The Korean Journal of Applied Statistics (2014) DOI: http://dx.doi.org/10.5351/KJAS.2014.27.7.1077
27(7), 1077–1095

Estimating Average Causal Effect in Latent Class Analysis

Gayoung Park[a] · Hwan Chung[a,1]

[a]Department of Statistics, Korea University

(Received August 25, 2014; Revised November 10, 2014; Accepted December 08, 2014)

Abstract

Unlike randomized trial, statistical strategies for inferring the unbiased causal relationship are required in the observational studies. Recently, new methods for the causal inference in the observational studies have been proposed such as the matching with the propensity score or the inverse probability treatment weighting. They have focused on how to control the confounders and how to evaluate the effect of the treatment on the result variable. However, these conventional methods are valid only when the treatment variable is categorical and both of the treatment and the result variables are directly observable. Research on the causal inference can be challenging in part because it may not be possible to directly observe the treatment and/or the result variable. To address this difficulty, we propose a method for estimating the average causal effect when both of the treatment and the result variables are latent. The latent class analysis has been applied to calculate the propensity score for the latent treatment variable in order to estimate the causal effect on the latent result variable. In this work, we investigate the causal effect of adolescents delinquency on their substance use using data from the 'National Longitudinal Study of Adolescent Health.'

Keywords: Average causal effect, causal inference, latent class analysis, propensity score.

1. 서론

관찰연구(observarional study)에서 인과관계를 추론할 경우 무작위 통제시험(randomized controlled trials, RCT)과는 달리 교란변수(confounder)로 인한 편향을 제어하기 위한 통계적 전략이 필요하다. 예를 들면, 짝짓기(matching)를 통한 교란변수의 보정 또는 회귀분석에 원인변수와 함께 교란변수를 공변수로 포함하는 공분산 분석 등을 사용할 수 있다. 최근에는 원인변수의 역확률을 가중치로 사용(inverse probability treatment weighting, IPTW)하는 주변구조모형(Robins 등, 2000)이나 성향점수(propensity score) 를 이용한 짝짓기(Rosenbaum, 2002)등이 제안되어 사용되고 있다. 이러한 인과관계 추론은 처치(treatment)가 명확히 주어진 경우에 교란변수를 통제하고 그 처치가 관측된 결과(outcome)에 미치는 영향을 평가하는 방법에 초점이 맞추어져 있다.

하지만 많은 관찰연구에서 결괏값 뿐만 아니라 그에 영향을 미치는 처치변수 또한 하나의 관측변수로 명확히 측정되기 어려운 경우가 있다. 예를 들어, 연구자가 처치변수인 청소년기의 비행(delinquency)정

This study was supported by a Korea University Grant (G1300030).
[1]Corresponding author: Associate Professor, Department of Statistics, Korea University, 145 Anam-Ro, Seongbuk-Gu, Seoul 136-701, Korea. E-mail: hwanch@korea.ac.kr

도가 약물사용(substance use)에 미치는 원인적 영향력(causal effect)에 관해 연구한다고 하자. 처치변수인 청소년의 비행행태는 부모님께 거짓말을 하는것 부터 상해, 절도 및 폭력 등 다양한 유형으로 나타나며 하나의 관측변수로 명확히 측정하기 어렵다. 따라서 청소년의 비행에 관한 여러 측정변수를 이용해 유사한 비행행태를 보이는 청소년을 분류하여야 할 것이며 이때 분류된 집단이 처치변수가 될 것이다. 결과변수인 약물사용 또한 개인의 현재 약물사용 행태를 일련의 단계발전(stage-sequential development)의 한 부분으로 파악해야 한다. 예를 들어, 니코틴 중독은 최초의 담배 흡연, 불규칙적 흡연, 규칙적 흡연, 그리고 니코틴 중독을 포함하는 일련의 행동양식 과정을 통해 이루어진다고 알려져 있으며 현재의 흡연 행태는 이러한 일련의 단계발전 현상의 한 부분으로 파악해야 한다는 것이다(Flay, 1993; Leventhal과 Cleary, 1980; Mayhew 등, 2000). 이러한 이론의 기저에는 약물사용 행위에 따라 각 개인들을 유사집단에 분류할 수 있다는 것을 전제로 하고 있다. 이와 같이 청소년 비행 혹은 약물사용 등 하나의 관측변수로 명확히 관측하기 어려우나 이와 관련된 여러 측정변수를 이용해 개체들을 몇개의 유사범주로 분류하기 위한 방법으로 잠재범주모형(latent class analysis, LCA)이 많은 연구에서 사용되고 있다. LCA 모형은 각 항목의 반응에 따라 구성원의 일부를 동질 집단으로 분류하기 위한 가장 간단한 혼합모형(mixture model) 중의 하나이며 직접 관찰 되지 않는 분류자의 존재를 가정함으로써 각 항목의 연관관계를 설명하고자 하는 모형이다(Clogg과 Goodman, 1984; Goodman, 1974).

원인적 영향력이란 한 개체가 다른 처치를 받는 경우에 일어날 결과(potential outcomes)들의 차이에 관한 추론을 근거로 하고 있다 (Rubin, 1974). 하지만 루빈의 인과모형(Rubin's causal model; RCM)이라 일컬어지는 이러한 접근법은 처치변수와 결과변수가 직접관측이 가능한 경우에만 사용할 수 있는 한계를 갖고 있다. 따라서 본 연구에서는 관찰연구에서 직접적으로 측정될 수 없는 처치변수와 결과변수 모두를 측정오차를 고려한 LCA 모형의 변수로 모형화함으로써 잠재범주 간의 원인적 영향력을 추정하는 방법을 RCM 의 틀 안에서 제시하고자 한다.

본 논문은 다음과 같이 구성되어 있다. 2.1절에서는 RCM과 교란변수로 인한 편향을 제어하기 위하여 사용되는 성향점수에 대해 설명하고, 2.2 절에서는 잠재처치변수와 잠재결과변수 추론을 위한 LCA 모형을 제시한다. 2.3절에서는 잠재결과변수에 대한 잠재처치변수의 원인적 영향력을 추정하는 방법을 제안하고, 3장에서는 2.3절에서 제안한 절차에 따라 약물사용에 대한 청소년기 비행의 원인적 영향력을 추론하고 이에 대한 적절한 해석을 제시할 것이다. 이때 자료는 미국의 The National Longitudinal Study of Adolescent Health 자료를 이용하며 4장에서는 본 논문의 결론 및 연구의 한계점을 기술할 것이다.

2. 잠재적 다항범주자료 분석을 위한 원인적 영향력 추론

2.1. 다항범주자료 분석을 위한 루빈의 인과모형(RCM) 및 성향점수

대부분의 관찰연구에서는 여러 처치의 결과를 동시에 관찰할 수 없으나 적절한 가정을 사용하여 현실에서 관찰된 값만으로 처치변수의 원인적 영향력을 추정할 수 있는 것이 RCM의 장점이라 할 수 있다. RCM의 구동원리는 주어진 결과의 원인을 직접 밝히는 것이 아니라 추정되는 원인의 영향을 가실험설계(pseudo experimental design)에 근거하여 밝히는 것이다. RCM의 쉬운 설명을 위해 다항범주형 처치변수와 다항범주형 결과변수를 사용하여 n명의 개체 자료를 표 2.1과 같이 나타내었다.

표 2.1에서 T_i는 i번째 개체가 실제로 받은 처치를 나타낸다. 예를 들어, 앞의 예와 같이 연구자가 처치변수인 청소년기 비행정도에 따라 약물사용의 정도가 다른지에 관해 연구한다고 하자. 이때 처치변수는 청소년기의 비행정도를 나타내며 다음과 같은 네 가지의 범주를 갖는다고 하자(미 비행청소년: $T_i = 1$, 언어적 비행청소년: $T_i = 2$, 절도형 비행청소년: $T_i = 3$, 전반적 비행청소년: $T_i = 4$). 이 경우 결과

Table 2.1. Data with multinomial treatment and response variables

개체*	교란변수	T_i	$Y_i(1)$	$Y_i(2)$	$Y_i(3)$	$Y_i(4)$
$i=1$	X_1	1	$Y_1(1)$	$Y_1(2)$	$Y_1(3)$	$Y_1(4)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$i=n_1$	X_{n_1}	1	$Y_{n_1}(1)$	$Y_{n_1}(2)$	$Y_{n_1}(3)$	$Y_{n_1}(4)$
$i=n_1+1$	X_{n_1+1}	2	$Y_{n_1+1}(1)$	$Y_{n_1+1}(2)$	$Y_{n_1+1}(3)$	$Y_{n_1+1}(4)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$i=n_2$	X_{n_2}	2	$Y_{n_2}(1)$	$Y_{n_2}(2)$	$Y_{n_2}(3)$	$Y_{n_2}(4)$
$i=n_2+1$	X_{n_2+1}	3	$Y_{n_2+1}(1)$	$Y_{n_2+1}(1)$	$Y_{n_2+1}(3)$	$Y_{n_2+1}(4)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$i=n_3$	X_{n_3}	3	$Y_{n_3}(1)$	$Y_{n_3}(2)$	$Y_{n_3}(3)$	$Y_{n_3}(4)$
$i=n_3+1$	X_{n_3+1}	4	$Y_{n_3+1}(1)$	$Y_{n_3+1}(2)$	$Y_{n_3+1}(3)$	$Y_{n_3+1}(4)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$i=n$	X_n	4	$Y_n(1)$	$Y_n(2)$	$Y_n(3)$	$Y_n(4)$

*$n_1 < n_2 < n_3 < n$

변수 $Y_i(t)$는 t번째 비행범주에 속한 i번째 청소년의 약물사용의 정도이다. 약물사용의 정도 역시 다음의 세 가지 범주를 갖는 다항범주 변수라 하자 (미사용자: $Y_i(t)=1$, 담배와 주류사용자: $Y_i(t)=2$, 모든 약물사용자: $Y_i(t)=3$). 이때, 만일 $T_i=1$ 이면 $Y_i(1)$ 은 관측되나 $Y_i(2)$, $Y_i(3)$ 및 $Y_i(4)$는 결측임을 알 수 있다. 따라서 어떤 개체에서도 모든 $t=2,3,4$에 대해 $Y_i(t)$와 $Y_i(1)$은 동시에 관측될 수 없으며 이러한 인과추론의 문제는 결측자료의 문제와 같게 된다. 무작위 통제시험 설계가 아닌 경우에는 이러한 문제를 해결하기 위하여 처치와도 관련이 있고 결과변수와도 관련이 있는 제3의 변수, 즉 교란변수를 고려하여야 한다. 만약 교란변수를 고려하지 않고 분석을 하게 되는 경우는 원인적 영향력의 추정치에 편향이 발생하게 된다. 원인적 영향력은 직접적으로 관찰되지 않기 때문에 통계적 인과관계 추론에서는 모집단에서의 평균 원인적 영향력(average causal effect; ACE)을 추정하는 방법을 이용한다. 만약 처치변수가 C개의 범주($t=1,\ldots,C$), 결과변수가 S개의 범주($y=1,\ldots,S$)를 가지며 결과변수의 첫 번째 범주($y=1$)를 기준범주(baseline category)로 사용한다면 처치범주가 t일때 결과범주 y의 오즈는 다음과 같이 정의된다.

$$Odds = \frac{P\left[Y_i(t)=y\right]}{P\left[Y_i(t)=1\right]}, \quad t=1,\ldots C, \;\; y=2,\ldots,S.$$

즉, 이때의 $Odds$는 모집단의 모든 개체가 처치변수 중 한 범주($t=1,\ldots,C$)에 속했을 때 결과변수의 기준범주($y=1$)의 확률과 다른 범주($y=2,\ldots,S$)의 확률의 비를 나타낸다. 처치변수의 첫 번째 범주($t=1$)를 기준범주로 사용한다면 이와 관련된 오즈비는 다음과 같이 구할 수 있으며 이를 ACE로 정의한다.

$$ACE = \frac{P\left[Y_i(t)=y\right]/P\left[Y_i(t)=1\right]}{P\left[Y_i(1)=y\right]/P\left[Y_i(1)=1\right]}, \quad t=2,\ldots,C, \;\; y=2,\ldots,S.$$

또한, C개 범주의 처치변수와 S개 범주의 결과변수가 있는 경우 모집단에 속하는 개체 중 하나의 처치범주에 속하는 개체들에 대해 ACE 를 추정할 수 있는데 이를 $ACE_t(t=1,\ldots,C)$라 하겠다. 예를 들어, 앞의 청소년 비행과 약물사용의 예의 경우, ACE_2의 경우 언어적 비행범주($t=2$)에 속한 청소년이 비행범주 중 한 범주($t=2,3,4$)에 속했을 때와 이들이 비행청소년이 아닌 기준범주($t=1$)에 속했을 때를 비교하여 약물사용 오즈에 대한 차이를 보는 것을 의미한다.

ACE 및 ACE_t를 추정하기 위해서는 교란변수의 벡터 $\mathbf{x}_i$가 주어진 경우 T_i는 $Y_i(t)$와 $Y_i(1)$이 모든 $t = 2, \ldots, C-1$에 대해 독립이라는 조건부 독립의 가정이 필요하며(Rosenbaum과 Rubin, 1983) 이런 가정을 만족하는 경우, 결측된 결과변수는 무작위 결측(missing at random; MAR)의 조건을 충족하게 된다(Rubin, 1976). Rosenbaum과 Rubin (1983)은 성향점수라는 방법을 이용하여 ACE를 추정하는 방법을 제시하였다. 성향점수는 i번째 개체의 교란변수의 벡터 $\mathbf{x}_i$가 주어진 경우 개체 i가 특정한 처치를 받는 확률로 $t = 1, \ldots, C$에 대해 $\pi_{it}(\mathbf{x}_i) = P(T_i = t \mid \mathbf{x}_i)$로 정의되며 이 성향점수는 각 처치에 할당될 성향을 의미한다. 무작위 통제 시험에서는 이 성향점수는 교란변수 $\mathbf{x}_i$에 영향을 받지 않으므로 $\pi_{it} = P(T_i = t)$이 된다. 성향점수는 π_{it}가 상수인 부집단(sub-poputlation)에서 $t = 1, \ldots, C$인 모든 범주에서 동일한 공변량 $\mathbf{x}_i$의 분포를 가지므로 성향점수는 모든 다항범주 사이의 교란변수를 균형있게 만들어 준다고 할 수 있다. 이러한 성향점수의 특성을 이용하면 짝짓기, 층화 및 처치변수의 역확률을 가중치로 사용하는 주변구조모형 등을 이용하여 ACE, ACE_t 및 이와 관련된 오즈비를 추정할 수 있다(Rosenbaum, 2002; Robins 등, 2000).

성향점수는 알려진 값이 아니므로 모형을 통하여 추정하여야 하는데 이 연구에서는 처치변수가 다항범주 자료이므로 로지스틱 회귀분석을 사용하여 추정할 수 있다. 모든 개체가 C개의 처치범주 중 하나의 범주를 부여받은 경우, i번째 개체의 알려진 처치 t에 대한 기준범주 로지스틱 회귀 모형을 통해 추정된 식 (2.1)의 $\hat{\pi}_{it}(\mathbf{x}_i)$을 성향점수로써 정의할 수 있고, 이때 처치에 대한 회귀 모형의 설명변수는 교란변수인 $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$가 된다.

$$\hat{\pi}_{it}(\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i{}'\hat{\boldsymbol{\alpha}}_t)}{\sum_{c=1}^{C} \exp(\mathbf{x}_i{}'\hat{\boldsymbol{\alpha}}_c)}. \tag{2.1}$$

이때, 식 (2.1)의 추정계수는 처치범주 $t = 2, \ldots, C$에 대해 $\hat{\boldsymbol{\alpha}}_t = (\hat{\alpha}_{1t}, \ldots, \hat{\alpha}_{pt})'$로 정의되며 처치변수의 기준범주인 $t = 1$ 의 계수는 항상 $\hat{\boldsymbol{\alpha}}_1 = \mathbf{0}$을 만족한다.

이와 같이 원인적 영향력 추론을 위해 각 처치범주에 속할 성향점수가 추정되면, 각 범주간 성향점수의 겹침(overlap)을 확인해야 한다. 만약 처치범주 간 성향점수의 분포가 겹치지 않는다면 각 범주에 속하는 개체들이 비슷하지 않다는 것을 의미하므로 성향점수를 사용하여 각 범주에 속하는 개체들을 무작위로 할당받은 것처럼 보이도록 조정하는 것이 불가능하다. 따라서 결과변수에 대한 처리변수의 원인적 영향력 추론을 할 수 없을 것이다. 처치범주 간 성향점수의 분포가 겹치는 것을 확인할 수 있다면, 교란변수로 인한 편향을 제어하기 위해 성향점수의 역수를 가중치로 주는 기법을 사용하여야 한다. 성향점수의 역수를 가중치로 주는 방법은 특정 처치범주에 속할 가능성이 높은(성향점수가 높은) 개체들의 영향을 줄여주고, 반대로 특정 처치범주에 속할 가능성이 낮은(성향점수가 낮은) 개체들의 영향은 높여준다는 점에서 일반적인 표본조사에서 가중치 사용하는 것과 의미가 비슷하다. 이와 같이 성향점수를 이용한 가중치 조정을 통해 비슷한 성향점수를 갖는 개체들은 어떤 범주에 속하는지에 관계없이 이들이 갖는 교란변수의 측정값을 유사하게 만들 수 있을 것이다. 성향점수를 이용한 가중치를 준 후, 처치범주 간 교란변수들 분포의 균형이 이루어졌는지를 평가해야 하며 이는 각각의 처치범주 간 교란변수의 가중치를 주기 전과 후의 표준화된 평균 차이(standardized mean difference; SMD)를 계산함으로써 평가할 수 있다. 일반적으로 SMD의 절댓값이 0.2를 넘지 않으면 균형이 이루어졌다고 평가한다(Cohen, 1988). 만약 균형이 이루어지지 않으면 성향점수에 대한 모형에 교호작용 항을 추가하거나, 고차 항을 추가하는 등의 방법을 통해서 균형을 맞춰야 한다. 앞의 청소년 비행과 약물사용 예와 같이 처치변수가 네 개의 범주를 갖는 경우, 모집단 전체에 대한 평균 원인적 영향력(ACE)과 처치범주 t에 속하는 개체들에 대한 평균 원인적 영향력(ACE_t, $t = 1, 2, 3, 4$)에 부여되는 가중치는 표 2.2와 같다.

Table 2.2. Weights for ACE estimation

	처치범주 1	처치범주 2	처치범주 3	처치범주 4
ACE	$1/\hat{\pi}_{i1}(\mathbf{x}_i)$	$1/\hat{\pi}_{i2}(\mathbf{x}_i)$	$1/\hat{\pi}_{i3}(\mathbf{x}_i)$	$1/\hat{\pi}_{i4}(\mathbf{x}_i)$
ACE_1	1	$\hat{\pi}_{i1}(\mathbf{x}_i)/\hat{\pi}_{i2}(\mathbf{x}_i)$	$\hat{\pi}_{i1}(\mathbf{x}_i)/\hat{\pi}_{i3}(\mathbf{x}_i)$	$\hat{\pi}_{i1}(\mathbf{x}_i)/\hat{\pi}_{i4}(\mathbf{x}_i)$
ACE_2	$\hat{\pi}_{i2}(\mathbf{x}_i)/\hat{\pi}_{i1}(\mathbf{x}_i)$	1	$\hat{\pi}_{i2}(\mathbf{x}_i)/\hat{\pi}_{i3}(\mathbf{x}_i)$	$\hat{\pi}_{i2}(\mathbf{x}_i)/\hat{\pi}_{i4}(\mathbf{x}_i)$
ACE_3	$\hat{\pi}_{i3}(\mathbf{x}_i)/\hat{\pi}_{i1}(\mathbf{x}_i)$	$\hat{\pi}_{i3}(\mathbf{x}_i)/\hat{\pi}_{i2}(\mathbf{x}_i)$	1	$\hat{\pi}_{i3}(\mathbf{x}_i)/\hat{\pi}_{i4}(\mathbf{x}_i)$
ACE_4	$\hat{\pi}_{i4}(\mathbf{x}_i)/\hat{\pi}_{i1}(\mathbf{x}_i)$	$\hat{\pi}_{i4}(\mathbf{x}_i)/\hat{\pi}_{i2}(\mathbf{x}_i)$	$\hat{\pi}_{i4}(\mathbf{x}_i)/\hat{\pi}_{i3}(\mathbf{x}_i)$	1

2.2. 잠재처치변수 및 잠재결과변수를 위한 잠재범주모형(LCA)

많은 관찰연구에서 결괏값 뿐만 아니라 그에 영향을 미치는 처치변수 또한 하나의 관측변수로 명확히 측정되기 어려운 경우가 있다. 앞의 예의 경우, 처치변수인 청소년의 비행정도는 유사한 비행행태를 보이는 몇개의 유사집단으로 분류하여야 한다. 하지만 이를 하나의 관측변수로는 명확히 측정하기 어려워 청소년의 비행에 관련된 여러 측정변수를 이용해 청소년들을 그들의 비행행태에 따라 몇개의 범주로 분류 하여야 할 것이다. 결과변수인 약물사용의 경우 또한 하나의 관측변수로는 약물사용 정도에 따른 유사집단으로 분류할 수 없으므로 약물사용과 관련된 여러 측정변수를 이용하여 분류하여야 한다. 본 연구에서는 이러한 자료의 분석을 위해 각 항목의 반응에 따라 구성원의 일부를 동질 집단으로 분류하기 위한 가장 간단한 혼합모형 중의 하나인 LCA 모형을 사용하고자 한다.

LCA 모형은 관측 가능한 범주형 변수들 사이의 관계를 직접 관측할 수 없는 몇 개의 잠재범주로 설명하고자 하는 분석 방법이다. 만약 우리가 C개의 잠재범주를 가지고 있는 처치변수를 측정하고자 M개의 다항문항을 사용한 LCA 모형을 구축한다고 가정하자. T는 C개의 잠재범주를 갖는 잠재처치변수이며 M개의 다항문항 $\mathbf{U} = (U_1, \ldots, U_M)$은 잠재처치변수 T를 측정하고자 사용된 처치관측변수이다. 처치관측변수 $\mathbf{U}$의 i번째 개체의 관측값을 $\mathbf{u}_i = (u_{i1}, \ldots, u_{iM})$이라고 할때 m번째 문항의 i번째 관측치 u_{im}의 가능한 응답범주는 $u_{im} = 1, \ldots, r_m$이라고 하자. 이때, 잠재처치변수 T와 처치관측변수 $\mathbf{U}$의 결합밀도함수는 다음과 같이 주어진다.

$$\begin{aligned} P(T = t, \mathbf{U} = \mathbf{u}_i) &= P(T = t)P(\mathbf{U} = \mathbf{u}_i \mid T = t) \\ &= P(T = t) \prod_{m=1}^{M} P(U_m = u_{im} \mid T = t) \\ &= \gamma_t \prod_{m=1}^{M} \prod_{k=1}^{r_m} \rho_{mk|t}^{I(u_{im}=k)}. \end{aligned} \tag{2.2}$$

식 (2.2)에서 주어진 결합밀도함수는 잠재처치범주의 출현율 $\gamma_t = P(T = t)$과 문항응답확률(item-response probability)인 $\rho_{mk|t} = P(U_m = k \mid T = t)$ 항들의 곱으로 나타낼 수 있다. 문답응답확률인 $\rho_{mk|t}$는 잠재처치범주 t에 속해있는 개체가 m 번째 처치관측 문항에 대해 k번째 범주로 답할 확률을 의미하며, 식 (2.2)에서는 잠재처치범주 t가 알려지면 처치관측변수는 독립이라는 지역독립성(local independence)의 가정을 하고 있음을 알 수 있다.

만약 모든 n개체의 잠재처치범주가 관측가능하다면, 식 (2.2)의 모수는 다항범주모형을 이용하여 쉽게 추정할 수 있으나 모든 개체의 잠재처치범주는 관측가능하지 않으므로 i번째 개체의 처치변수에 관한

LCA 모형의 우도함수는 식 (2.3)과 같이 주어진다.

$$\begin{aligned} L_{i(treatment)} &= P(\mathbf{U}=\mathbf{u}_i) \\ &= \sum_{t=1}^{C} P(T=t, \mathbf{U}=\mathbf{u}_i) \\ &= \sum_{t=1}^{C} \gamma_t \prod_{m=1}^{M} \prod_{k=1}^{r_m} \rho_{mk|t}^{I(u_{im}=k)}. \end{aligned} \tag{2.3}$$

식 (2.3)에서 추론한 잠재처치범주는 다음의 식 (2.4)의 사후확률(posterior probability)에 따라 각 개체의 관측처치변수 문항반응형태에 따라 무작위로 모든 개체에 대해 잠재처리범주를 부여함으로써 구체화 할 수 있다.

$$\begin{aligned} \theta_{it} &= P(T=t \mid \mathbf{U}=\mathbf{u}_i) \\ &= \frac{\gamma_t \prod_{m=1}^{M} \prod_{k=1}^{r_m} \rho_{mk|t}^{I(u_{im}=k)}}{\sum_{t=1}^{C} \gamma_t \prod_{m=1}^{M} \prod_{k=1}^{r_m} \rho_{mk|t}^{I(u_{im}=k)}}. \end{aligned} \tag{2.4}$$

처치변수와 마찬가지로 결과변수 또한 직접관측이 어려운 잠재범주변수이다. 만약 우리가 C'개의 잠재범주를 가지고 있는 결과변수를 측정하고자 M'개의 다항문항을 사용한 LCA 모형을 구축한다고 가정하자. Y는 $y = 1, \ldots, C'$개의 잠재범주를 갖는 잠재결과변수이고 M' 개의 다항문항 $\mathbf{W} = (W_1, \ldots, W_{M'})$은 잠재결과변수 Y를 측정하고자 사용된 결과관측변수이다. 이때, 결과관측변수 $\mathbf{Y}$ 의 i번째 개체의 관측값을 $\mathbf{w}_i = (w_{i1}, \ldots, w_{iM'})$ 이라고 할때 m번째 문항의 i번째 관측치 w_{im}의 가능한 응답범주는 $w_{im} = 1, \ldots, r'_m$이라고 하자. 또한, 위의 잠재처치변수 분석을 위해 사용된 LCA 모형에서 추정된 i번째 개체의 잠재처치변수를 t_i, $t_i = 1, \ldots, C$라고 하자. 이때, t_i는 식 (2.4)의 확률로 각 개체에게 부여된 잠재처치범주이다. 잠재처치변수 t_i 가 주어졌을 때 잠재결과변수 Y와 결과관측변수 $\mathbf{W}$의 결합밀도함수는 다음과 같이 주어진다.

$$\begin{aligned} P(Y=y, \mathbf{W}=\mathbf{w}_i \mid t_i) &= P(Y=y \mid t_i) P(\mathbf{W}=\mathbf{w}_i \mid Y=y) \\ &= P(Y=y \mid t_i) \prod_{m=1}^{M'} P(W_m = w_{im} \mid Y=y) \\ &= \gamma_y(t_i) \prod_{m=1}^{M'} \prod_{k=1}^{r'_m} \eta_{mk|y}^{I(w_{im}=k)}. \end{aligned} \tag{2.5}$$

식 (2.2)와 같이 식 (2.5)에서도 결합밀도함수는 잠재결과범주의 출현율 $\gamma_y(t_i) = P(Y = y \mid t_i)$와 문항응답확률인 $\eta_{mk|y} = P(W_m = k \mid Y = y)$의 곱으로 나타낼 수 있으며 잠재결과범주 y가 주어지면 결과관측변수는 독립이라는 지역독립성의 가정을 하고 있음을 알 수 있다. 문답응답확률인 $\eta_{mk|y}$는 잠재결과범주 y에 속해있는 개체가 m번째 결과관측변수 문항에 대해 k번째 범주로 답할 확률을 의미한다. 또한, 로짓모형(Agresti, 2002)을 이용하여 다음과 같이 식 (2.6)에 의해 i 번째 개체가 잠재결과범주 y에 속할 확률을 추정할 수 있다(Dayton과 Macready, 1988).

$$\gamma_y(t_i) = \gamma_y(\mathbf{z}_i) = \frac{\exp(\mathbf{z}_i{}'\boldsymbol{\beta}_y)}{\sum_{c=1}^{S} \exp(\mathbf{z}_i{}'\boldsymbol{\beta}_c)}. \tag{2.6}$$

식 (2.6)의 $\mathbf{z}_i$는 i번째 개체의 잠재처리범주를 나타내는 더미변수(dummy variable)로서 길이가 $C-1$인 벡터이며 추정계수는 잠재결과범주 $y = 1, \ldots, S-1$에 대해 $\boldsymbol{\beta}_y = (\beta_{0y}, \beta_{1y}, \ldots, \beta_{S-1,y})'$로 정의되며 결과변수의 기준범주인 S 의 계수는 항상 $\boldsymbol{\beta}_S = \mathbf{0}$을 만족한다.

만약 모든 n개체의 잠재결과변수가 관측가능하다면, 식 (2.5)의 모수는 로짓모형과 다항범주모형을 이용하여 쉽게 추정할 수 있을 것이다. 그러나 모든 개체의 잠재결과변수는 관측가능하지 않으므로 i번째 개체의 결과변수에 관한 LCA 모형의 우도함수는 식 (2.7)과 같이 주어진다.

$$\begin{aligned} L_{i(outcome)} &= P(\mathbf{W} = \mathbf{w}_i \mid \mathbf{z}_i) \\ &= \sum_{y=1}^{S} P(Y = y, \mathbf{W} = \mathbf{w}_i \mid \mathbf{z}_i) \\ &= \sum_{y=1}^{S} \gamma_y(\mathbf{z}_i) \prod_{m=1}^{M'} \prod_{k=1}^{r'_m} \eta_{mk|y}^{I(w_{im}=k)}. \end{aligned} \tag{2.7}$$

2.3. 잠재결과변수에 대한 잠재처치변수의 원인적 영향력 추론

이 절에서는 앞에서 기술한 바와 같이 처치변수와 결과변수가 모두 잠재범주의 자료형태를 갖는 경우 결과변수에 대한 처치변수의 ACE 및 $ACE_t(t = 1 \ldots, C)$의 추정을 위한 7단계의 분석 절차를 제안하고자 한다. 1단계와 2단계가 수행된 후 3단계의 사후확률에 따라 개체들을 임의로 하나의 처치범주에 할당하는 것을 N번 반복하며, 4단계부터 6단계까지는 3단계에서 할당된 N개의 처치범주 자료의 각각에 대해 과정을 수행한다. 이렇게 추정된 N개의 원인적 영향력을 종합하여 할당된 처치범주에 따른 약물사용에 대한 비행의 ACE 및 $ACE_t(t = 1 \ldots, C)$를 7단계에 의해 추정할 수 있다.

1. **변수선택**

   먼저 분석에 사용될 변수들을 선택해야 한다. 본 연구에서는 교란변수, 잠재처치범주를 측정하기 위한 처치관측변수 및 잠재결과범주를 측정하기 위한 결과관측변수 등을 선택한다.

2. **다중대체(multiple imputation)**

   결측 자료에 대해 다중대체와 같은 기법을 고려하지 않은 경우, 단지 하나의 교란변수에 대해서 결측값이 존재하는 개체에 대해서도 성향점수 추정치가 결측이 된다. 이러한 개체들은 그 이후의 분석 과정에서도 누락되기 때문에 표본수가 상당히 감소하게 되거나, 평균 원인적 영향력의 추정치가 편향될 수 있다. 그러므로 결측치에 대해 다중 대체를 실시한다. 이 방법의 장점은 변수들의 결측이 일반적인 대체 모형하에서 고려된다는 점이다. 결측 자료에 대해 다중대체를 실시함으로써 잠재범주모형을 이용한 원인적 영향력 추론이 완전한 표본에서 이루어지게 되어, 통계학적 검정력을 최대화할 수 있고 원인적 영향력 추정에 대한 편향을 줄일 수 있게 된다. 결측치는 MAR 가정을 하고 있으며 결측 모형은 다변량 정규분포를 사용하였고 결측치에 대한 사후예측확률 분포에 따라 결측치를 생성하였다. 사용된 알고리즘은 최대우도추정치의 값을 초기값으로 하여 데이터 확대기법(data augmentation)을 통해 매 20번 반복마다 한 번씩 대체하여 다섯개의 자료를 만들었다.

3. **처치에 대한 잠재범주분석 및 사후확률에 따른 잠재범주 할당**

   이 단계에서는 잠재처치변수를 측정하기 위해 선택된 처치관측변수를 이용하여 처치범주에 대한 LCA 식 (2.3)를 수행한다. 이때 잠재범주의 수를 결정하는 것이 매우 중요한데, LCA의 경우 일반적인 가능도비 검정을 통해서 잠재범주의 수를 결정할 수 없으나 AIC와 BIC를 이용하여 잠재범주의 수를 결정할 수 있다. 모형선택 과정을 통해 잠재처치범주의 수가 결정되면, 식 (2.4)에 주어진 사후확률의 추정치에 따라 각 개체에게 임의로 하나의 잠재처치범주를 할당해준다.

4. **성향점수 추정 및 겹침(overlap) 평가**

1단계에서 선택된 교란변수를 설명변수로 하여 3단계에서 할당된 잠재처치범주에 대한 로지스틱 회귀모형을 적합시켜 성향점수를 식 (2.1)과 같이 추정한다. 각 개체에 대한 성향점수가 추정되면 원인적 영향력 추론이 가능한지 확인하기 위해 겹침을 평가해야 한다. 성향점수의 역수를 가중치로 주는 방법을 적용하기 위해서는 처치범주 간 성향점수의 분포가 어느 정도 겹쳐야 하기 때문이다. 이를 위해 처치범주 간 성향점수의 상자 그림을 통해 분포를 비교한다.

5. **가중치 계산 및 균형(balance) 평가**
   교란변수로 인한 편향을 조정하기 위해 표 2.2와 같이 각 평균 원인적 영향력(ACE)에 대한 가중치를 계산한다. ACE를 추정하기 위한 가중치는 i번째 개체가 범주 t에 속하는 경우 $1/\hat{\pi}_{it}(\mathbf{x}_i)$을 부여한다. 또한 모든 $t = 1, \ldots, C$에 대한 ACE_t을 추정하기 위한 가중치는 범주 t에 속하는 개체의 경우 1을 부여하고, t가 아닌 범주(t')에 속하는 경우 $\hat{\pi}_{it}(\mathbf{x}_i)/\hat{\pi}_{it'}(\mathbf{x}_i)$을 부여함으로써 교란변수들의 특성이 범주 t에 속하는 개체들의 특성과 비슷해지도록 만들어 준다. 표 2.2와 같이 가중치가 계산되면 처치범주 간에 가중치를 주기 전과 후의 교란변수의 표준화된 평균차이를 계산하여 비교한다. 가중치를 준 후에 교란변수들의 잠재처치범주 간 표준화된 평균 차이의 절댓값이 0.2(Cohen, 1988)보다 작은 경우 균형이 잘 이루어졌다고 볼 수 있으며, 이는 가중치를 부여한 표본이 무작위 통제시험의 표본과 유사하게 되었다고 간주할 수 있다.
6. **가중치를 이용하여 결과에 대한 잠재범주분석**
   이 단계에서는 잠재결과변수를 측정하기 위해 1단계에서 선택된 결과관측변수를 이용하여 결과범주에 대한 LCA 모형을 수행한다. 잠재결과변수에 대한 LCA 모형을 수행함에 있어서 역시 잠재범주의 수를 결정하는 것이 중요한 문제이다. 3단계의 잠재처치변수의 모형선택과 동일한 과정을 거쳐 잠재결과범주의 수를 결정한 후, 5단계에서 계산된 가중치와 함께 3단계에서 각 개체에게 할당된 처치범주에 대한 더미변수를 공변량으로 고려한 LCA 모형 식 (2.7)을 수행한다. 이때 평균 원인적 영향력의 추정치는 식 (2.6)에서 주어진 회귀계수들의 추정치이고, 이에 대해 지수승을 취하여 오즈비를 계산함으로써 결과를 쉽게 해석할 수 있다.
7. **부여된 처치범주에 따른 N번의 결과 종합**
   3단계부터 6단계까지를 N번 반복하여 얻은 N개의 회귀계수와 표준오차를 루빈의 규칙(Rubin's rules)을 통해 종합한다(Rubin, 2004). 각 회귀계수의 전체적인 추정치는 평균으로 계산되고, 표준오차의 추정치는 할당된 처치범주 내에서의 분산과 처치범주 간 분산의 가중 합으로써 추정된다.

위의 7단계를 수행하기 위해 다음과 같은 통계 패키지가 사용되었다. 먼저 결측치에 대한 다중대체를 수행하기 위해 SAS 9.3 의 PROC MI 프로시저를 사용하였다. 그리고 처치변수와 결과변수에 대한 잠재범주분석을 수행하기 위해 PROC LCA(Lanza 등, 2007; Lanza 등, 2013)를 사용하여 LCA 모형을 적합시켰으며, 마지막으로 로지스틱 회귀모형을 통해 성향점수를 추정하기 위하여 SAS 9.3 PROC LOGISTIC을 사용하였다.

3. 약물사용에 대한 청소년기의 비행의 원인적 영향력 추론

3.1. 자료

청소년기의 비행에 대한 잠재범주가 처치변수인 경우 약물사용의 잠재범주에 대한 원인적 영향력을 추정하기 위해 미국의 청소년 위험 행동의 대표적인 조사인 The National Longitudinal Study of Adolescent Health(Udry, 2003) 의 자료를 사용하였다. 이 조사의 표본은 지역, 도시성, 학교규모, 학교유

Table 3.1. Manifest items for adolescent delinquency

항목	응답 범주	
거짓말	1=거짓말 한 적 없음	2=거짓말 한 적 있음
고성방가	1=해본 적 없음	2=해본 적 있음
상해	1=상해를 입힌 적 없음	2=상해를 입힌 적 있음
절도	1=훔친 적 없음	2=훔친 적 있음
절도< $50	1=훔친 적 없음	2=훔친 적 있음
집단 폭력	1=가담한 적 없음	2=가담한 적 있음

Table 3.2. Manifest items for substance abuse

항목	응답 범주		
담배	1=피워본 적 없음	2=피워본 적 있음	3=지금도 피우고 있음
술	1=마셔본 적 없음	2=마셔본 적 있음	3=많이 마심
마리화나	1=피워본 적 없음	2=피워본 적 있음	3=지금도 피우고 있음
기타 마약	1=사용한 적 없음	2=사용한 적 있음	3=지금도 사용하고 있음

형 등의 층 내에서 132개의 고등학교를 추출하였고, 각 학교 내에서 학교의 규모와 학생의 특성에 따라 학생들을 확률 추출하였다. 1차 조사는 1994-95년에 중학교 1학년부터 고등학교 3학년까지의 학생들을 대상으로 하였고, 해당 학생들에 대해서 1년 후인 1996년에 2차 조사를 하였다. 각 조사는 건강과 관련된 태도나 행동 등에 대한 광범위한 질문을 포함하고 있다.

본 연구에서의 처치변수는 비행에 대한 잠재범주이고 이는 1994-95에 시행된 1차 조사에서 얻어진 데이터를 기반으로 추정하였다. 결과변수인 약물사용에 대한 잠재범주는 1996 년에 시행된 2차 조사에서 얻어진 응답을 토대로 추정하였고, 교란변수로는 1차 조사에서 얻어진 성별, 가족의 소득을 포함한 11개의 변수를 사용하였다. 분석에 사용된 개체는 고등학교 1학년부터 3 학년까지의 학생들이고, 개체 수는 총 2,013명이다. 이때 비행을 측정하기 위해 사용된 처치관측변수 항목은 표 3.1 과 같고, 약물사용을 측정하기 위해 사용된 결과관측변수 항목은 표 3.2 와 같다.

비행을 측정하기 위한 항목들의 경우 각각 지난 1년간 거짓말을 한 횟수, 공공장소에서 크게 소리를 지른 횟수, 남의 물건에 고의로 상해를 입힌 횟수, 물건을 사고 돈을 내지 않은 횟수, 미화 $50 이하에 해당하는 돈이나 물건을 훔친 횟수, 집단 간 폭력을 행사한 횟수를 조사하여 0회는 1 = '없음'으로 1회 이상은 2 = '있음'으로 정의하여 이항변수로 변환하였다. 약물 사용을 측정하기 위한 항목들에 대해서는 담배의 경우, '담배를 피워본 적이 있는가'와 '최근 30일 동안 몇 개비의 담배를 피웠는가'의 두 문항을 이용해 담배 사용에 관한 변수를 만들었다. ' 담배를 피워본 적이 있는가'에 '그렇지 않다'고 답한 경우 1 = '피워본 적 없음'으로, '그렇다'고 답했으나 최근 30일 동안 피운 담배가 한 개비도 없는 경우 2 = '피워본 적 있음'으로, 최근 30일 동안 한 개비 이상 피운 경우 3 = '지금도 피우고 있음'으로 정의하였다. 마리화나와 기타 마약에 대해서도 담배와 마찬가지로 세 개의 응답 범주를 갖는 변수로 정의하였다. 그리고 술에 관한 항목에 대해서는 '술을 마셔본 적 있는가'와 '최근 1년 동안 다섯 잔 이상의 술을 마신 날이 며칠이나 되는가'의 두 질문에 대한 응답을 바탕으로, 마셔본 적 없는 경우 1 = ' 마셔본 적 없음', 마셔본 적은 있으나 최근 1년간 다섯 잔 이상의 술을 마신 적이 없는 경우 2 = '적당히 마심', 최근 다섯 잔 이상의 술을 한 번이라도 마신 경우 3 = '많이 마심'으로 정의하였다.

3.2. 결과

이 장에서는 미국의 Add Health의 자료를 이용하여 약물사용 패턴에 대한 비행 유형의 원인적 영향력

Table 3.3. Goodness-of-fit for model selection

처치잠재범주의 수	추정된 모수의 수	우도비	자유도	AIC	BIC	우도값
2	13	573.40	50	599.40	678.51	−9383.36
3	20	133.26	43	173.26	294.96	−9163.29
4	27	48.64	36	102.64	266.94	−9120.98
5	34	33.04	29	101.04	307.94	−9113.18
6	41	22.22	22	104.22	353.72	−9107.78

Table 3.4. Item response probabilities for adolescent delinquency

	잠재처치범주			
항목	(범주 1) 미 비행 청소년	(범주 2) 언어적 비행 청소년	(범주 3) 절도형 비행 청소년	(범주 4) 전반적 비행 청소년
거짓말	0.35	0.86	0.78	0.89
	(0.02)	(0.04)	(0.03)	(0.04)
고성방가	0.22	0.83	0.59	0.99
	(0.02)	(0.04)	(0.05)	(0.04)
상해	0.01	0.26	0.22	0.84
	(0.01)	(0.03)	(0.04)	(0.12)
절도	0.02	0.04	0.95	0.89
	(0.01)	(0.06)	(0.04)	(0.06)
절도 $<$ \$50	0.00	0.04	0.73	0.89
	(0.00)	(0.03)	(0.07)	(0.05)
집단 폭력	0.05	0.31	0.22	0.62
	(0.01)	(0.03)	(0.04)	(0.07)
출현율	0.51	0.25	0.17	0.07

을 추정하기 위하여 2.3 장의 분석 절차에 따라 분석한 결과를 보여준다. 청소년의 비행정도를 측정하기 위해 사용한 문항은 표 3.1과 같고 교란변수로는 성별, 가족의 소득 등 11개의 변수를 고려하였으며 모두 1차 조사에서 측정되었다. 잠재처치범주의 수를 결정하기 위해 잠재범주모형에 대한 적합통계량을 표 3.3에 정리하였다. 잠재범주의 수가 두 개인 모형부터 여섯 개인 모형까지 총 다섯 개의 LCA 모형을 적합 시킨 결과 AIC를 기준으로 잠재범주의 수가 다섯 개인 모형이 가장 좋고, BIC를 기준으로는 잠재범주의 수가 네 개인 모형이 가장 좋다고 볼 수 있다. 두 모형 중 잠재처치범주의 수가 네 개인 모형에서 각 범주의 해석이 적절하게 이루어지므로 잠재범주의 수는 네 개로 결정하였다.

잠재범주의 수가 네 개인 모형을 적합시킨 결과 각 범주에 대한 문항응답확률과 이에 대한 표준오차 및 출현율을 표 3.4에 정리하였다. 범주 1의 문항응답확률을 살펴보면 범주 1에 속해 있는 대부분의 청소년들은 6개의 모든 비행항목에서 '경험없음'으로 응답하여 비행을 저지르지 않는 일반적인 '미 비행 청소년'에 해당함을 알 수 있다. 이들의 출현율은 전체 모집단의 약 51%에 이른다. 범주 2에 속한 청소년 중 약 86%가 거짓말의 경험이 있으며 약 83%가 공공장소에서 소리를 지르는 등 ' 언어적 비행 청소년'을 의미하며 전체의 약 25%가 이 범주에 속한다. 범주 3은 범주 2의 언어적 비행외에 가게에서 물건을 사고 돈을 지불하지 않거나(73%), 미화 \$50이하의 가치에 해당하는 절도(73%)를 한 적이 있는 '절도형 비행 청소년'이며 전체의 17%이다. 마지막으로 범주 4는 거짓말, 도벽, 집단 폭력 등 전체적으로 나쁜 행동을 하고 있는 학생들로써 '전반적인 비행 청소년'으로 볼 수 있고 전체의 7%가 이 범주에 해당한다.

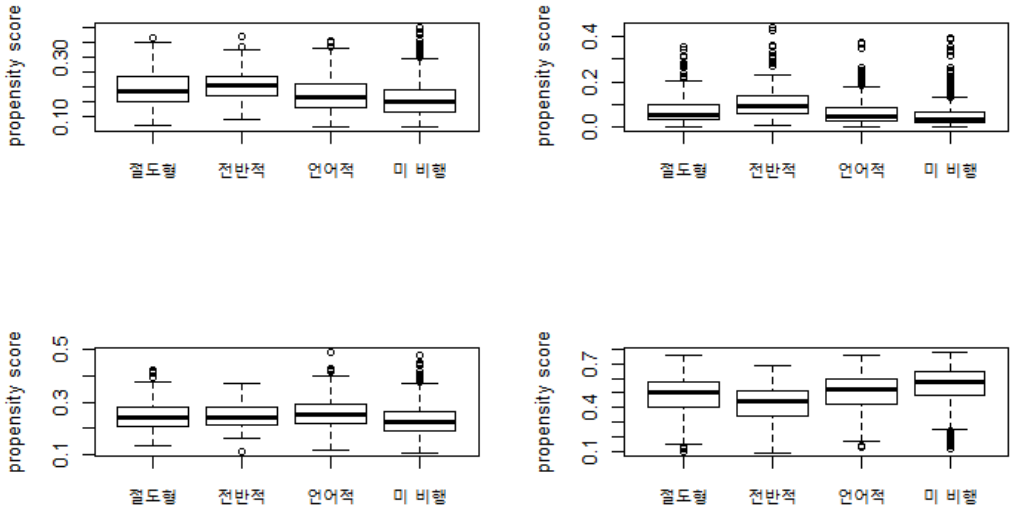

Figure 3.1. Boxplot for latent classes of adolescent delinquency

다음으로 잠재범주분석 결과 추정된 사후확률에 따라 개체마다 비행에 대한 네 개의 범주 중 하나의 범주를 무작위로 할당하였다. 할당된 비행 유형에 대해 기준범주 로지스틱 회귀모형을 적합 시켜 각 비행 유형에 대한 성향점수를 추정하였다. 그리고 원인적 영향력 추론의 적합성을 판단하기 위하여 성향점수 분포 간 겹침을 평가하였다. 이를 위해 비행의 네 범주에 대해 각 범주에 속할 성향점수의 상자그림을 그려서 성향점수 분포의 겹친 정도를 그림 3.1에서 비교하였다. 상자그림을 비교해 볼 때 비록 전체적인 성향점수의 분포가 완벽히 겹치는 것은 아니지만, 평균 원인적 영향력을 추정할 때 필요한 수준의 겹침을 만족한다고 볼 수 있다. 그림 3.1의 결과는 $N = 100$번의 반복 중 하나의 자료의 내용을 정리한 것으로서 나머지 99번의 반복된 자료에도 비슷한 결과가 나와 겹침의 수준에 문제가 없음을 알 수 있다. 따라서 표 2.2과 같이 가중치를 계산해 주고, 계산된 가중치에 따라 교란변수들의 효과가 조정되었는지 파악하기 위해 균형을 평가하였다.

본 연구에서는 $N = 100$번의 반복을 사용하였으나 기존 연구에서는 약 3-10회의 반복만으로도 만족할 수 있는 결과를 얻어낼 수 있다고 알려져있다(Graham 등, 2007; Rubin, 2004; Schafer, 1997). 그러나 Graham 등 (2007)의 논문에서는 필요 반복횟수는 결측정보와 검정력 모두와 관련이 있어 3-10회 보다는 많은 반복횟수를 사용하는 것을 추천하고 있다. 이에 따라 본 논문에서는 100회의 반복을 시행하였으며 PROC MI를 사용할 경우 1분 미만의 계산시간이 소요되었다.

그림 3.2와 그림 3.3은 가중치를 주기 전의 비행범주 간 교란변수의 표준화된 평균 차와 가중치를 준 후 비행범주 간 교란변수의 표준화된 평균차(SMD)를 계산하여 비교한 그림으로써 각각 ACE 및 $ACE_t (t = 1, 2, 3, 4)$를 추정하기 위한 가중치를 적용한 그림을 나타낸다. 그림 3.2와 그림 3.3을 살펴보면 성향점수를 이용하여 가중치를 주기 전에는 비행범주 간 SMD가 -0.8부터 0.8까지 나타났으나 가중치를 부여한 후에는 그 차이가 급격히 줄어들어 -0.3부터 0.3 이내로 나타나는 것을 확인할 수 있다.

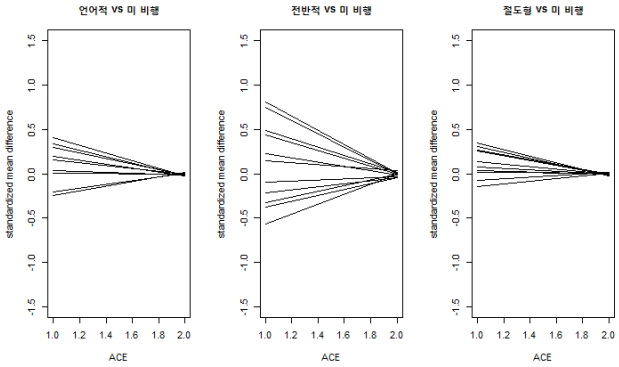

Figure 3.2. Standardized mean difference for ACE

Table 3.5. Frequencies for the best latent class models with weights

	AIC					BIC				
	2	3	4	5	6	2	3	4	5	6
ACE	0	0	0	75	25	0	95	5	0	0
ACE_1	0	0	0	72	28	0	100	0	0	0
ACE_2	0	0	0	63	37	0	100	5	0	0
ACE_3	0	0	0	56	44	0	89	11	0	0
ACE_4	0	0	0	90	10	0	100	0	0	0

다음으로 표 3.2의 네 개의 항목을 사용하여 약물사용에 대한 LCA 모형을 수행하였다. 이때 표 2.2의 가중치를 고려한 모형을 적합시켰으며 잠재처치변수와 마찬가지로 잠재결과범주의 수를 결정하기 위한 모형선택의 과정이 필요하다. 여기서는 모형 선택 과정에서부터 가중치를 고려해주어야 하며 $N = 100$번의 반복 중 각각의 반복에 대해 모형 선택 과정을 실시하여 반복마다 가장 좋은 모형에 대한 잠재범주의 수의 빈도를 정리하여 표 3.5에 나타내었다.

다섯 개의 가중치 구조에 대해 잠재범주의 수가 두 개인 모형부터 여섯 개인 모형까지 총 다섯 개씩의 잠재범주모형을 적합 시킨 결과 AIC를 기준으로 잠재범주의 수가 다섯 개 혹은 여섯 개인 모형이 가장 좋고, BIC를 기준으로는 잠재범주의 수가 세 개인 모형이 가장 좋다고 볼 수 있다. 세 모형 중 잠재범주의 수가 세 개인 모형에서 각 범주의 해석이 적절하게 이루어졌고, ACE와 ACE_t에 대해 잠재결과범주가 동일하게 측정되었으며, 처치범주 부여를 $N = 100$번 반복했을 때에도 잠재범주의 의미가 동일하게 측정되었다. 그러나 잠재범주의 수가 다섯 개 혹은 여섯 개인 경우 ACE 및 $ACE_t(t = 1, 2, 3, 4)$ 에 대해 잠재범주의 의미가 동일하게 측정되지 않고, $N = 100$번의 반복에 따라 잠재범주의 의미가 서로 다르게 측정되어 모형이 불안정하고 해석에 어려움이 있으므로 약물사용에 대한 잠재범주의 수는 세 개로 결정하였다.

먼저 ACE를 추정하기 위한 가중치를 적용하여 세 개의 잠재범주를 갖는 LCA 모형을 $N = 100$번 적합시킨 결과를 루빈의 규칙에 의해 종합하여 표 3.6에 나타내었다. 범주 1의 문항응답확률을 살펴보면 범주 1에 속한 대부분의 청소년들은 네 개의 모든 약물사용 항목에서 '경험없음'으로 응답하여 약물을 사용하지 않는 '미사용자'에 해당함을 알 수 있다. 이들의 출현율은 전체 모집단의 약 47%에 해당한다. 범주 2 에 속한 청소년의 약 65%가 현재 담배를 피고 있으며 약 68%가 최근 1년간 5잔이상의 술을 마신 날이 있는 것으로 응답하여 '담배+폭음'을 하는 청소년을 의미하며 전체의 약 41%에 이른다. 마지

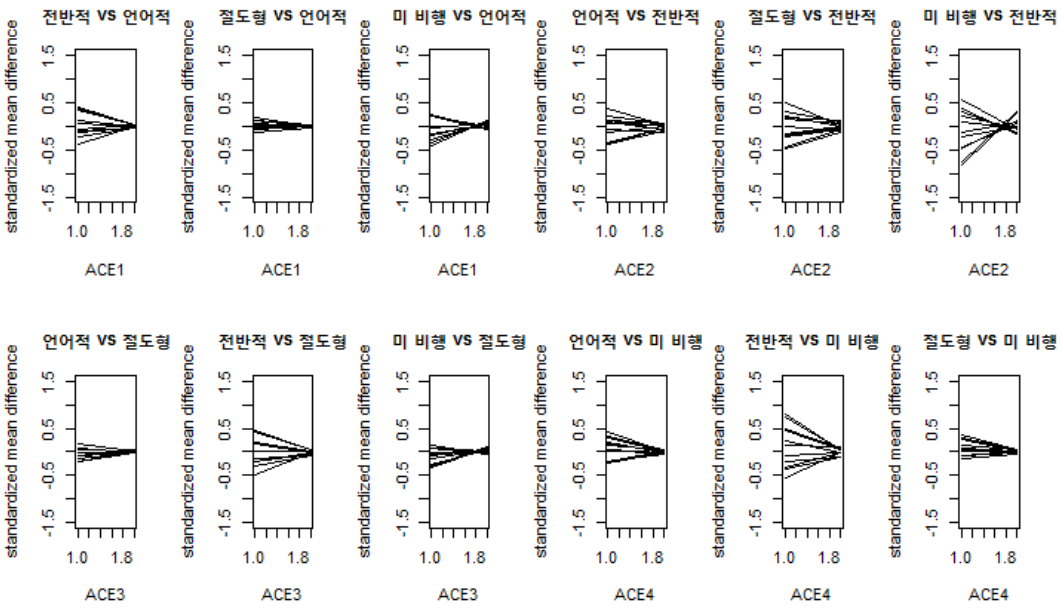

Figure 3.3. Standardized mean difference for ACE_t

막으로 범주 3은 범주 2의 담배와 폭음외에 마리화나와 기타 마약을 사용한 적이 있거나 지금도 사용하는 청소년으로써 모든 '약물 사용자'에 해당하며 전체의 약 12%가 이 범주에 해당한다.

$ACE_t(t = 1, 2, 3, 4)$를 추정하기 위한 가중치를 적용하여 세 개의 잠재범주를 갖는 LCA 모형을 $N = 100$번 적합시킨 결과를 루빈의 규칙에 의해 종합하여 표 3.7과 표 3.8에 나타내었다. 표 3.7과 표 3.8를 살펴보면 ACE를 추정하기 위한 가중치를 적용하여 얻은 표 3.6의 결과와 비슷함을 알 수 있다.

마지막으로 비행에 대해 LCA 모형을 적합시켜 얻은 사후확률에 따라 무작위로 할당한 잠재처치범주를 공변량으로 사용하여 약물사용에 대한 LCA 모형을 수행하였다. 이때 사용된 결과관측변수는 표 3.2에 나타나 있다. 이와 같은 과정을 무작위로 할당된 $N = 100$개의 잠재처치범주 자료를 사용하여 그 결과를 루빈의 규칙에 의해 종합하여 표 3.9에 평균 원인적 영향력을 정리하였다. 표 3.9을 살펴보면 전체적으로 미비행 청소년에 비하여 언어적, 절도형, 전반적 비행 청소년으로 갈수록 약물 미사용 청소년에 대한 약물사용의 오즈가 점점 커지는 것을 볼 수 있다. 예를 들어, 전체 모집단에 대한 ACE를 살펴보면 미 비행 청소년에 비하여 언어적, 절도형, 전반적 비행 청소년인 경우 담배와 폭음(범주 2) 을 함께할 오즈가 각각 약 1.9배, 2.5배, 4.6배 이며 모든 약물사용(범주 3)을 함께할 오즈는 각각 약 1.6배, 4.5배, 8.8배에 이르는 등 언어적, 절도형, 전반적 비행 청소년으로 갈수록 약물 미사용 청소년에 대한 약물사용의 오즈가 점점 커지며 이러한 현상은 모든 잠재처치범주의 도메인에 나타남을 알 수 있다. 특히 이러한 경향은 미 비행 청소년을 모집단으로 할 경우(ACE_1) 더욱 두드러지게 나타나 전반적 비행 청소년의 경우 담배와 폭음(범주 2)을 함께할 오즈 및 모든 약물사용(범주 3)을 함께할 오즈가 각각 약 5.2배와 11.1배가 됨을 알 수 있다. 이는 언어적 비행, 절도형 비행 및 전반적인 비행에 속하는 경우의 오즈비보다 월등히 큼을 알 수 있다. 즉, 비행을 저지르지 않는 청소년들이 비행을 저지를 경우 약물 사용에 대한 오즈가 비행을 저지르는 청소년들이 비행을 저지르는 경우의 약물사용에 대한 오즈보다 높다는 것을 의미한다.

Table 3.6. Item response probabilities for substance abuse from LCA for ACE

		잠재결과범주		
		(범주 1)	(범주 2)	(범주 3)
항목	응답범주	약물 미사용	담배+폭음	약물 사용
담배	피운 적 없음	0.74	0.18	0.12
		(0.21)	(0.11)	(0.23)
	피운 적 있음	0.08	0.16	0.07
		(0.02)	(0.04)	(0.06)
	지금 피움	0.18	0.65	0.82
		(0.21)	(0.11)	(0.23)
술	마신 적 없음	0.60	0.13	0.09
		(0.17)	(0.05)	(0.23)
	마신 적 있음	0.19	0.19	0.07
		(0.04)	(0.05)	(0.09)
	많이 마심	0.21	0.68	0.84
		(0.19)	(0.08)	(0.26)
마리화나	피운 적 없음	0.88	0.40	0.10
		(0.26)	(0.15)	(0.28)
	피운 적 있음	0.05	0.29	0.03
		(0.07)	(0.07)	(0.10)
	지금 피움	0.07	0.31	0.87
		(0.20)	(0.13)	(0.30)
기타 마약	사용한적 없음	0.96	0.92	0.31
		(0.12)	(0.06)	(0.32)
	사용한 적 있음	0.01	0.06	0.20
		(0.05)	(0.04)	(0.15)
	지금 사용	0.03	0.02	0.49
		(0.07)	(0.04)	(0.26)
출현율		0.47	0.41	0.12

4. 결론 및 토의

본 논문에서는 처치변수와 결과변수 모두 잠재범주인 경우 LCA 모형을 이용하여 원인적 영향력을 추론하는 방법을 RCM의 틀 안에서 제안하였다. 일반적인 관찰연구에서는 여러 처치의 결과를 동시에 관찰할 수 없으나 RCM을 통해 적절한 가정을 사용하면 현실에서 관찰된 값만으로 처치변수의 원인적 영향력을 추정할 수 있다. 그러나 무작위 통제시험 설계가 아닌 경우, 처치변수의 원인적 영향력의 불편 추정치를 얻기 위하여 교란변수를 고려해주어야 한다. 이 때 i번째 개체의 교란변수의 벡터 $\mathbf{x}_i$가 주어진 경우 개체 i가 특정한 처치를 받는 확률을 의미하는 성향점수를 모형을 통해 추정하여 ACE를 추정하는 것이 가능하다. 성향점수는 π_{it}가 상수인 부집단의 모든 범주에서 동일한 공변량의 분포를 가지므로 성향점수는 모든 다항범주 사이의 교란변수를 균형있게 만들어 준다고 할 수 있기 때문이다. 이러한 성향점수의 특성을 이용하면 짝짓기, 층화 및 처치변수의 역확률을 가중치로 사용하는 주변구조모형 등을 이용하여 ACE, ACE_t 및 이와 관련된 오즈비를 추정할 수 있다. 이 때 원인적 영향력은 직접 관찰되지 않기 때문에 모집단에서의 평균 원인적 영향력을 추정하는 방법을 이용하며 범주형 결과변수에 대한 원인적 영향력은 오즈비를 통해 ACE를 정의할 수 있다. 그러나 많은 관찰연구에서 결괏값 뿐만 아니라 그에 영향을 미치는 처치변수 또한 하나의 관측변수로 명확히 측정되기 어려운 경우가 있다. 이러한 경

Table 3.7. Item response probabilities for substance abuse from LCAs for ACE_1 and ACE_2

		ACE_1			ACE_2		
항목	응답범주	(범주 1) 약물 미사용	(범주 2) 담배+폭음	(범주 3) 약물 사용	(범주 1) 약물 미사용	(범주 2) 담배+폭음	(범주 3) 약물 사용
담배	피운 적 없음	0.74	0.18	0.12	0.65	0.18	0.16
		(0.17)	(0.09)	(0.18)	(0.23)	(0.12)	(0.25)
	피운 적 있음	0.08	0.17	0.05	0.10	0.13	0.03
		(0.02)	(0.04)	(0.05)	(0.03)	(0.03)	(0.03)
	지금 피움	0.17	0.66	0.86	0.24	0.69	0.81
		(0.16)	(0.10)	(0.19)	(0.24)	(0.13)	(0.27)
술	마신 적 없음	0.60	0.13	0.09	0.53	0.13	0.14
		(0.14)	(0.05)	(0.18)	(0.18)	(0.06)	(0.27)
	마신 적 있음	0.19	0.20	0.06	0.17	0.20	0.03
		(0.04)	(0.05)	(0.07)	(0.05)	(0.06)	(0.05)
	많이 마심	0.20	0.67	0.87	0.30	0.68	0.83
		(0.15)	(0.07)	(0.22)	(0.21)	(0.09)	(0.29)
마리화나	피운 적 없음	0.91	0.41	0.07	0.83	0.40	0.15
		(0.21)	(0.14)	(0.22)	(0.29)	(0.18)	(0.30)
	피운 적 있음	0.04	0.30	0.03	0.05	0.25	0.02
		(0.05)	(0.07)	(0.09)	(0.08)	(0.08)	(0.07)
	지금 피움	0.06	0.30	0.90	0.12	0.14	0.83
		(0.16)	(0.12)	(0.24)	(0.22)	(0.12)	(0.30)
기타 마약	사용한 적 없음	0.97	0.91	0.34	0.94	0.90	0.29
		(0.10)	(0.05)	(0.28)	(0.16)	(0.57)	(0.31)
	사용한 적 있음	0.01	0.07	0.15	0.02	0.09	0.12
		(0.04)	(0.04)	(0.11)	(0.07)	(0.06)	(0.10)
	지금 사용	0.02	0.02	0.50	0.04	0.01	0.60
		(0.06)	(0.03)	(0.24)	(0.10)	(0.01)	(0.29)
출현율		0.46	0.41	0.13	0.46	0.41	0.13

우 여러 측정변수를 이용하여 각 항목의 반응에 따라 구성원의 일부를 동질 집단으로 분류하기 위한 가장 간단한 혼합모형 중의 하나인 LCA 모형을 사용할 수 있다. LCA 모형은 관측 가능한 범주형 변수들 사이의 관계를 직접 관측할 수 없는 몇 개의 잠재범주로 설명하고자 하는 분석방법으로써, 잠재처치범주가 알려지면 처치관측변수는 독립이라는 지역독립성의 가정을 하고 있다. 또한 로짓모형을 이용하여 각각의 개체가 잠재결과범주에 속할 확률을 추정하는 것이 가능하다.

이를 통해 본 논문에서 잠재결과변수에 대한 잠재처치변수의 원인적 영향력을 추론하기 위해 제안한 절차는 다음과 같다. 먼저 처치변수에 대해 잠재범주분석을 하였고 사후확률에 따라 각 개체가 속하는 범주를 무작위로 할당해주었다. 이어서 성향점수를 추정하기 위해 할당받은 처치범주에 대해 교란변수들의 로지스틱 회귀모형을 적합시킨 후 처치범주 간 성향점수의 분포가 적당히 겹치는 것을 확인하였다. 이때 계산된 성향점수를 이용해 ACE 및 $ACE_t(t = 1, \ldots, C)$를 추정하기 위한 가중치를 계산해 주었다. 또한, 교란변수 분포의 균형을 평가하기 위해 가중치를 주기 전의 처치범주 간 교란변수의 표준화된 평균 차이와 가중치를 준 후의 표준화된 평균 차이를 비교하였다. 균형이 이루어지는 것을 확인한 후, 결과변수에 대해 LCA 모형을 적합 시키는데, 이때 처치범주에 대한 가변수들을 예측변수로 고려함으로써 결과변수의 범주에 대한 처치범주의 원인적 영향력을 추정할 수 있다. 마지막으로 처치에 대한 잠재범주모형의 사후확률에 따라 각 개체가 속하는 범주를 부여하는 단계부터 원인적 영향력을 추정하는 것

Table 3.8. Item response probabilities for substance abuse from LCAs for ACE_3 and ACE_4

		ACE_3			ACE_4		
항목	응답범주	(범주 1) 약물 미사용	(범주 2) 담배+폭음	(범주 3) 약물 사용	(범주 1) 약물 미사용	(범주 2) 담배+폭음	(범주 3) 약물 사용
담배	피운 적 없음	0.66	0.22	0.21	0.75	0.19	0.10
		(0.30)	(0.14)	(0.33)	(0.21)	(0.16)	(0.22)
	피운 적 있음	0.08	0.17	0.07	0.08	0.17	0.08
		(0.03)	(0.04)	(0.05)	(0.03)	(0.04)	(0.07)
	지금 피움	0.26	0.61	0.72	0.17	0.64	0.82
		(0.30)	(0.15)	(0.32)	(0.20)	(0.17)	(0.23)
술	마신 적 없음	0.54	0.14	0.18	0.61	0.14	0.07
		(0.24)	(0.06)	(0.33)	(0.18)	(0.09)	(0.22)
	마신 적 있음	0.17	0.21	0.07	0.19	0.18	0.10
		(0.06)	(0.06)	(0.08)	(0.04)	(0.05)	(0.12)
	많이 마심	0.29	0.65	0.75	0.20	0.68	0.83
		(0.28)	(0.10)	(0.36)	(0.20)	(0.12)	(0.26)
마리화나	피운 적 없음	0.78	0.47	0.20	0.88	0.40	0.08
		(0.37)	(0.19)	(0.39)	(0.26)	(0.20)	(0.26)
	피운 적 있음	0.66	0.27	0.04	0.05	0.30	0.03
		(0.12)	(0.09)	(0.08)	(0.09)	(0.08)	(0.11)
	지금 피움	0.16	0.26	0.77	0.06	0.30	0.90
		(0.31)	(0.12)	(0.38)	(0.20)	(0.17)	(0.29)
기타 마약	사용한 적 없음	0.91	0.94	0.41	0.96	0.93	0.27
		(0.19)	(0.04)	(0.36)	(0.10)	(0.05)	(0.39)
	사용한 적 있음	0.04	0.05	0.18	0.01	0.05	0.24
		(0.08)	(0.04)	(0.15)	(0.05)	(0.03)	(0.18)
	지금 사용	0.05	0.01	0.41	0.02	0.01	0.49
		(0.12)	(0.01)	(0.28)	(0.06)	(0.03)	(0.29)
출현율		0.42	0.42	0.16	0.48	0.40	0.12

까지의 과정을 N번 반복하여 추정된 N개의 추정치를 루빈의 규칙에 따라 종합시킴으로써 평균 원인적 영향력을 추정할 수 있다.

이처럼 본 논문에서는 처치변수와 결과변수가 모두 잠재범주인 경우 원인적 영향력 추론을 위해 처치변수에 대한 잠재범주를 할당하여 결과범주에 대한 원인적 영향력을 추정하는 과정을 N번 반복하여 수행하고 있다. 이는 매우 번거로운 과정이며 계산이 복잡하고 분석에 따른 시간도 매우 오래걸린다. 또한 잠재범주모형에서 중요한 문제 중 하나가 적절한 잠재범주의 수를 선택하는 것인데, 추정하고자하는 원인적 영향력의 성격에 따라, 그리고 분석 절차를 N번 반복함에 따라 가장 적절한 잠재범주의 수가 달라질 수 있으므로, N의 수를 선택하는데 있어서 어려움이 있을 수 있다. 그리고 본 논문에서 제안하는 방법의 경우 처치에 대한 잠재범주가 알려진 경우에만 사용할 수 있어 효율성이 낮으므로 이 또한 본 논문의 한계점이라고 볼 수 있다. 이를 해결하기 위해 처치와 결과를 동시에 고려한 가능도 함수를 모형화하여 모수를 추정하는 것이 향후 진행되어야 할 연구 과제가 될 수 있을 것이다. 모형을 통해 원인적 영향력을 추론할 경우, 처치에 대한 범주가 알려져 있어야 한다는 가정이 불필요하며 복잡한 계산을 반복적으로 수행하지 않아도 되므로 효율성이 뛰어날 것이다.

그러나 지금까지 처치변수와 결과변수가 모두 잠재범주인 경우 원인적 영향력을 추론하는 방법이 제안된 바가 없다는 점에서 본 논문은 의의가 있으며, 본 논문에서 다룬 주제 외에도 다양한 주제들을 다루

Table 3.9. Average causal effect of adolescent delinquency to substance abuse and their 95% confidence interval

		잠재결과범주*			
		(범주 2) 담배+폭음		(범주 3) 약물 사용	
도메인	잠재처치범주†	오즈비	95% 신뢰구간	오즈비	95% 신뢰구간
전체 모집단 (ACE)	언어적 비행 청소년	1.886	[0.948, 3.752]	1.619	[0.475, 5.518]
	절도형 비행 청소년	2.499	[0.759, 8.231]	4.455	[0.516, 38.441]
	전반적 비행 청소년	4.586	[0.640, 32.888]	8.751	[0.277, 276.123]
미 비행 청소년 (ACE_1)	언어적 비행 청소년	2.031	[0.995, 4.148]	1.605	[0.407, 6.332]
	절도형 비행 청소년	2.740	[0.763, 9.835]	4.886	[0.560, 42.675]
	전반적 비행 청소년	5.239	[0.465, 58.964]	11.124	[0.329, 376.505]
언어적 비행 청소년 (ACE_2)	언어적 비행 청소년	1.617	[0.628, 4.165]	1.350	[0.297, 6.145]
	절도형 비행 청소년	1.917	[0.388, 9.466]	2.900	[0.152, 55.178]
	전반적 비행 청소년	2.815	[0.240, 32.951]	4.316	[0.044, 427.273]
절도형 비행 청소년 (ACE_3)	언어적 비행 청소년	1.849	[0.977, 3.500]	1.758	[0.549, 5.630]
	절도형 비행 청소년	2.505	[0.942, 6.665]	4.665	[0.792, 27.467]
	전반적 비행 청소년	4.483	[0.879, 22.852]	8.376	[0.550, 127.658]
전반적 비행 청소년 (ACE_4)	언어적 비행 청소년	1.571	[0.426, 5.793]	1.536	[0.261, 9.028]
	절도형 비행 청소년	1.717	[0.345, 8.556]	3.286	[0.237, 45.578]
	전반적 비행 청소년	2.560	[0.268, 24.501]	5.218	[0.190, 143.796]

† 잠재처치범주의 기준범주는 (범주 1) '미 비행 청소년' 임.

* 잠재결과범주의 기준범주는 (범주 1) '약물 미사용' 임.

는 데 유용하게 쓰일 수 있다. 예를 들어 다이어트 패턴이 심리 상태에 미치는 원인적 영향력, 또는 사춘기 시기가 약물 사용에 미치는 영향력 등을 추론하고자 하는 연구자들에게 본 논문이 제안하는 분석 방법이 도움이 될 것이다.

References

Agresti, A. (2002). *Categorical Data Analysis* (Second ed.). Hoboken, NJ: Wiley.

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 Behavioral Sciences (2nd ed.).* Hillsdale, NJ: Erlbaum.

Collins, L. M. and Lanza, S. T. (2010). *Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences.* New York, NY: Wiley.

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

Flay, B. R. (1993). *Youth tobacco use: risks, patterns, and control. In C. Orleans & J. Slade (Eds.), Nicotine Addiction,* (pp. 360-384). New York, NY: Oxford University Press.

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

Kang, J. and Schafer, J. L. (2010). Estimating Average Treatment Effects When the Treatment is a Latent Class. *Department of Statistics, The Pennsylvania State University,* Technical Report, 10-05.

Lanza, S. T., Collins, L. M., Lemmon, D. R., and Schafer, J. L. (2007). PROC LCA: a SAS procedure for latent class analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, **14**, 671–694.

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.

Leventhal, H. and Cleary, P. D. (1980). The smoking problem: A review of the research and theory in behavioral risk modification. *Psychological Bulletin*, **88**, 370–405.

Mayhew, K. P., Flay, B. R. and Mott, J. A. (2000). Stages in the development of adolescent smoking. *Drug and Alcohol Dependence*, **59**, S61–S81.

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

Rosenbaum, P. R. (2002). *Observational Studies* (Second ed.). New York, NY: Springer.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies. *Journal of Educational Psychology*, **66**, 688–701.

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

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

Rubin, D. B. (2004). *Multiple Imputation for Nonresponse in Surveys.* New York, NY: John Wiley and Sons.

Rubin, D. B. (2005). Causal inference using potential outcomes: design, modeling, decisions. *Journal of American Statistical Association*, **469**, 322–331.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data.* London: Chapman and Hall.

Schafer, J. L. and Kang, J. (2008). Average causal effects from non-randomized studies: a practical guide and simulated example. *Psychological Methods*, **13**, 279–313.

Graham, J.W., Olchowski, A.E. and Gilreath, T.D. (2007) How many imputations are really needed Some practical clarifications of multiple imputation theory. *Prevention Science*, **8**, 206–213.

Udry, J. R. (2003). *The National Longitudinal Study of Adolescent Health (Add Health), Waves Ⅰ and Ⅱ, 1994-1996.* Chapel Hill, NC: Carolina Population Center, University of North Carolina at Chapel Hill.

잠재범주분석을 이용한 원인적 영향력 추론에 관한 연구

박가영[a] · 정환[a,1]

[a]고려대학교 통계학과

(2014년 08월 25일 접수, 2014년 11월 10일 수정, 2014년 12월 08일 채택)

요 약

관찰연구를 이용하여 인과관계를 추론할 경우 무작위 통제시험과는 달리 교란변수로 인한 편향을 제어하기 위한 통계적 전략이 필요하다. 최근에는 성향점수(propensity score) 를 이용한 짝짓기나 원인변수의 역확률을 가중치로 사용하는 주변구조모형이 제안되어 사용되고 있다. 이러한 인과관계 추론은 처치(treatment)가 명확히 주어진 경우에 교란변수를 통제하고 그 처치가 결과에 미치는 영향을 평가하는 방법에 초점이 맞추어져 있다. 하지만 기존의 방법의 경우 원인변수인 처치가 직접관측이 가능한 범주형 변수이고 결과변수 또한 직접관측이 가능한 변수인 경우에만 사용할 수 있는 한계를 갖고 있다. 본 연구에서는 원인변수인 처치와 결과변수의 결괏값의 직접적인 관측이 어려운 경우, 측정오차를 고려한 잠재범주모형(latent class analysis)의 변수로 모형화 함으로써 잠재범주 간의 원인적 영향력을 추정하는 방법을 제시하고자 한다. 그리고 미국의 The National Longitudinal Study of Adolescent Health 자료를 이용하여, 약물사용의 잠재범주에 대한 청소년기의 비행(delinquency)이라는 잠재범주의 원인적 영향력을 추정하였다.

주요용어: 평균 원인적 영향력, 인과 관계 추론, 잠재범주분석, 성향점수.

이 연구는 2012년도 고려대학교 특별연구비(G1300030)의 지원을 받아 수행되었으며, 제 1저자 박가영의 석사학위논문의 축약본임.

[1]교신저자: (136-701) 서울특별시 성북구 안암로 145, 고려대학교 통계학과, 부교수.
E-mail: hwanch@korea.ac.kr