2에 해당하는 교란변수의 SMD를 의미하며, 오른쪽 다섯 개 점은 가중치를 적용한 후 교란변수의 SMD를 의미한다. 그래프의 수평 점선은 -0.2와 0.2를 의미하며 다섯 개의 점들이 이 구간 안에 포함되어 있으면 처치 간 교란변수의 균형이 이루어졌음을 의미한다. Figure 3.2를 살펴보면 PLogit과 PLCA 두 방법 모두 ACE 및 ACE_t의 가중치를 적용한 후 교란변수의 SMD가 -0.2와 0.2의 구간 안에 포함되어 있어 교란변수의 균형이 이루어졌음을 확인할 수 있다. Figure 3.2은 100개의 데이터 중 하나의 데이터에 대한 처치 간 교란변수의 SMD를 비교하기 위한 그림이며 나머지 99개의 데이터에서도 유사한 결과를 보여 처치 간 교란변수의 균형에 문제가 없음을 알 수 있다.

Table 3.1은 처치변수가 잠재변수인 경우에 100개의 자료에 대한 참값의 평균 및 ACE와 ACE_t 추정치의 평균을 나타내며 가로안에 있는 수치는 상대적 편향의 평균을 나타낸다. 100개의 자료에 대해 각각 참값을 계산하고 ACE와 ACE_t를 추정하기 때문에 Table 3.1은 100개의 값에 대한 평균값으로 나



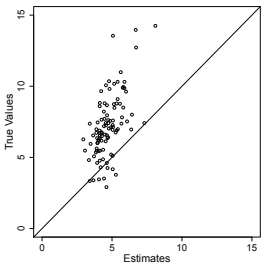
(a) Manifest outcome (PLogit) (b) Manifest outcome (PLogit)

(c) Manifest outcome (PLCA) (d) Manifest outcome (PLCA)

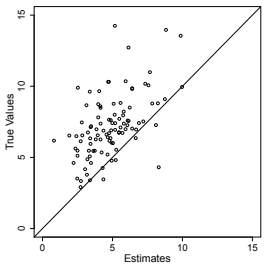
(e) Latent outcome (PLogit) (f) Latent outcome (PLogit)

(g) Latent outcome (PLCA) (h) Latent outcome (PLCA)

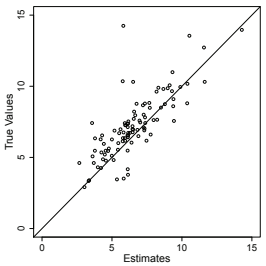
Figure 3.2. Plots of standardized mean difference (SMD) in the confounders when treatment is latent (Each of five colored lines represents confounder's line connecting two SMDs between treatments before (left side points) and after (right side points) using ACE and ACE$_t$ weights). PLogit = propensity for logit model; PLCA = propensity for latent class analysis.



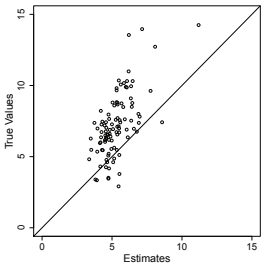
(a) Non-adjusted ACE in Manifest outcome



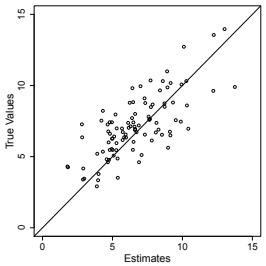
(b) ACE from PLogit in Manifest outcome



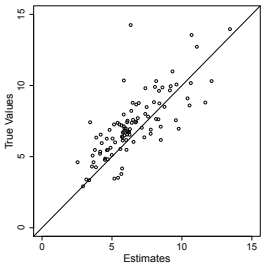
(c) ACE from PLCA in Manifest outcome



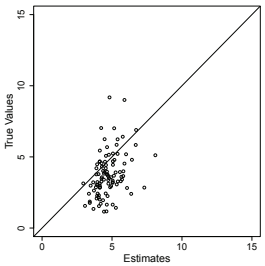
(d) Non-adjusted ACE in Latent outcome



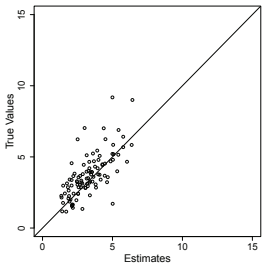
(e) ACE from PLogit in Latent outcome



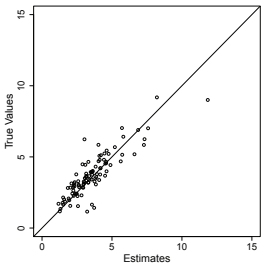
(f) ACE from PLCA in Latent outcome



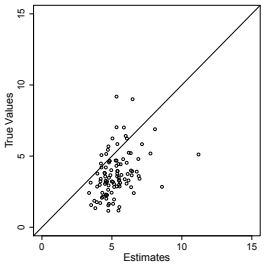
(g) Non-adjusted ACE_t in Manifest outcome



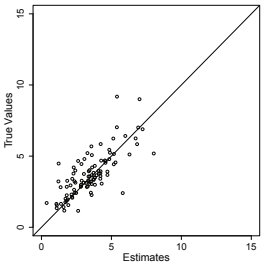
(h) ACE_t from PLogit in Manifest outcome



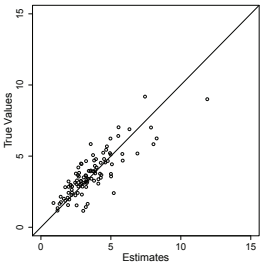
(i) ACE_t from PLCA in Manifest outcome



(j) Non-adjusted ACE_t in Latent outcome



(k) ACE_t from PLogit in Latent outcome



(l) ACE_t from PLCA in Latent outcome

Figure 3.3. Scatter plots of true values of ACE and ACE_t *versus* their estimates from three different methods (non-adjusted, PLogit, and PLCA) for 100 samples. ACE = average causal effect; ACE_t = ACE on the treated; PLogit = propensity for logit model; PLCA = propensity for latent class analysis.

타내었다. 생성된 100개의 자료에서 ACE의 참값의 평균은 7.209이며 ACE_t의 참값의 평균은 3.715이다. 보정 전의 ACE의 추정치의 평균과 ACE_t의 추정치의 평균은 결과변수가 관측변수일 경우 4.720, 잠재변수일 경우 5.298이며 상대적 편향의 평균값도 가장 큰 것을 알 수 있다. PLogit과 PLCA의 경우 가중치를 주지 않았을 때와 비교하여 결과변수의 종류에 관계없이 ACE와 ACE_t의 추정치의 평균은 참값의 평균과 유사함을 확인할 수 있다. 그러나 ACE 추정치의 평균에서 PLCA가 PLogit에 비해 상대적 편향의 평균이 작은 추정치를 제공하며, ACE_t 추정치에 대한 평균의 경우 결과변수가 관측변수일때 PLCA가 PLogit에 비해 상대적 편향의 평균이 작은 추정치를 제공하는 것을 확인할 수 있다.

Figure 3.3은 100개 독립적인 데이터에 대한 참값과 추정치를 비교한 그림이다. 보정 전 ACE 추정치는 결과변수 종류에 관계없이 과소추정되는 경향이 있고 ACE_t 추정치는 과대추정되는 경향을 보임을 알 수 있다. 전반적으로 관측결과변수와 잠재결과변수인 상황에서 가중치를 주지 않은 경우와 비교하여 PLogit의 추정치의 편향이 감소하였지만 PLCA에 비해 여전히 편향이 존재하는 것을 알 수 있다. 즉, PLCA의 추정치가 상대적으로 안정적인 것을 확인할 수 있다.

4. 사례분석

4.1. 자료 소개

PLCA의 실증적 분석을 위하여 미국 여성 청소년을 대상으로 약물사용에 대한 청소년기 신체성숙도의 인과효과를 추론하고자 한다. 청소년기 신체적 성숙 정도가 약물사용에 미치는 영향을 추정하기 위해 'National Longitudinal Study of Adolescents Health (Add Health)' (Udry, 2003)의 자료를 이용하였다. Add Health는 1994년에서 1995년에 미국의 7–12학년 청소년 11,796명을 대상으로 시작한 종단연구이며, 응답자의 사회적, 경제적, 심리적, 신체적 상태 등 포괄적 내용을 다루고 있다. 2차 조사는 1996년에 시행되었으며 3차 조사는 청소년부터 청년까지의 변화를 알아보기 위하여 2001에서 2002년에 걸쳐 시행되었다.

본 장에서의 처치변수는 신체성숙도에 대한 잠재범주이며 1차 조사와 2차 조사에서 얻어진 자료를 사용하여 추정하였다. 결과변수는 약물복용에 대한 잠재범주이며 1996년에 시행된 2차 조사에서 얻어진 자료를 사용하여 추정하였다. 교란변수로는 영어와 수학 성적, 아침 식사 여부, 몸무게, 키, 인종, 부모님과의 관계, 소득 등 22개의 항목을 사용하였으며 1차 조사에서 얻어진 자료를 사용하였다. 분석 대상은 여성 청소년 3,356명이다.

여성 청소년의 신체성숙도를 측정하기 위해 '초등학생 대비 가슴발달 정도(Breast)'와 '체형 변화정도(Body curvy)' 두 문항을 사용하였다. 가슴발달 정도는 1 = '많이 변화함'(많이 크다, 훨씬 크다), 2 = '약간 변화함'(같은 크기, 약간 크다, 크다)로 정의하였으며 체형 변화정도는 1 = '많이 변화함'(많이 굴곡지다, 훨씬 굴곡지다), 2 = '약간 변화함'(같은 굴곡, 약간 굴곡지다, 굴곡지다)로 정의하여 이항변수로 변환하였다.

약물복용을 측정하기 위해 '한 달 동안 술을 마신 횟수(AlcUse)', '한 달 동안 흡연한 횟수(CigUse)','일년 동안 한번에 5잔 이상 술을 마신 횟수(5 + Drinks)', '한 달 동안 취한 횟수(Drunk)'를 조사하여 1회 이상이면 1 = '있음'으로, 0회는 응답했으면 2 = '없음'으로 정의하여 이항변수로 변환하였다.

4.2. 신체성숙도: 처치변수에 대한 잠재범주분석

본 장에서는 Add Health 자료를 이용하여 약물복용에 대한 청소년 신체성숙도의 인과효과를 추정하기 위해여 PLCA 분석 방법을 적용한 결과를 보여준다. 신체성숙과 관련된 관측문항, 약물복용과 관련된

Table 4.1. Model selection for adolescent puberty

청소년 성장	G^2	df	AIC	BIC	CAIC	aBIC
2개 범주 모형	406.20	6	424.20	479.23	488.23	450.63
3개 범주 모형	137.10	1	165.10	250.70	264.70	206.22

AIC = Akaike information criterion; BIC = Bayesian information criterion; CAIC = consistent Akaike information criterion; aBIC = adjusted Bayesian information criterion.

Table 4.2. Item-response probabilities for adolescents puberty

성장 범주	출현율	첫 번째 조사		두 번째 조사	
		가슴발달	체형변화	가슴발달	체형변화
조숙	0.414	0.995	0.773	0.848	0.751
보통성숙	0.185	0.001	0.432	0.490	0.997
만숙	0.402	0.106	0.150	0.137	0.009

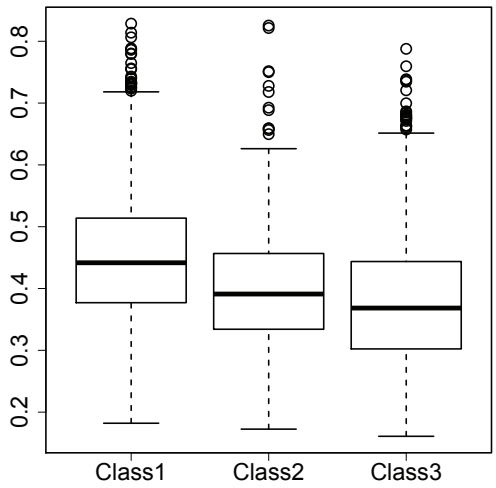


Figure 4.1. The overlap for distribution of propensity score between latent classes of puberty.

관측문항, 교란변수에 결측이 존재한다. 교란변수의 결측은 MAR로 가정하여 다중대체를 실시하였다. 10개의 대체 데이터를 Rubin의 규칙에 의해 종합한 결과를 제시하였다.

본 절에서는 신체성숙도와 관련된 문항을 사용하여 LCA를 수행한 결과를 보여준다. 청소년 신체성숙도와 관련된 관측문항은 가슴발달 정도와 신체변화 정도이며 1차, 2차 조사에서 측정되었다.

Table 4.1는 잠재처치범주의 수를 결정하기 위해 잠재범주모형에 대한 적합통계량을 정리한 결과이며 모든 적합통계량에서 세 개의 잠재범주를 가지는 모형이 좋기 때문에 잠재처치범주 수는 세 개로 결정하였다.

Table 4.2는 신체성숙도와 관련하여 세 개의 범주를 가지는 잠재범주모형을 적합시킨 결과를 나타낸다. 첫 번째 범주의 문항응답확률을 살펴보면, 첫 번째 범주에 속해 있는 대부분의 청소년은 네 개의 모든 신체성숙 항목에서 '많이 변화'했다고 응답하여 신체적 성숙이 빠른 시기에 이루어진 '조숙(early puberty)' 청소년임을 알 수 있다. 전체 모집단의 약 41%가 이 범주에 속해있다. 두 번째 범주에 속

해 있는 청소년은 1차 조사에서 가슴발달(0.001)과 체형변화(0.432)가 없었지만 2차 조사에서 가슴발달(0.490)과 체형변화(0.997)가 급격히 이루어진 '보통성숙(mid-puberty)' 청소년임을 알 수 있다. 이 그룹의 출현율은 전체 모집단의 19% 정도이다. 세 번째 범주에 속해 있는 대부분의 청소년은 네 개의 성장 항목에서 '약간 변화'했다고 응답하여 '만숙(late puberty)' 청소년임을 알 수 있다. 전체 모집단의 약 40% 정도가 이 범주에 속해있다.

다음으로 LCA로 추정된 사후확률에 따라 개체들에게 잠재처치범주를 할당하고 식 (2.4)의 $(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{L-1})$로 성향점수를 추정하였다. 성향점수 추정 후, 할당된 처치범주 간 성향점수 분포의 겹침을 확인해야하며 Figure 4.1은 분포의 겹침을 확인하기 위한 박스플롯이다. Figure 4.1을 보면 세 개의 잠재범주에서 전반적으로 겹침이 이루어졌음을 확인할 수 있다. Figure 4.1은 10개의 대체 자료 중 하나의 결과를 나타내며 나머지 9개의 자료에서도 유사한 결과를 보여준다. 따라서 식 (2.8)과 같이 ACE 가중치와 ACE_t 가중치를 계산한 후 교란변수로 인한 편향이 조정되었는지 확인하기 위해 잠재처치범주 간 교란변수의 균형을 평가하였다.

Figure 4.2는 ACE와 ACE_t의 가중치를 주기 전과 후의 잠재처치범주 간 교란변수의 SMD를 계산하여 비교한 그림이다. 왼쪽 그림은 조숙 청소년과 만숙 청소년을 비교한 것으로 그림의 왼쪽 점은 가중치를 적용하기 전 조숙 청소년과 만숙 청소년에 해당하는 교란변수의 SMD를 의미하며 오른쪽 점은 가중치를 적용한 후 교란변수의 SMD를 의미한다. 오른쪽 그림은 보통성장 청소년과 만숙 청소년을 비교한 것으로 그림의 왼쪽 점은 가중치를 적용하기 전 보통성장 청소년과 만숙 청소년에 해당하는 교란변수의 SMD를 의미하며 오른쪽 점은 가중치를 적용한 후 교란변수의 SMD를 의미한다. 그래프의 수평 점선은 -0.2와 0.2를 의미하며 가중치를 적용한 후 교란변수의 SMD가 이 구간 안에 포함되어 있으면 처치간 교란변수의 균형이 이루어졌음을 의미한다. Figure 4.2를 살펴보면 ACE 및 ACE_t의 가중치를 적용한 후 교란변수의 SMD가 -0.2와 0.2의 구간 안에 포함되어 있어 교란변수의 균형이 이루어졌음을 확인할 수 있다. Figure 4.2은 10개의 대체 자료 중 하나의 결과를 나타내며 나머지 9개의 자료에서도 유사한 결과를 보여준다.

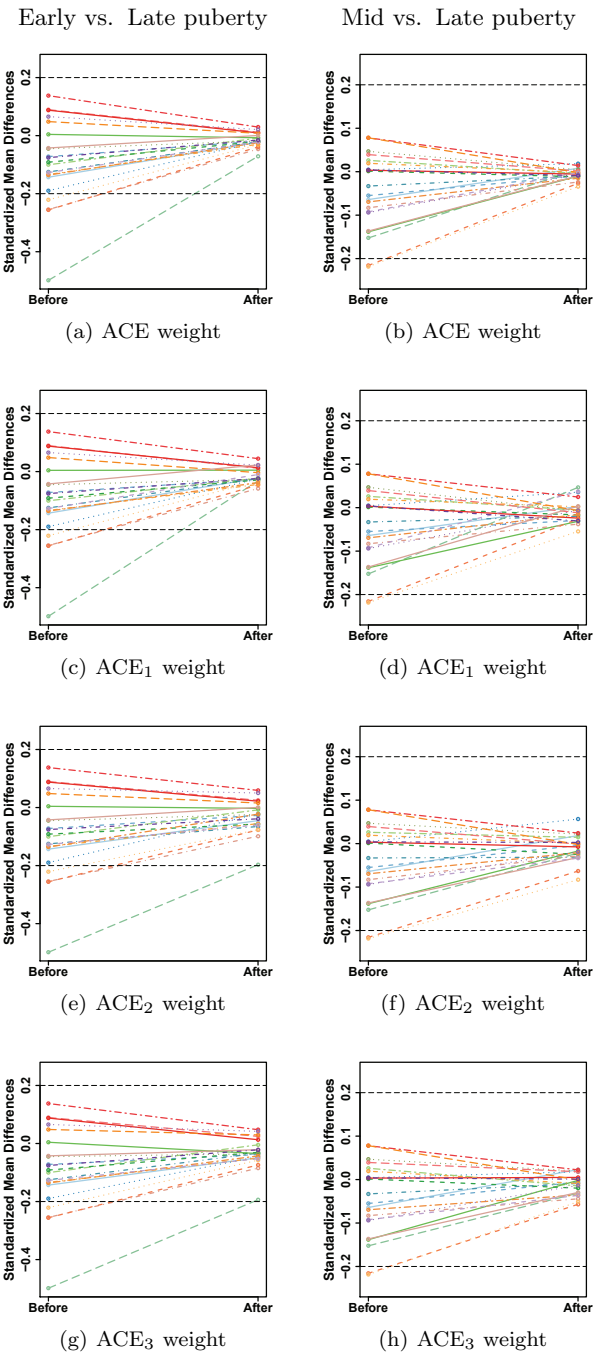
4.3. 약물복용: 결과변수에 대한 잠재범주분석

본 절에서는 약물복용과 관련된 네 개의 문항으로 LCA를 수행한 후 평균인과효과를 추정하고자 한다. 이 때 LCA 모형에서 앞서 할당된 잠재처치범주를 공변량으로 고려하였다.

Table 4.3은 잠재결과범주의 수를 결정하기 위해 잠재범주모형에 대한 적합통계량을 정리한 결과이며 AIC를 제외한 모든 적합통계량에서 세 개의 범주를 가지는 모형이 적합한 것을 확인할 수 있다. 또한, 잠재범주가 세 개인 모형에서 해석이 적절하게 이루어졌기 때문에 잠재결과범주 수는 세 개로 결정하였다.

Table 4.4는 약물복용에 대한 세 개의 범주를 가지는 잠재범주모형을 적합시킨 결과를 나타낸다. 이때 잠재처치범주를 공변량으로, ACE와 ACE_t 가중치를 각각 적용하여 가중 LCA를 수행하였다. ACE 가중치를 적용한 모형과 ACE_t 가중치를 적용한 모형의 해석이 동일하므로 ACE 가중치를 준 LCA 결과만 나타냈다.

첫 번째 범주의 문항응답확률을 살펴보면, 첫 번째 범주에 속해 있는 대부분의 청소년은 네 개의 모든 약물복용 항목에서 '그런적이 없다'고 응답하여 약물을 복용하지 않는 '약물 미복용' 그룹에 해당함을 알 수 있다. 두 번째 범주에 속해 있는 대부분의 청소년이 흡연을 한 적이 있다고 응답(0.999)하여 '흡연' 그룹에 해당하는 것을 알 수 있다. 세 번째 범수에 속해 있는 청소년은 네 개의 모든 약물복용 항목에서 '그런적이 있다'고 응답하여 '일반적 약물복용' 그룹임을 알 수 있다.



(a) ACE weight (b) ACE weight

(c) ACE_1 weight (d) ACE_1 weight

(e) ACE_2 weight (f) ACE_2 weight

(g) ACE_3 weight (h) ACE_3 weight

Figure 4.2. Plots of standardized mean difference (SMD) in the confounders between latent classes of puberty using ACE (total population), ACE_1 (early puberty), ACE_2 (mid puberty), and ACE_3(late puberty) weights from PLCA (Each of colored lines represents a confounder's line connecting two SMDs between treatments before (left side points) and after (right side points) using ACE, ACE_1, ACE_2, and ACE_3 weights).

Table 4.3. Model selection for adolescents substance use

약물복용	G^2	df	AIC	BIC	CAIC	aBIC
2개 범주 모형	98.21	36	120.21	184.34	195.34	149.39
3개 범주 모형	30.53	29	66.53	171.49	189.49	114.29
4개 범주 모형	14.66	22	64.66	210.43	235.43	131.00
5개 범주 모형	2.27	15	66.27	252.86	284.86	151.18
6개 범주 모형	1.30	8	79.30	306.70	345.70	182.79

AIC = Akaike information criterion; BIC = Bayesian information criterion; CAIC = consistent Akaike information criterion; aBIC = adjusted Bayesian information criterion.

Table 4.4. Item-response probabilities for adolescents substance use

약물복용 범주	약물복용 관련 문항			
	음주	흡연	음주 5+잔	취함
약물 미복용	0.148	0.048	0.000	0.000
흡연	0.372	0.999	0.177	0.240
일반적 약물복용	0.693	1.000	0.896	0.921

Table 4.5. ACE and ACE_t estimates for puberty on substance use

	성장 범주	약물복용 범주	
		흡연	일반적 약물복용
ACE	조숙 청소년	1.831	1.651
	보통성숙 청소년	1.527	1.420
ACE_1(조숙 청소년)	조숙 청소년	2.137	2.619
	보통성숙 청소년	1.560	1.508
ACE_2(보통성숙 청소년)	조숙 청소년	4.568	3.071
	보통성숙 청소년	1.407	1.552
ACE_3(만숙 청소년)	조숙 청소년	2.088	2.212
	보통성숙 청소년	1.562	1.497

ACE = average causal effect; ACE_t = ACE on the treated.

Table 4.5는 청소년 신체성숙 정도에 따른 청소년 약물복용의 평균인과효과를 추정한 결과이다. Table 4.5를 살펴보면, 만숙 청소년과 비교하여 성장이 빠를수록 약물복용의 오즈가 커지는 것을 확인할 수 있다. 예를 들어, 전체 모집단에 대한 ACE를 살펴보면 만숙 청소년에 비해 조숙 청소년과 보통성장 청소년이 흡연을 할 오즈는 각각 1.8배, 1.5배이며 모든 약물을 복용할 오즈는 1.7배, 1.4배에 이른다. 이처럼 성장이 빨리 이루어질수록 약물 미복용 그룹에 대한 약물 복용의 오즈가 점점 커짐을 알 수 있으며 이러한 현상은 모든 잠재처치범주의 도메인에 나타난다. 특히, 이러한 경향은 보통성숙 청소년을 모집단으로 할 경우, 조숙 청소년에서 두드러지게 나타난다. 조숙 청소년이 흡연을 할 오즈와 모든 약물을 복용할 오즈는 각각 만숙 청소년의 4.6배, 3.1배가 됨을 알 수 있다. 즉, 보통성숙 청소년이 신체성숙시기가 빠른 경우 다른 청소년에 비해 약물 복용에 대한 위험성이 증가함을 의미한다.

5. 결론

기존의 연구에서 처치변수와 결과변수가 가치 태도와 같은 관측할 수 없는 잠재변수일 경우, LCA를 활용하여 인과효과를 추론하는 방법을 제안하였다. 그러나 이러한 연구들은 예제를 기반으로 하여 참값을 알 수 없어 제안된 방법의 성능을 파악하기 어렵다는 한계가 있다. 따라서 본 논문에서 모의실험연구를

통해 기존의 제안된 방법들의 성능을 평가하였다. 특히, 잠재처치변수인 상황에서 PLogit과 PLCA의 성능을 알아보았다. PLogit과 PLCA는 성향점수를 추정하는 방법에서 차이가 있다. PLogit은 LCA를 통해 처치를 할당하여 로지스틱 회귀분석을 적합하여 성향점수를 추정하며, PLCA는 공변량을 고려한 LCA를 수행한 결과로 성향점수를 추정한다.

처치변수가 잠재변수인 상황에서 두 방법의 ACE와 ACE_t 추정치를 비교한 결과를 살펴보면, 결과변수의 종류에 관계없이 가중치를 적용하기 전 ACE와 ACE_t 추정치에 비해 두 방법 모두 편향이 감소하였다. PLogit과 PLCA의 성능을 비교하였을 때, PLogit에 비해 PLCA가 잠재결과변수에서 ACE_t 추정치를 제외하고 모든 경우에 상대적 편향의 평균이 작았으며, 100개의 데이터에 대한 참값과 추정치를 비교한 그림을 살펴보았을 때 PLCA가 편향을 제어하는 것으로 나타났다. 즉, PLCA는 PLogit에 비해 훨씬 안정적인 추정치를 제공하며 추정방법의 절차도 간단하다. 따라서 PLogit보다 PLCA를 사용하는 것이 더 적절하다.

모의실험 연구를 통해 ACE 추정방법의 성능을 파악한 후, PLCA를 적용하여 사례분석을 시행하였다. Add Health 자료를 이용하여 청소년 신체성숙도와 약물복용의 인과효과를 추정하였다. 청소년 성숙도와 관련된 문항은 가슴발달 정도와 체형변화 정도이며 첫 번째 조사와 두 번째 조사에서 각각 측정되었다. 약물복용과 관련된 문항은 네 개의 문항으로 이루어져 있으며 세 개의 문항은 술과 관련되어 있고 하나의 문항은 담배와 관련된 문항이다. LCA 수행결과 신체성숙도의 경우 세 개의 잠재처치범주를 가지며, 조숙, 보통성숙, 만숙 청소년으로 분류하였다. 약물복용의 경우 세 개의 잠재결과범주를 가지며, 약물 미복용, 흡연, 일반적 약물복용 그룹으로 분류하였다. 청소년 신체성숙도와 약물복용의 ACE 및 ACE_t 추정 결과, 성숙이 빠른 시기에 이루어질수록 약물복용의 위험성이 높은 것으로 나타났다.

본 논문에서 처치변수가 잠재변수인 상황에서 ACE를 추정하는 방법인 PLCA를 새롭게 제안하였다. 모의실험 연구를 통해 잠재처치변수인 경우 기존방법이 불안정한 성능을 가지는 반면 PLCA는 보다 안정적인 성능을 가지는 것을 확인하였다. 또한, 다항범주형 잠재변수를 포함하는 다양한 상황에서 ACE 추정방법을 제시하고 있어 잠재변수를 이용하여 인과효과를 추론하고자 하는 많은 연구에서 범용적인 활용이 가능할 것이다.

PLCA가 기존의 PLogit보다 우수하지만 PLCA에서도 처치를 할당하게 되며 이로 인해 정보의 손실이 발생한다. 이를 보완하기 위해 잠재범주를 할당하지 않도록 처치와 결과를 하나의 모형으로 설정하여 인과효과를 추정하는 방법이 연구되어야 할 것이다.

References

Clogg, C. C. and Goodman, L. A. (1984). Latent structure analysis of a set of multidimensional contingency tables, *Journal of the American Statistical Association*, **79**, 762–771.

Cohen, J. (1988). *Statistical Power Analysis for the Behavior Science*, Lawrance Eribaum Association.

Dayton, C. M. and Macready, G. B. (1988). Concomitant-variable latent-class models, *Journal of the American Statistical Association*, **83**, 173–178.

Frölich, M. (2004). Programme evaluation with multiple treatments, *Journal of Economic Surveys*, **18**, 181–224.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models, *Biometrika*, **61**, 215–231.

Lanza, S. T., Coffman, D. L., and Xu, S. (2013). Causal inference in latent class analysis, *Structural Equation Modeling: A Multidisciplinary Journal*, **20**, 361–383.

Lazarsfeld, P. and Henry, N. (1968). *Latent Structure Analysis*, Houghton, Mifflin, New York.

McCaffrey, D. F., Griffn, 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**, 3388–3414.

Park, G. and Chung, H. (2014). Estimating average causal effect in latent class analysis, *Korean Journal of Applied Statistics*, **27**, 1077–1095.

Robins, J. M., Hernan, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology, *Epidemiology*, **11**, 550–560.

Rosenbaum, P. R. (2002). Observational studies. In *Observational Studies* (pp. 1–17), Springer.

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

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

Rubin, D. B. (1976). Inference and missing data, *Biometrika*, **63**, 581–592.

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate, *Journal of Educational and Behavioral Statistics*, **2**, 1–26.

Rubin, D. B. (1978). Bayesian inference for causal effects: the role of randomization, *The Annals of Statistics*, **6**, 34–58.

Udry, J. R. (2003). *The National Longitudinal Study of Adolescent Health (Add Health)*, Wave I, 1994–1995.

잠재범주회귀모형의 성향점수를 이용한 잠재변수의 원인적 영향력 추론 연구

이미솔[a] · 정환[a,1]

[a]고려대학교 통계학과

(2017년 7월 5일 접수, 2017년 8월 22일 수정, 2017년 8월 23일 채택)

요 약

무작위 통제시험에서와 달리, 관찰연구에서는 편향되지 않은 인과관계를 추론하기 위한 통계적 전략이 필요하다. 최근 잠재범주분석(latent class analysis; LCA)에서 처치의 평균인과효과(average causal effect; ACE)를 추정하기 위한 새로운 방법들이 제안되었으나 이러한 방법들은 실제 데이터를 분석하는 응용 연구에 초점이 맞춰있다. 따라서 ACE의 참값을 알 수 없어 추정 방법의 성능을 평가하는 데 한계가 있다. 본 연구에서는 Park과 Chung (2014)이 제안한 방법을 개선하여, 다항범주형 처치변수가 잠재변수인 상황에서 다항범주형 결과변수에 미치는 인과효과 추정방법을 제안하고 처치변수와 결과변수가 잠재변수 또는 관측변수를 포함하는 여러 상황에서 본 연구가 제안한 인과효과 추정방법의 성능을 모의실험연구를 통하여 평가하고자 한다. 더불어 'National Longitudinal Study of Adolescents Health'자료를 사용하여 미국 여성 청소년 성장과 약물사용에 대한 인과효과를 추론하고자 한다.

주요용어: 성향점수, 잠재범주분석, 평균인과효과.

이 논문은 2017년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(No. 2015R1D1A1A01056846)이며, 제 1저자 이미솔의 석사논문 축약본임.

[1]교신저자: (02841) 서울특별시 성북구 안암로 145, 고려대학교 통계학과. E-mail: hwanch@korea.ac.kr