미국 청소년 데이터를 활용한 범주형 자료 분석

STA518 Final project

통계학과 2021021352 윤아연

- 목차 -

1. 서론

1-1. 연구 주제 및 목적

청소년기는 아동기를 벗어나 성인이 되기 전 신체적/정신적 성숙이 진행되는 시기이다. 이 시기의 청소년은 자아 정체감을 형성하면서 자신에 대해 깊게 탐구해보거나 미래를 계획하며 고민에 빠지기도 한다. 더불어 '질풍노도의 시기'라고 불릴 만큼 청소년기에서는 주위 환경에 대한 불만과 일탈을 경험하기도 한다. 따라서 청소년이 올바른 가치관을 성립할 수 있도록 사회와 교육기관의 지도가 요구되며 적절한 지도 방향을 설정하기 위해서는 청소년 집단의 특성을 파악해야 할 것이다.

청소년 지도 방향의 핵심은 정신적으로 건강하고 사회적 일탈 행동을 하지 않도록 하는 것이다. 이러한 목적에 따라 본 연구는 청소년의 삶의 만족도와 일탈 행동 중 하나인 약물 사용에 집중해 연구를 진행하고 연구 결과에 따른 청소년 지도 방향을 제시하고자 했다. 또한 청소년의 인구통계학적 요소 및 정치·종교적 가치관을 함께 살펴봄으로써 청소년의 특성을 파악하고 삶의 만족도 및 약물 사용 행태와 어떠한 연관성이 있는지 분석했다.

1-2. 데이터 설명

본 연구에 사용된 데이터는 'MONITORING THE FUTURE: A CONTINUING STUDY OF AMERICAN YOUTH'(이하 MTF) 연구의 2004년 데이터이다. 해당 데이터는 사회 및 행동과학에 관련된 데이터를 제공하는 ICPSR에서 다운로드 받을 수 있다.[1] MTF 연구는 미시간 대학교의 사회 연구소(University of Michigan's Institute for Social Research)에서 1975년부터 미국의 청소년을 대상으로 매년 진행해온 연구이다. MTF 연구는 당년의 미국 청소년의 중요한 가치, 행동, 그리고 지향하는 생활 방식 등을 광범위하게 탐구하고 매년 이러한 부분이 어떻게 변화하는지 파악하기 위한 연구이다.

구체적으로 MTF 연구는 약 130개의 미국 공립·사립 고등학교를 선택해 해당 학교의 12학년 학생(12th grade, senior)을 대상으로 진행된다. MTF 연구의 표본 수집은 지역-학교-학생의 3단계로 진행된다. 먼저 1단계(지역)에서는 미국의 각 지역 중 연구를 진행할 지역을 primary sampling unit으로 선택한다. 다음 2단계(학교)에서는 1단계에서 선택된 지역에 포함된 고등학교 중 연구를 진행할 지역을 선택한다. 이때 각 고등학교의 선택 가중치는 해당 고등학교의 상급반(senior class)의 크기에 비례하도록 설정한다. 일반적으로 각 지역마다 한 개의 고등학교가 선택되며 대도시의 경우 두 개 이상의 고등학교가 선택되기도 한다. 마지막 3단계(학생)에서는 2단계에서 선택된 고등학교에서 연구의 대상이 될 12학년 학생을 선택한다. 고등학교마다 400명의 학생을 선택하며, 학생 선택 시 학급을 무작위로 선택하는 등 편향이 일어나지 않는 방법으로 추출한다. 만약 전체 12학년 학생이 400명 이하인 고등학교의 경우, 모든 상급반 학생을 연구 대상으로 추출한다. 2004년 연구에서는 총 128개의 고등학교에서 15,222명의 12학년 학생이 선택되었다. MTF 연구는 고등학교를 다니는 청소년만을 대상으로 하기 때문에 학교 밖 청소년을 배제하게 된다는 단점이 있다. 하지만 MTF 연구는 학교 밖 청소년은 전체 청소년의 15~20%인 적은 비율이므로 이들이 연구에서 제외되더라도 연구 결과에 일정 이상의 편향을 가져오지 않는다고 설명한다. 또한 학교 밖 청소년까지 연구 대상으로 포함하는 것은 연구 비용과 노력의 막대한 증가를

1) 데이터 출처 : https://www.icpsr.umich.edu/web/ICPSR/studies/4264

가져온다는 점도 지적한다.

MTF 연구의 데이터는 설문지에 대한 학생들의 답변으로 구성된다. 설문지의 중심이 되는 주제는 약물 사용이지만, 그 밖에도 청소년의 생활습관 및 가치관을 파악할 수 있는 다양한 주제가 포함되어 있다. MTF 연구의 설문조사는 아래와 같은 20개의 주제를 담고 있다.

- MFT 연구의 설문조사 주제 -

1. 약물 (Drugs)
2. 교육 (Education)
3. 일과 여가 (Work and Leisure)
4. 성 역할과 가족 (Sex Role and Famliy)
5. 인구 문제 (Population Concerns)
6. 환경보호, 유물론, 형평성 등 (Conservation, Materialism, Equity, etc)
7. 종교 (Religion)
8. 정치 (Politics)
9. 사회적 변화 (Social Change)
10. 사회적 문제 (Social Problem)
11. 주요 사회 기관 (Major Social Institution)
12. 군대 (Military)
13. 대인 관계 (Interpersonal Relationship)
14. 인종 관계 (Race Relationship)
15. 타인에 대한 관심 (Concern for Others)
16. 행복 (Happiness)
17. 기타 성격 변수 (Other Personality Variables)
18. 배경 (Background)
19. 일탈적 행동과 피해 (Deviant Behavior and Victimization)
20. 건강 (Health)

이처럼 광범위한 주제를 다루기 위해서는 많은 설문 문항이 필요하므로 전체 문항은 6개의 세트로 나누어져 있다. 연구에 참여하는 학생은 6개 세트 중 무작위로 선택된 한 가지 설문 문항 세트에만 답하게 된다. 각 설문 문항 세트에는 공통적인 핵심(core) 변수-약물 사용 및 인구통계학적 변수-가 포함되어 분석에 도움을 줄 수 있다.

2. 본론

2-1. 연구 방법론 설명

2-1-1. 데이터 전처리

본 연구에서는 6개의 설문 문항 세트 중 첫 번째 세트를 사용하였다. 해당 세트는 총 618개의 변수와 2,563개의 관측치를 포함한다. 본 연구의 주요 관심사는 청소년의 삶의 만족도, 약물 사용, 그리고 인구통계학적 특징 및 정치·종교적 가치관이므로 이에 해당하는 13개의 변수를 선택했다.

변수코드 (변수명)	세부사항	답변 형태
V1646 (SAT OWN FRIENDS)	친구 및 지인에 대한 만족도	<1 - 7 범위의 답변> 1 : 매우 불만족 4 : 중간 7 : 매우 만족 ---------- -9 : 결측치
V1647 (SAT GT ALNG PRN)	부모님에 대한 만족도	
V1648 (SAT YOURSELF)	자기 자신에 대한 만족도	
V1650 (SAT TIME FR THG)	자유 시간에 대한 만족도	
V1653 (SAT GOVT OPRTNG)	정부 운영에 대한 만족도	
V1102 (#CIGS SMKD/30DA)	지난 30일 동안 흡연 정도	1 : 전혀 하지 않음 2 : 하루에 한 개비 이하 3 : 하루에 1~5개비 4 : 하루에 반 갑 5 : 하루에 한 갑 6 : 하루에 한 갑~한 갑 반 7 : 하루에 두 갑 이상 ---------- -9 : 결측치
V1216 (#X ALC/30D SIPS)	지난 30일 동안 음주 횟수	1 : 전혀 하지 않음 2 : 1~2회 3 : 3~5회 4 : 6~9회 5 : 10~19회 6 : 20~39회 7 : 40회 이상 ---------- -9 : 결측치
V1254 (#X MARJ/LAST 30D)	지난 30일 동안 마리화나 흡연 횟수	
V1712 (#X DIETPILL/30D)	지난 30일 동안 불법 처방받은 다이어트 약 섭취 횟수	
V1150 (R'S SEX)	성별	1 : 남성 2 : 여성 ---------- -9 : 결측치
V1151 (R'S RACE)	인종	0 : 백인 1 : 흑인

		---------- -9 : 결측치 (백인, 흑인 외 다른 인종은 결측치로 처리됨)
V1167 (R'S POL BLF RADICAL)	정치적 성향	1 : 매우 보수적 2 : 보수적 3 : 중립(온건) 4 : 진보적 5 : 매우 진보적 6 : 급진적(=radical) 8 : 없음/모름 ---------- -9 : 결측치
V1170 (RLGN IMP R'S LF)	인생에서 종교의 중요도	1 : 중요하지 않음 2 : 조금 중요함 3 : 꽤 중요함 4 : 매우 중요함 ---------- -9 : 결측치

Table 1. 연구에 사용된 변수 설명 및 답변 형태 (raw data)

Table 1.에는 연구에 사용된 13개의 변수와 해당 변수에 대한 답변 형태가 설명되어있다. 여기서 5개의 변수 V1646, V1647, V1648, V1650, V1653은 삶의 만족도를 측정하는 변수, 4개의 변수 V1102, V1216, V1254, V1712는 약물 사용 행태를 측정하는 변수, 그리고 마지막 4개의 변수 V1150, V1151, V1167, V1170은 인구통계학적 정보 및 정치·종교적 가치관을 측정하는 변수이다.

분석을 위해 결측치 제거 및 리코딩 작업을 진행했다. 모든 변수에 결측치가 존재했지만 V1150, V1151, V1167, V1170의 결측치만 제거해주었다. 나머지 9개 변수(삶의 만족도, 약물 사용 행태 관련 변수)의 결측치를 제거하지 않은 이유는 2-1-2. 분석방법론 부분에서 설명할 것이다. 분석의 편의를 위해 적절한 리코딩 과정도 진행되었다. 먼저 삶의 만족도 변수(V1646, V1647, V1648, V1650, V1653)에서는 1~2를 '불만족'(=1), 3~4를 '중간'(=2), 6~7을 '만족'(=3)으로 리코딩 했다. 약물 사용 변수(V1102, V1216, V1254, V1712)에서는 1을 '사용 안함'(=1), 2~4를 '약간 사용'(=2), 5~7을 '많이 사용'(=3)으로 리코딩 했다. 성별(V1150)과 인종(V1151)변수는 리코딩을 하지 않았다. 정치적 성향(V1647) 변수는 1~2는 '보수'(=1), 3은 '중립'(=2), 4~6은 '진보'(=3), 8은 '없음/모름'(=4)로 리코딩 했다. 마지막으로 종교의 중요도(V1170) 변수는 1~2는 '중요하지 않음'(=1), 3~4는 '중요함'(=2)으로 리코딩

했다. Table 2.에는 전처리를 마친 데이터의 변수 설명 및 답변 형태가 설명되어있다.
전처리를 마친 최종 데이터는 13개의 변수와 1,443개의 관측치가 포함되어있다.

변수코드 (변수명)	세부사항	답변 형태
V1646 (SAT OWN FRIENDS)	친구 및 지인에 대한 만족도	1 : 불만족 2 : 중간 3 : 만족 ---------- NA : 결측치
V1647 (SAT GT ALNG PRN)	부모님에 대한 만족도	
V1648 (SAT YOURSELF)	자기 자신에 대한 만족도	
V1650 (SAT TIME FR THG)	자유 시간에 대한 만족도	
V1653 (SAT GOVT OPRTNG)	정부 운영에 대한 만족도	
V1102 (#CIGS SMKD/30DA)	지난 30일 동안 흡연 정도	1 : 사용 안함 2 : 약간 사용 3 : 많이 사용 ---------- NA : 결측치
V1216 (#X ALC/30D SIPS)	지난 30일 동안 음주 횟수	
V1254 (#X MARJ/LAST 30D)	지난 30일 동안 마리화나 흡연 횟수	
V1712 (#X DIETPILL/30D)	지난 30일 동안 불법 처방받은 다이어트 약 섭취 횟수	
V1150 (R'S SEX)	성별	1 : 남성 2 : 여성 ---------- 결측치 없음
V1151 (R'S RACE)	인종	1 : 백인 2 : 흑인 ---------- 결측치 없음
V1167 (R'S POL BLF RADICAL)	정치적 성향	1 : 보수 2 : 중립 3 : 진보 4 : 없음/모름 ---------- 결측치 없음
V1170 (RLGN IMP R'S LF)	인생에서 종교의 중요도	1 : 중요하지 않음 2 : 중요함

		---------- 결측치 없음

Table 2. 전처리 후 최종 데이터의 변수 설명 및 답변 형태

2-1-2. 분석방법론

본 연구에 사용된 데이터는 설문조사에 대한 답변이므로 모두 명목형 범주형 변수로 이루어져 있다. 따라서 다음과 같은 범주형 자료 분석방법론을 사용해 연구를 진행했다.

(1) 카이제곱 독립성 검정

카이제곱 독립성 검정은 두 범주형 변수의 연관성에 대한 검정이다. 해당 검정의 귀무가설은 '두 변수는 서로 독립이다.'이다. 카이제곱 독립성 검정은 성별(V1150), 인종(V1151), 정치적 성향(V1167), 종교의 중요도(V1170) 변수 사이의 연관성을 알아보기 위해 사용되었다. 또한 다음에 설명할 잠재 층 분석 후 청소년의 삶의 만족도에 대한 잠재 층과 약물 사용에 대한 잠재 층 사이의 연관성을 알아보기 위해서도 사용되었다.

(2) 로그-선형 모형

로그-선형 모형은 둘 이상의 범주형 변수들 간의 독립성과 종속성을 알아볼 수 있는 모형이다. 다른 통계모형과 달리 로그-선형 모형은 모형에 사용되는 모든 변수를 반응변수로 간주한다. 로그-선형 모형은 모든 교호작용이 포함된 포화모형부터 모든 변수가 독립인 상호독립 모형까지 변수들의 독립성과 종속성에 따라 다양한 모형을 적합할 수 있으며, 우도비 검정 통계량을 사용해 내포(nested)관계에 있는 두 모형을 비교할 수 있다.

로그-선형 모형은 청소년의 인구통계학적 자료 및 정치·종교적 가치관 변수 사이의 독립성과 종속성을 구체적으로 알아보기 위해 사용되었다.

(3) 잠재 층 회귀 모형

잠재 층 분석은 일련의 택일형 문항에 대한 범주형 응답 자료를 분석할 수 있는 모형이다. 응답자들의 답변을 바탕으로 응답자들을 몇 개의 집단으로 분리할 수 있다는 것이 잠재 층 분석의 기본 전제이며, 이때 응답자들의 집단은 겉으로 드러나지 않고 잠재되어있으므로 '잠재 층'이라고 부른다.

잠재 층 회귀 모형은 응답자들의 답변뿐만 아니라 응답자들의 공변량을 분석에 활용해 잠재 층을 추정하는 모형이다. 기본적인 잠재 층 모형의 확장이라고 할 수 있다. 본 연구에서는 청소년의 삶의 만족도에 대한 잠재 층과 약물 사용 행태에 대한 잠재 층을 추정하기 위해 잠재 층 회귀 모형을 사용했으며 이때 성별, 인종, 정치적 성향, 종교의 중요도 4개의 변수는 공변량으로 사용하였다. 잠재 층 회귀 모형을 설명하면 다음과 같다.

(3-1) 표기 및 모수 설정

- n : 표본의 크기
- p : 공변량의 수
- C : 잠재 층의 개수
- M : 문항의 개수
- r_m : m번째 문항에 대해 가능한 응답 수

- $\boldsymbol{y}_i$: i번째 응답자의 응답, 데이터, M차원 벡터 $(i = 1, \cdots n)$
- $\boldsymbol{x}_i$: i번째 응답자의 공변량, 데이터, p차원 벡터
- $\gamma_l(\boldsymbol{x}_i)$: 공변량이 $\boldsymbol{x}_i$로 주어졌을 때 l번째 잠재 층에 속할 확률, 모수, C차원 벡터 $(i = 1, \cdots n,\ l = 1, \cdots, C)$
- $\boldsymbol{\beta}_l$: l번째 잠재 층에 대한 회귀 계수, p차원 벡터 $(l = 1, \cdots, C)$
- $\rho_{mk|l}$: 개체의 잠재 층이 l번째 잠재 층일 때, 해당 개체의 m번째 문항에 k라고 응답할 확률, 모수, $(m = 1, \cdots, M,\ k = 1, \cdots, r_m)$

(3-2) 확률 모형의 구조 및 추정

$$P(\boldsymbol{Y}_i = \boldsymbol{y}_i) = \sum_{l=1}^{C} \gamma_l(\boldsymbol{x}_i) \prod_{m=1}^{M} \prod_{k=1}^{r_m} \rho_{mk|l}^{I(y_{im}=k)} \quad \text{이때, } \gamma_l(\boldsymbol{x}_i) = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_l)}{1 + \sum_{j=1}^{C-1} \exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)}$$

잠재 층 회귀분석은 위의 확률 모형을 사용해 주어진 응답 $\boldsymbol{y}_i$가 나올 확률을 최대화하는 모수 $\boldsymbol{\beta}_l$, $\rho_{mk|l}$를 추정한다. 확률 모형의 로그 가능도에 대한 해답이 closed-form으로 주어지지 않으므로 잠재 층 회귀분석에서는 모수의 추정을 위해 EM 알고리즘과 Newton-Raphson 알고리즘과 같은 최적화 방법을 사용한다.

(3-3) 결측치의 처리

잠재 층 회귀분석의 모수 추정은 응답자의 사후확률을 정의하면서 진행된다. i번째 응답자의 응답이 $\boldsymbol{y}_i$로 주어졌을 때 해당 응답자의 잠재 층 L이 l번째 잠재 층일 사후확률 θ_{il}은 다음과 같다.

$$P(L = l \mid Y_i = \boldsymbol{y}_i) = \theta_{il} = \frac{\gamma_l(\boldsymbol{x}_i) \prod_{m=1}^{M} \prod_{k=1}^{r_m} \rho_{mk|l}^{I(y_{im}=k)}}{\sum_{j=1}^{C} \gamma_j(\boldsymbol{x}_i) \prod_{m=1}^{M} \prod_{k=1}^{r_m} \rho_{mk|j}^{I(y_{im}=k)}}$$

만약 응답의 결측치가 있다면 응답자의 사후확률은 어떻게 정의될 수 있을까? 잠재 층 회귀분석에서는 응답에 결측치가 존재하더라도 사후확률 정의 및 모수 추정이 가능하다. 결측치가 존재할 경우, 잠재 층 회귀분석에서는 관측된 응답만을 사용해 다음과 같은 사후확률 θ_{il}^{obs}를 정의한다.

$$\theta_{il}^{obs} = \frac{\gamma_l(\boldsymbol{x}_i) \prod_{m \in obs_i} \prod_{k=1}^{r_m} \rho_{mk|l}^{I(y_{im}=k)}}{\sum_{j=1}^{C} \gamma_j(\boldsymbol{x}_i) \prod_{m \in obs_i} \prod_{k=1}^{r_m} \rho_{mk|j}^{I(y_{im}=k)}}$$

위의 식에서 obs_i는 i번째 응답자가 응답한 문항을 말한다.

이처럼 잠재 층 회귀분석에서는 응답에 대한 결측치를 제거하지 않고도 모수의 추정이 가능하기 때문에 위의 전처리 과정에서 삶의 만족도 변수(V1646, V1647, V1648, V1650, V1653)와 약물 사용 변수(V1102, V1216, V1254, V1712)의 결측치는 제거하지 않은 것이다. 또한 위의 식을 통해 모든 문항에 응답하지 않은 응답자의 정보는 사후확률 정의에 사용될 수 없다는 것을 알 수 있는데, 연구에 사용된 데이터 중 삶의 만족도 변수와 약물 사용 변수

모두 응답하지 않은 응답자는 없었으므로 결측치 제거가 발생하지 않았다. 반면 공변량으로 사용되는 성별, 인종, 정치적 성향, 종교의 중요도 변수에 대해서는 결측치를 제거하였는데 이는 잠재 층 회귀분석에서 공변량에 대한 결측치는 다룰 수 없기 때문이다.

(3-4) 기본 잠재 층 모형과 잠재 층 회귀 모형의 관계

잠재 층 회귀 모형과 공변량이 없는 기본 잠재 층 모형의 관계는 잠재 층 분석의 'Marginalization property'라는 개념으로 설명할 수 있다. 이는 공변량의 분포 공간에서 잠재 층 회귀모형의 평균을 계산하면 기본 잠재 층 모형을 얻을 수 있다는 것이다. 잠재 층 회귀 모형의 marginalization property를 수식으로 설명하면 다음과 같다.

$$
\begin{aligned}
P(\boldsymbol{Y}_i = \boldsymbol{y}_i) &= \int \sum_{l=1}^{C} \gamma_l(\boldsymbol{x}_i) \prod_{m=1}^{M} \prod_{k=1}^{r_m} \rho_{mk|l}^{I(y_{im}=k)} dF(\boldsymbol{x}_i) \\
&= \sum_{l=1}^{C} \left(\int \gamma_l(\boldsymbol{x}_i) dF(\boldsymbol{x}_i) \right) \prod_{m=1}^{M} \prod_{k=1}^{r_m} \rho_{mk|l}^{I(y_{im}=k)} \\
&= \sum_{l=1}^{C} \gamma_l^* \prod_{m=1}^{M} \prod_{k=1}^{r_m} \rho_{mk|l}^{I(y_{im}=k)}
\end{aligned}
$$

(이때, γ_l^* 는 잠재 층 회귀 모형에서 l번째 잠재 층에 속할 평균적 확률)

잠재 층 회귀 모형에서 연구자가 선택해야 할 사항은 잠재 층의 개수이다. 연구자는 잠재 층의 개수를 2개에서부터 점차 늘려가며 여러 잠재 층 모형을 적합한 후, 각 모델을 평가하면서 주어진 데이터에 몇 개의 잠재 층이 적절한지 결정할 수 있다. 이때, marginalization property를 사용하면 잠재 층 개수 선택에 도움이 될 수 있다. 연구자는 모형 선택 과정에서 잠재 층 회귀 모형이 아닌 공변량을 제외한 잠재 층 모형을 적합함으로써 적절한 잠재 층의 개수를 선택할 수 있다. 이는 모형에서 추정해야 할 모수의 개수를 획기적으로 줄일 수 있으므로 모형 적합에 걸리는 시간을 단축할 수 있고 모수가 많아서 발생하는 identification problem의 발생 가능성을 낮출 수 있다.

(3-5) 모형 평가 및 선택

본 연구에서는 앞서 설명한 것과 같이 공변량을 제외한 잠재 층 모형을 적합 후 각 모형에 대한 평가를 바탕으로 잠재 층의 개수를 선택했다. 모형의 평가는 절대적 평가와 상대적 평가로 나눌 수 있다.

먼저 절대적 평가의 기준으로는 identification problem의 유무와 우도비 검정통계량 G^2가 있다. 이론적으로 잠재 층 회귀 모형이 identifiable하기 위한 조건은 모형의 자유도가 1 이상이라는 것이다. 하지만 실제로는 자유도가 1 이상이더라도 identification problem이 종종 발생한다. Identification problem을 쉽게 할 수 있는 한 가지 방법은 모형 적합 시 여러 개의 initial value를 사용해 모수를 추정하는 것이다. 만약 잠재 층 회귀분석은 모수 추정에 있어서 최적화 알고리즘을 사용하기 때문에 만약 모형이 불안정하다면 initial value를 달리했을 때 모수 추정 및 모형 적합 결과가 매번 달라질 것이다. 본 연구에서는 모형의 identification problem을 확인하기 위해 각 잠재 층 개수마다 서로 다른 50개의 initial value를 사용해 모형을 적합한 후 initial value에 따라 로그 가능도(log likelihood) 값이 얼마나 달라지는지 살펴보았다. 또 다른 모델의 절대적 평가 기준인 우도비 검정통계량 G^2는 모델이 주어진 데이터에 얼마나 적합한지를 나타낸다. G^2를 사용한 모형 적합도 검정은 G^2

검정 통계량이 카이제곱 분포를 따른다는 사실을 사용한 검정이다. 하지만 잠재 층 회귀 모형의 자유도가 너무 클 경우 카이제곱 분포로 근사가 잘 되지 않아 모형 적합도 검정을 시행하기 어려워진다. 따라서 본 연구에서는 붓스트랩 G^2를 사용한 모형 적합도 검정을 실시하였다. 붓스트랩 G^2를 사용한 모형 적합도 검정 방법은 다음과 같다. : 1) 잠재 층 모형을 적합해 G^2값을 얻는다; 2) 1)에서 적합한 모형의 모수 추정치로 B개의 붓스트랩 응답 데이터를 생성한 후 붓스트랩 데이터를 사용해 잠재 층 모형을 적합한다; 3) B개의 붓스트랩 G^2를 얻고, 그 중에서 1)에서 얻은 G^2값보다 큰 붓스트랩 G^2의 비율을 계산한다(=p-value). 붓스트랩 G^2를 사용한 모형 적합도 검정에서 p-value가 0.05보다 크다면 해당 모형이 데이터에 적합하다고 판단한다.

두 번째로 모델의 상대적 평가 기준은 AIC와 BIC가 있다. 잠재 층의 개수가 서로 다른 잠재 층 모형은 내포 관계가 아니므로 G^2를 사용한 모형 비교가 불가능하다. 따라서 여러 모델을 비교하기 위해서는 내포 관계와 상관없는 AIC와 BIC를 사용한다. AIC(또는 BIC)의 값이 작은 모형을 선택한다.

본 연구에서는 2개~6개의 잠재 층을 사용한 모형을 적합해보고 절대적·상대적 평가 기준과 더불어 모형의 해석도 고려해 최종적으로 잠재 층의 개수를 선택했다.

(4) 코크란-맨텔-핸젤 검정

코크란-맨텔-핸젤 검정은 여러 개의 그룹이 있을 때 그룹의 효과를 제어한 후 두 변수 간의 연관성을 살펴볼 수 있는 검정이다. 본 연구에서는 삶의 만족도/약물 사용 잠재 층의 효과를 제어한 후 성별, 인종, 정치적 성향, 종교의 중요도 사이의 연관성을 살펴보기 위해 코크란-맨텔-핸젤 검정이 사용되었다.

본 연구에서는 이러한 분석방법론을 사용하기 위해 통계 분석 프로그램 R을 사용했다. 카이제곱 검정을 위해서는 stat 패키지의 chisq.test 함수를 사용했다. 로그-선형 모형을 적합하기 위해서는 stat 패키지의 glm 함수를 사용했다. 잠재 층 회귀 모형을 적합하기 위해서는 poLCA 패키지의 poLCA 함수를 사용했다. 마지막으로 코크란-맨텔-핸젤 검정을 위해서는 stat 패키지의 mantelhaen.test 함수를 사용했다.

2-2. 공변량 사이의 연관성

2-2절에서는 카이제곱 독립성 검정과 로그-선형 모형을 통해 성별, 인종, 정치적 성향, 종교의 중요도 4가지 변수 사이의 연관성을 알아보고자 했다.

2-2-1. 카이제곱 독립성 검정

카이제곱 독립성 검정은 두 변수의 연관성을 검정하는 것이므로 4개 변수에서 가능한 총 6가지 조합에 대해 카이제곱 독립성 검정을 실시했다. 검정에 대한 귀무가설은 두 변수가 서로 독립이라는 것이다.

Table 3.을 통해 카이제곱 독립성 검정 결과를 살펴볼 수 있다. 성별과 인종, 인종과 정치적 성향의 조합에 대해서는 p-value가 0.05보다 커 유의수준 0.05에서 두 변수가 서로 독립이라는 귀무가설을 기각할 수 없었다. 그 외의 변수들의 조합에 대해서는 p-value가 0.05보다 작았으므로 귀무가설을 기각하게 되었다.

변수 조합	자유도	X^2	p-value
(성별, 인종)	1	0.129	0.720
(성별, 정치적 성향)	3	11.415	0.010
(성별, 종교의 중요도)	1	18.854	< 0.001
(인종, 정치적 성향)	3	3.089	0.379
(인종, 종교의 중요도)	1	36.389	< 0.001
(정치적 성향, 종교의 중요도)	3	58.497	< 0.001

Table 3. 성별, 인종, 정치적 성향, 종교의 중요도 4개 변수에 대한 카이제곱 독립성 검정 결과

성별과 인종의 경우 상식적으로 독립적인 변수이므로 분석에 큰 의미는 없다. 분석결과에서 주목해야 할 것은 다른 변수들끼리는 연관성이 보이는 반면 인종과 정치적 성향은 서로 독립이라는 것이다.

2-2-2. 로그-선형 모형

카이제곱 독립성 검정 후 성별, 인종, 정치적 성향, 종교의 중요도 4개의 변수의 연관성을 구체적으로 파악하기 위해서 로그-선형 모형을 사용했다. 모형 표기의 편의를 위해서 성별=X, 인종=Y, 정치적 성향=Z, 종교의 중요도=W로 표기하도록 하겠다.

로그-선형 모형 적합 과정에서는 적절하다고 선택된 M_1 모형과 M_1과 내포 관계에 있는 축소(reduced) 모형 M_2을 우도비 검정을 통해 비교했다. 우도비 검정의 귀무가설은 M_2가 데이터에 적합하다는 것이다. 첫 번째 M_1로 포화모형 ($XYZW$)을 설정했으며 그 후 우도비 검정에서 귀무가설을 기각할 수 없는 축소모형 M_2가 M_1이 된다.

Table 4.에는 로그-선형 모형 적합 결과가 나타나 있다. 우도비 검정 과정에서 축소모형 M_2가 적절하다고 판단될 경우 해당 모델을 M_1로 설정하는 부분이 빨간 글씨로 적혀있다. 가장 먼저 M_1로 설정된 포화모형과 비교할 축소 모형으로 (XZW, YZW) 모형을 적합했다. 해당 모형은 포화모형에서 성별(=X)과 인종(=Y)의 교호작용이 포함되는 모든 항을 제거한 모형이다. 이는 성별과 인종은 상식적으로 독립적인 변수이기에 둘의 교호작용을 제거하는 것이 합리적이라고 판단했기 때문이다. 포화모형과 (XZW, YZW) 모형을 비교하는 우도비 검정 결과, p-value = 0.874로 유의수준 0.05에서 귀무가설을 기각할 수 없었다. 즉, 우리의 데이터에는 4차 교호작용이 유의하지 않았다.

(XZW, YZW) 모형이 M_1으로 설정된 후에는 3차 교호작용을 하나씩 제거해보며 우도비 검정을 실시했다. 그 결과 XZW(성별*정치적 성향*종교의 중요도) 항을 제거한 모형은 적합하지 않았지만, YZW(인종*정치적 성향*종교의 중요도) 항을 제거한 모형은 p-value = 0.081로 유의수준 0.05에서 귀무가설을 기각할 수 없었다. 우리의 데이터에는 인종, 정치적 성향, 종교의 중요도의 3차 교호작용이 유의하지 않았다.

(XZW, YZW) 모형에서 YZW 항이 제거된 (XZW, YZ, YW) 모형이 M_1으로 설정된 후에는 먼저 나머지 3차 교호작용 항인 XZW 항을 제거해 우도비 검정을 실시했다. 하지만 3차 교호작용 항이 아예 없는 모델은 우리 데이터에 적합하지 않았다. 다음으로는 (XZW, YZ, YW) 모형에서 YZ(인종*정치적 성향) 항을 제거한 모형을 적합해 보았다. YZ

항을 제거한 이유는 앞선 카이제곱 독립성 검정에서 인종과 정치적 성향이 독립이라는 결과를 얻었기 때문이다. 우도비 검정 결과, p-value = 0.120으로 유의수준 0.05에서 귀무가설을 기각할 수 없었다. 즉, 우리의 데이터에는 인종과 정치적 성향의 교호작용이 유의하지 않았다.

(XZW, YZ, YW) 모형에서 YZ 항을 제거한 (XZW, YW) 모형이 M_1으로 설정된 후에는 아직 남아있는 3차 교호작용을 다시 한번 제거한 모형을 적합해 보았다. 3차 교호작용이 있는 모형은 해석의 어려움이 있기 때문에 최대한 단순한 형태의 모형을 적합하고자 한 것이다. 하지만 우도비 검정 결과 p-value = 0.035로 유의수준 0.05에서 3차 교호작용 항을 제거한 (XZ, XW, YW, ZW) 모형이 적합하다는 귀무가설을 기각하게 되었다.

그 외에 (XZW, YW) 모형보다 간단한 모형을 적합했을 때는 모형의 적합도가 현저히 떨어졌다. 따라서 우리의 데이터에 적절한 로그-선형 모형으로는 최종적으로 (XZW, YW) 모형이 선택되었다.

[1] $G^2(M_2|M_1) = G^2(M_2) - G^2(M_1)$, 자유도 = M_2의 자유도 - M_1의 자유도

로그-선형 모형	$G^2(M_2\|M_1)$ (자유도) [1]	p-value
$(XYZW)$	-	-
(XZW, YZW)	3.810 (8)	0.874
M_1 : (XZW, YZW)		
(YZW, XZ, XW)	8.603 (3)	0.035
(XZW, YZ, YW)	6.741 (3)	0.081
M_1 : (XZW, YZ, YW)		
(XZ, XW, YZ, YW, ZW)	8.603 (3)	0.035
(XZW, YW)	5.835 (3)	0.120
M_1 : (XZW, YW)		
(XZ, XW, YW, ZW)	8.603 (3)	0.035

Table 4. 성별(=X), 인종(=Y), 정치적 성향(=Z), 종교의 중요도(=W)의 로그-선형 모형 적합 결과

Table 5.에는 최종적으로 선택된 모형인 (XZW, YW) 모형이 정리되어있다. 모형의 모수 추정이 가능하도록 Table 5.에 나타난 모수를 제외한 나머지 모수들은 0으로 고정되었다.

로그-선형 모형에서 중요한 부분은 주효과가 아닌 교호작용 부분이므로 교호작용의 모수 추정치 및 유의성을 위주로 모형을 해석하고자 한다. 먼저 (XZW, YW) 모형에는 인종(=Y)과 정치적 성향(=Z)의 교호작용이 전혀 포함되어있지 않다. 이는 앞서 진행한 카이제곱 독립성 검정의 결과가 로그-선형 모형에도 반영되는 것이라고 할 수 있다. 다음으로 2차 교호작용과 3차 교호작용을 비교해보자면 2차 교호작용은 대부분 유의한 것에 비해서 3차 교호작용은 모두 유의하지 않았다. 모형의 설명력을 높이기 위해 3차 교호작용이 포함되었지만 모형을

해석할 때에는 3차 교호작용을 고려하지 않는 것이 해석의 편의성에 도움이 될 것이다.

2차 교호작용에 집중해 모형을 해석해본다면 다음과 같다. 먼저 남성과 정치적 성향이 진보, 중립, 없음/모름의 교호작용 항은 모두 음수로 나타났다. 이를 통해 남성은 정치적 성향이 보수적일 경향이 있음을 알 수 있다. 남성과 종교가 중요하지 않음의 교호작용 항은 양수로 나타났다. 이를 통해 남성은 종교가 중요하지 않은 경향이 있음을 알 수 있다. 백인과 종교가 중요하지 않음의 교호작용 역시 양수로 나타났다. 이를 통해 백인은 종교가 중요하지 않은 경향이 있음을 알 수 있다. 마지막으로 종교가 중요하지 않음과 정치적 성향이 진보, 중립, 없음/모름의 교호작용이 양수로 나타났다. 이를 통해 다른 정치적 성향에 비해 보수적 정치 성향을 가진 사람이 종교의 중요도를 높이 평가하는 경향이 있음을 알 수 있다.

[1] * 표시는 유의수준 0.05에서 유의한 효과를 나타냄

효과	추정치	p-value[1]
Intercept	2.890	< 0.001 *
성별(남성)	0.011	0.941
인종(백인)	1.401	< 0.001 *
정치적 성향(진보)	-0.116	0.446
정치적 성향(중립)	0.372	0.006 *
정치적 성향(없음/모름)	0.677	< 0.001 *
종교의 중요도(중요하지 않음)	-2.431	< 0.001 *
성별(남성)*정치적 성향(진보)	-0.435	0.059
성별(남성)*정치적 성향(중립)	-0.463	0.023 *
성별(남성)*정치적 성향(없음/모름)	-0.428	0.024 *
성별(남성)*종교의 중요도(중요하지 않음)	0.829	0.007 *
인종(백인)*종교의 중요도(중요하지 않음)	0.998	< 0.001 *
정치적 성향(진보)*종교의 중요도(중요하지 않음)	1.638	< 0.001 *
정치적 성향(중립)*종교의 중요도(중요하지 않음)	0.490	0.109
정치적 성향(없음/모름)*종교의 중요도(중요하지 않음)	1.284	< 0.001 *
성별(남성)*정치적 성향(진보)*종교의 중요도(중요하지 않음)	-0.440	0.259
성별(남성)*정치적 성향(중립)*종교의 중요도(중요하지 않음)	0.223	0.566
성별(남성)*정치적 성향(없음/모름)*종교의 중요도(중요하지 않음)	-0.555	0.120

Table 5. 로그-선형 분석의 최종 모형 : (XZW, YW) 모형의 모수 추정치

2-3. 잠재 층 회귀 모형

2-3절에서는 잠재 층 회귀 모형을 사용해 MTF 연구의 대상인 12학년 청소년들을 몇 개의 집단으로 나눠보고자 한다. 본 연구에서는 잠재 층 회귀 분석을 통해 청소년의 삶의 만족도에 대한 잠재 층과 약물 사용 행태에 대한 잠재 층을 분석하였으며 이때 성별(V1150), 인종(V1151), 정치적 성향(V1167), 종교의 중요도(V1170)가 공변량으로 사용되었다.

2-3-1. 청소년의 삶의 만족도

첫 번째로 삶의 만족도에 대한 청소년의 잠재 층을 분석했다. 삶의 만족도 분석에 사용된 주요 변수는 V1646 (SAT OWN FRIENDS), V1647(SAT GT ALNG PRN), V1648(SAT

YOURSELF), V1650(SAT TIME FR THG), V1653(SAT GOVT OPRTNG)의 5개 변수이다. 편의를 위해 V1646는 친구 만족도, V1647는 부모 만족도, V1648는 자기 만족도, V1650는 자유 시간 만족도, 마지막으로 V1653는 정부 만족도라고 칭한다.

Table 6.에는 5개의 삶의 만족도 문항에 대한 응답 비율이 나타나있다. 친구 만족도, 부모 만족도, 자기 만족도는 3(만족)의 응답 비율이 가장 높으며 자유 시간 만족도, 정부 만족도는 2(중간)의 응답 비율이 가장 높게 나타났다. 5개의 문항 중 3(만족)의 응답 비율은 친구 만족도에서 가장 높고 정부 만족도에서 가장 낮다. 5개의 문항에서 결측치는 매우 적은 비율로 나타났다.

	1 (불만족)	2 (중간)	3 (만족)	결측치
친구 만족도	0.012	0.246	0.740	0.001
부모 만족도	0.080	0.383	0.535	0.001
자기 만족도	0.058	0.362	0.572	0.009
자유 시간 만족도	0.155	0.525	0.317	0.002
정부 만족도	0.231	0.626	0.140	0.003

Table 6. 삶의 만족도 문항의 각 응답 별 응답 비율

우리의 데이터에 적절한 잠재 층 개수를 선택하기 위해 2-1-2 분석방법론에서 언급한 바와 같이 2개~6개의 잠재 층을 사용한 공변량이 없는 잠재 층 모형을 비교·평가했다.

먼저 identification problem을 살펴보기 위해 각 모형마다 50개의 서로 다른 initial value를 사용했을 때 로그 가능도 값이 얼마나 달라지는지 살펴보았다.

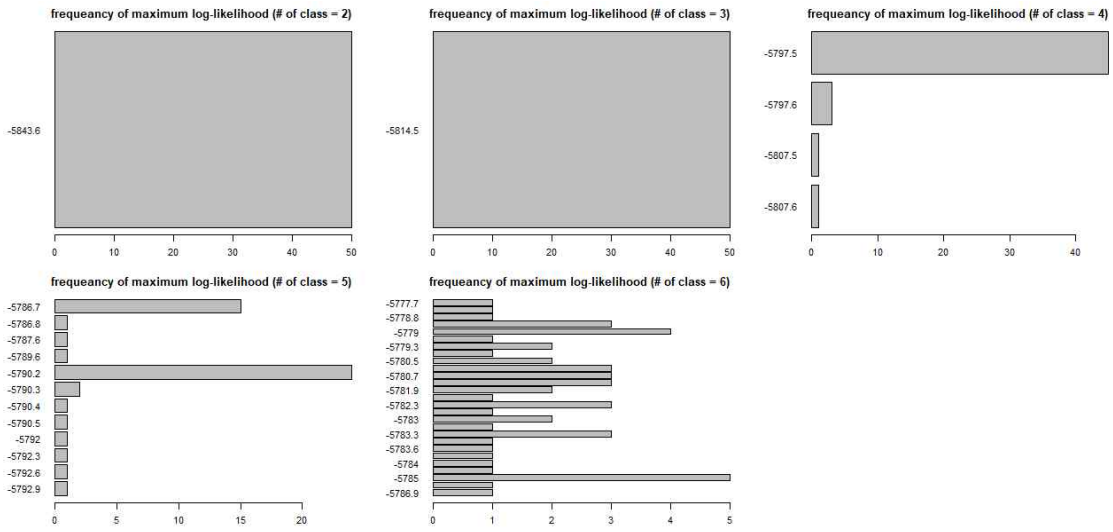

Figure 1. 삶의 만족도 잠재 층 모형에서 잠재 층 개수 별 최대 로그 가능도(maximum likelihood) 값의 히스토그램

Figure 1.을 통해 각 모형의 identification problem을 살펴볼 수 있다. 잠재 층 개수가 2개 또는 3개인 모형은 50개의 initial value에서 모두 같은 최대 로그 가능도 값으로

수렴했다. 잠재 층의 개수가 4개인 경우 50번의 모형 적합에서 90%의 모형이 최대 로그 가능도 값 -5797.5로 수렴했다. 잠재 층이 2개~4개인 경우 큰 identification problem이 발생하지 않는다고 할 수 있다. 하지만 잠재 층이 5개 이상이 되면 identification problem이 생겨나는데, 잠재 층이 5개인 모형은 50번의 적합 중 48%가 최대 로그 가능도 값 -5790.2로 수렴했으며 잠재 층이 6개인 모형은 오직 10%만이 최대 로그 가능도 값 -5785로 수렴했다. 잠재 층이 5개~6개인 경우 initial value에 따라 최대 로그 가능도 값이 크게 달라지기 때문에 모형이 불안정하다고 판단할 수 있다.

다음으로는 절대적 평가 기준인 G^2의 붓스트랩 p-value와 상대적 평가 기준인 AIC, BIC를 살펴보았다.

number of class	bootstrap p-value	AIC	BIC
2	< 0.01	11729.28	11840.04
3	0.06	11693.04	11861.82
4	0.19	11681.09	11907.90
5	0.41	11681.34	11966.16
6	0.48	11685.38	12028.22

Table 7. 삶의 만족도 잠재 층 모형에서 잠재 층 개수 별 G^2의 붓스트랩 p-value (100회의 붓스트랩), AIC, BIC

Table 7.을 통해 그 결과를 살펴볼 수 있다. 먼저 G^2의 붓스트랩 p-value를 확인해보면 잠재 층 개수가 2인 경우는 p-value가 0.05보다 작으므로 주어진 데이터에는 잠재 층 개수가 2개인 모형은 적합하지 않다는 것을 알 수 있다. 모형의 상대적 평가에서는 AIC를 기준으로 본다면 잠재 층이 4개인 모형이 가장 좋고 BIC를 기준으로 본다면 잠재 층이 2개인 모형이 가장 좋다.

본 연구에서는 identification problem, G^2의 붓스트랩 p-value, AIC, BIC를 종합적으로 평가해보았을 때 잠재 층이 3개 또는 4개인 모형이 가장 적합하다고 판단했다. 두 모형의 해석을 비교해봤을 때는 잠재 층이 4개인 모형의 해석이 직관적이고 쉬웠다. 따라서 최종적으로 삶의 만족도에 대한 잠재 층 회귀 모형은 4개의 잠재 층을 가진 모형으로 선택되었다.

Table 8.에는 잠재 층이 삶의 만족도 4개인 잠재 층 회귀 모형에서의 모수 γ^*, ρ의 추정 결과가 나타나 있다. 모수 γ^*는 잠재 층 회귀 모형에서 평균적으로 해당 잠재 층에 속할 확률을 말하며 모수 ρ는 잠재 층이 주어졌을 때 각 문항에 대한 응답 별 확률을 말한다. 우선 모수 ρ를 살펴보면서 4개의 잠재 층에 대한 해석을 진행하였다. 첫 번째 잠재 층은 친구 만족도에서는 3(만족)의 확률이 가장 높고 나머지 4개의 문항에서는 모두 2(중간)의 확률이 가장 높다. 이를 통해 첫 번째 잠재 층은 '친구를 좋아하는 평균 집단'이라고 해석할 수 있다. 두 번째 잠재 층은 정부 만족도에서는 1(불만족)로 응답할 확률이 가장 높았고 나머지 4개의 문항에서는 모두 2(중간)로 응답할 확률이 가장 높았다. 첫 번째 잠재 층과 비슷한 결과이지만 두 번째 잠재 층이 첫 번째 잠재 층보다 친구 만족도와 정부 만족도가 떨어진다는 차이점이 있다. 이를 통해 두 번째 잠재 층은 '정부에 불만족 집단'이라고 해석할 수 있다. 세 번째 잠재 층은 정부 만족도에서는 2(중간)의 확률이 가장 높았고 나머지 4개의 문항에서는 모두 3(만족)의 확률이 가장 높았다. 이를 통해 세 번째 잠재 층은 '만족 집단'이라고 해석할 수 있다. 마지막으로 네 번째 잠재 층은 친구 만족도, 부모 만족도, 자기 만족도에서는 3(만족)의

확률이 가장 높았고 자유 시간 만족도와 정부 만족도에서는 2(중간)의 확률이 가장 높았다. 비슷해 보이는 세 번째 잠재 층과 네 번째 잠재 층을 비교해보며 네 번째 잠재 층에 대한 해석을 진행하겠다. 세 번째 잠재 층과 네 번째 잠재 층은 모두 정부 만족도 문항에서 2(중간)으로 응답할 확률이 약 59%로 비슷하지만, 세 번째 잠재 층은 3(만족)으로 응답할 확률이 약 34%로 두 번째로 높은 반면 네 번째 잠재 층은 1(불만족)으로 응답할 확률이 약 40%로 두 번째로 높았다. 이를 통해 세 번째 잠재 층보다 네 번째 잠재 층에서 정부 만족도가 떨어진다는 것을 알 수 있다. 특히 네 번째 잠재 층의 정부 만족도 문항에서 1(불만족)로 응답할 확률은 4개의 잠재 층 전체와 비교해보아도 꽤 높은 응답 확률이다. 따라서 네 번째 잠재 층은 '대체로 만족, 정부 불만족 집단'이라고 해석할 수 있다.

- 삶의 만족도 잠재 층 회귀 모형의 4가지 잠재 층 해석 -

잠재 층 1 : 친구를 좋아하는 평균 집단
잠재 층 2 : 정부 불만족 집단
잠재 층 3 : 만족 집단
잠재 층 4 : 대체로 만족, 정부 불만족 집단

[1)] 진한 글씨는 각 잠재 층에서 문항별로 가장 높은 확률의 응답을 나타냄

		잠재 층			
		1	2	3	4
모수 γ^*의 추정치		0.232	0.157	0.330	0.281
문항 / 응답		모수 ρ의 추정치			
친구 만족도	1	0.000	0.060	0.006	0.004
	2	0.400	**0.498**	0.049	0.211
	3	**0.600**[1)]	0.442	**0.946**	**0.785**
부모 만족도	1	0.082	0.286	0.000	0.059
	2	**0.627**	**0.560**	0.114	0.401
	3	0.291	0.154	**0.886**	**0.540**
자기 만족도	1	0.033	0.309	0.005	0.000
	2	**0.777**	**0.629**	0.128	0.154
	3	0.190	0.062	**0.866**	**0.845**
자유 시간 만족도	1	0.148	0.371	0.061	0.152
	2	**0.691**	**0.565**	0.413	**0.502**
	3	0.161	0.065	**0.526**	0.346
정부 만족도	1	0.057	**0.534**	0.063	0.405
	2	**0.858**	0.425	**0.592**	**0.594**
	3	0.086	0.041	0.344	0.002

Table 8. 잠재 층이 4개인 삶의 만족도 잠재 층 회귀 모형에서 모수 γ^*, ρ의 추정 결과

모수 γ^*의 추정 결과로는 각 잠재 층의 분포(prevalance)를 추정할 수 있다. γ^*의 추정치를 살펴보면 세 번째와 네 번째 잠재 층의 비율이 비슷하게 가장 많고, 다음으로는 첫 번째 잠재 층이, 마지막으로는 두 번째 잠재 층의 비율이 가장 적게 나타났다. 이를 통해 대부분의 청소년의 전반적 삶의 만족도는 양호하다고 판단된다. 전반적으로 삶의 만족도가 떨어지고 특히 정부에 대한 불만족이 큰 두 번째 잠재 층(약 15%)에 대한 관심이 필요할 것으로 보인다.

[1)] * 표시는 유의수준 0.05에서 유의한 변수를 나타냄

2/1	β 추정치	오즈비	p-value[1)]
Intercept	-0.533	0.587	0.220
성별(여성)	-0.073	0.930	0.816
인종(흑인)	1.676	5.346	0.031 *
정치적 성향(보수)	-0.645	0.524	0.195
정치적 성향(중립)	-0.817	0.442	0.055
정치적 성향(진보)	2.001	7.396	0.001 *
종교의 중요도(중요함)	-0.167	0.846	0.602
3/1	β 추정치	오즈비	p-value
Intercept	0.342	1.408	0.228
성별(여성)	-0.784	0.457	< 0.001 *
인종(흑인)	0.267	1.305	0.722
정치적 성향(보수)	0.480	1.617	0.088
정치적 성향(중립)	-0.396	0.673	0.121
정치적 성향(진보)	0.304	1.356	0.614
종교의 중요도(중요함)	0.647	1.910	0.004 *
4/1	β 추정치	오즈비	p-value
Intercept	0.055	1.057	0.889
성별(여성)	-0.545	0.580	0.086
인종(흑인)	3.170	23.815	< 0.001 *
정치적 성향(보수)	-0.852	0.427	0.111
정치적 성향(중립)	-0.550	0.577	0.153
정치적 성향(진보)	1.965	7.138	0.001 *
종교의 중요도(중요함)	-0.371	0.690	0.263

Table 9. 잠재 층이 4개인 삶의 만족도 잠재 층 회귀 모형의 잠재 층 별 오즈 비 (base line 잠재 층 : 첫 번째 잠재 층)

Table 9.를 통해 삶의 만족도 잠재 층 회귀 모형을 해석하면 다음과 같다. 먼저 두 번째 잠재 층('정부 불만족 집단')과 첫 번째 잠재 층('친구를 좋아하는 평균 집단')의 오즈비이다. 흑인의 경우 백인에 비해 '친구를 좋아하는 평균 집단'보다 '정부 불만족 집단'에 속할 확률이 높다. 또한 정치적 성향이 진보인 경우 그렇지 않은 경우에 비해 '친구를 좋아하는 평균

집단'보다 '정부 불만족 집단'에 속할 확률이 높다. 정치적 성향이 중립인 경우도 마찬가지이다. 다음으로는 세 번째 잠재 층('만족 집단')과 첫 번째 잠재 층('친구를 좋아하는 평균 집단')의 오즈비이다. 여성인 경우 남성에 비해 '만족 집단'보다 '친구를 좋아하는 평균 집단'에 속할 확률이 높다. 정치적 성향이 보수인 경우 그렇지 않은 경우에 비해 '친구를 좋아하는 평균 집단'보다 '만족 집단'에 속할 확률이 높다. 종교가 중요한 경우 중요하지 않은 경우에 비해 '친구를 좋아하는 평균 집단'보다 '만족 집단'에 속할 확률이 높다. 마지막으로 네 번째 잠재 층('대체로 만족, 정부 불만족 집단')과 첫 번째 잠재 층('친구를 좋아하는 평균 집단')의 오즈비이다. 여성인 경우 남성에 비해 '대체로 만족, 정부 불만족 집단'보다 '친구를 좋아하는 평균 집단'에 속할 확률이 크다. 흑인인 경우 백인에 비해 '친구를 좋아하는 평균 집단'보다 '대체로 만족, 정부 불만족 집단'에 속할 확률이 높다. 정치적 성향이 진보인 경우 그렇지 않은 경우에 비해 '친구를 좋아하는 평균 집단'보다 '대체로 만족, 정부 불만족 집단'에 속할 확률이 높다.

이러한 모형 해석을 통해 다음과 같은 시사점을 발견할 수 있다. 여성의 경우 남성에 비해 만족도가 평균인 집단에 속하는 경향이 있다. 종교를 중요하게 여기는 것이 삶의 만족도에 긍정적인 영향을 줄 수 있다. 흑인의 경우 정부에 불만족하는 경향이 있다. 정치적 성향이 보수인 경우 정부에 만족하는 경향이 있고 진보 또는 중립의 경우 정부에 불만족하는 경향이 있다. 여기서 정치적 성향과 관련한 결과는 당연한 것으로 생각되는데, 그 이유는 해당 연구가 이루어진 2004년 당시 미국의 대통령은 보수 정당이 공화당 소속의 조지 W.부시 전 대통령이었기 때문이다. 모형 해석에서 주목할 만한 점은 흑인의 정부 불만족 경향이다. 앞선 카이제곱 검정과 로그-선형 모형에서는 인종과 정치적 성향이 독립이라는 결과를 얻을 수 있었다. 하지만 삶의 만족도 잠재 층 회귀 모형을 분석한 결과 인종과 정치적 성향 사이의 연관성이 있을 것으로 의심된다.

2-3-2. 청소년의 약물 사용 행태

두 번째로 약물 사용 대한 청소년의 잠재 층을 분석했다. 약물 사용 행태 분석에 사용된 주요 변수는 V1102 (#CIGS SMKD/30DA), V1216(#X ALC/30D SIPS), V1254(#X MARJ/LAST 30D), V1712(#X DIETPILL/30D)의 4개 변수이다. 편의를 위해 V1102는 담배, V1216은 술, V1254는 마리화나, 마지막으로 V1712는 다이어트 약이라고 칭한다.

	1 (사용 안함)	2 (약간 사용)	3 (많이 사용)	결측치
담배	0.731	0.215	0.033	0.021
술	0.479	0.428	0.070	0.023
마리화나	0.773	0.125	0.078	0.025
다이어트 약	0.920	0.040	0.017	0.023

Table 10. 약물 사용 문항의 각 응답 별 응답 비율

Table 10.에는 4개의 약물 사용 문항에 대한 응답 비율이 나타나 있다. 모든 문항에서 1(사용 안함), 2(약간 사용), 3(많이 사용)의 순서로 응답 비율이 높았다. 3(많이 사용)의 응답 비율이 모든 문항에서 10% 이내인 것을 보아 청소년의 약물 사용 실태가 심각한 상태는 아니라는 것을 알 수 있다. 응답 비율을 살펴보면 청소년의 약물 사용은 술, 담배, 마리화나,

다이어트 약의 순서로 잦음을 알 수 있는데, 술, 담배, 그리고 마리화나는 약물의 강도에 따라 약물 사용 비율이 줄어드는 것을 확인할 수 있다. 다이어트 약의 사용 비율은 매우 낮은데, 이는 다이어트 약의 강도가 강해서라기 보다는 다이어트 약을 불법 처방받아 섭취하는 상황 자체가 잘 일어나지 않는다고 해석하는 것이 적절할 것이다. 4개의 문항에서 결측치의 비율은 삶의 만족도 문항보다는 높은 편이지만 여전히 2% 전후의 낮은 결측치 비율을 보이고 있다.

삶의 만족도에 대한 잠재 층 분석에서와 같이 약물 사용에 대한 잠재 층 분석에서도 2개~6개의 잠재 층을 사용한 공변량이 없는 잠재 층 모형을 비교·평가했다.

먼저 identification problem을 살펴보기 위해 각 모형마다 50개의 서로 다른 initial value를 사용했을 때 로그 가능도 값이 얼마나 달라지는지 살펴보았다.

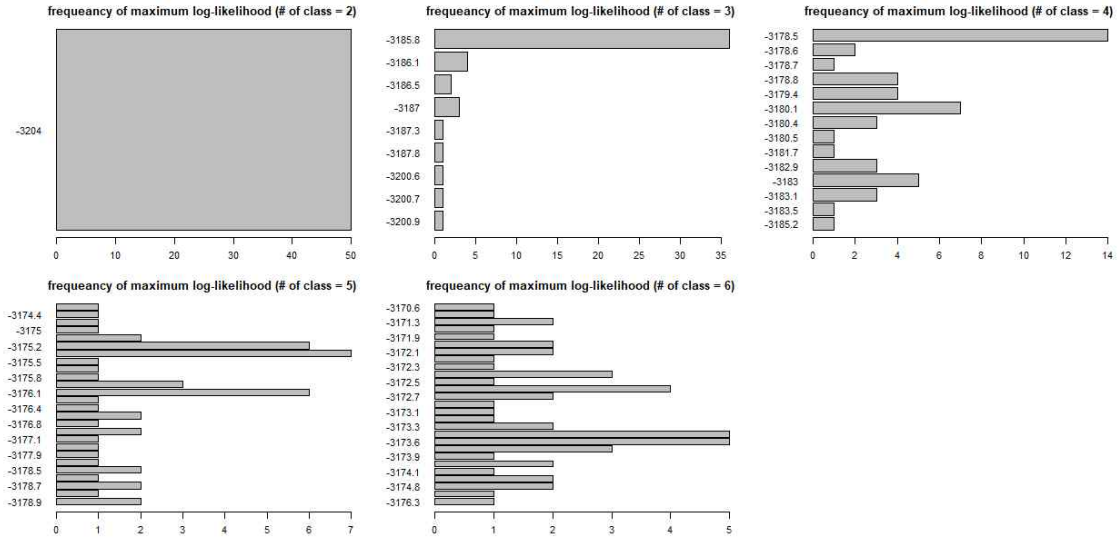

Figure 2. 약물 사용 잠재 층 모형에서 잠재 층 개수 별 최대 로그 가능도(maximum likelihood) 값의 히스토그램

Figure 2.을 통해 각 모형의 identification problem을 살펴볼 수 있다. 잠재 층 개수가 2개인 모형은 50개의 initial value에서 모두 같은 최대 로그 가능도 값으로 수렴했다. 잠재 층의 개수가 3개인 경우 50번의 모형 적합에서 72%의 모형이 최대 로그 가능도 값 -3185.8로 수렴했다. 잠재 층이 2개~3개인 경우 큰 identification problem이 발생하지 않는다고 할 수 있다. 하지만 잠재 층이 4개 이상이 되면 identification problem이 발생하는 것을 확인할 수 있다. 잠재 층이 4개인 모형은 50번의 적합 중 28%가 최대 로그 가능도 값 -3178.5로 수렴했으며 잠재 층이 5개인 모형은 14%가 최대 로그 가능도 값 -3175.3로 수렴했다. 잠재 층이 6개인 모형은 50번의 적합 중 각각 10%가 최대 로그 가능도 값 -3173.6과 -3173.5로 수렴했다. 잠재 층이 4개~6개인 경우 initial value에 따라 최대 로그 가능도 값이 크게 달라지기 때문에 모형이 불안정하다고 판단할 수 있다.

다음으로는 절대적 평가 기준인 G^2의 붓스트랩 p-value와 상대적 평가 기준인 AIC, BIC를 살펴보았다. Table 11.을 통해 그 결과를 살펴볼 수 있다. 먼저 G^2의 붓스트랩 p-value를 확인해보면 잠재 층 개수가 2인 경우는 p-value가 0.05보다 작으므로 주어진 데이터에는 잠재 층 개수가 2개인 모형은 적합하지 않다는 것을 알 수 있다. 모형의 상대적 평가에서는 AIC를 기준으로 본다면 잠재 층이 3개인 모형이 가장 좋고 BIC를 기준으로 본다면 잠재 층이 2개인 모형이 가장 좋다.

number of class	bootstrap p-value	AIC	BIC
2	< 0.01	6441.998	6531.665
3	0.12	6423.589	6560.726
4	0.42	6426.933	6611.539
5	0.45	6435.967	6668.044
6	0.54	6447.277	6726.824

Table 11. 잠재 층 개수 별 G^2의 붓스트랩 p-value (100회의 붓스트랩), AIC, BIC

본 연구에서는 identification problem, G^2의 붓스트랩 p-value, AIC, BIC를 종합적으로 평가해보았을 때 잠재 층이 3개인 모형이 가장 적합하다고 판단했다. 따라서 최종적으로 약물 사용에 대한 잠재 층 회귀 모형은 3개의 잠재 층을 가진 모형으로 선택되었다.

Table 12.에는 잠재 층이 4개인 약물 사용 잠재 층 회귀 모형에서의 모수 γ^*, ρ의 추정 결과가 나타나 있다. 우선 모수 ρ를 살펴보면서 4개의 잠재 층에 대한 해석을 진행하였다. 첫 번째 잠재 층은 모든 약물 사용 문항에서 1(사용 안함)이라고 응답할 확률이 가장 높았다. 따라서 첫 번째 잠재 층은 '약물 비(非)사용 집단'으로 해석할 수 있다. 두 번째 잠재 층은 담배와 술 문항에서는 2(약간 사용)의 응답 확률이 가장 높았고 마리화나와 다이어트 약에서는 1(사용 안함)이라고 응답할 확률이 가장 높았다. 첫 번째 잠재 층과 두 번째 잠재 층의 마리화나 문항에 대한 응답 확률을 비교해보면, 두 잠재 층 모두 1(사용 안함)이라고 응답할 확률이 높았지만 첫 번째 잠재 층에서는 해당 확률이 약 98%인 반면 두 번째 잠재 층에서는 해당 확률이 약 62%로 첫 번째 잠재 층보다는 낮게 나타났다. 또한 두 번째 잠재 층이 마리화나를 2(약간 사용)이라고 답할 확률은 약 37%로 나타났는데 이는 무시할 수 없는 꽤 높은 확률이다. 따라서 두 번째 잠재 층은 첫 번째 잠재층 보다 담배, 술, 마리화나 사용량이 높다는 것을 알 수 있다. 또한 두 번째 잠재 층에서 주목해야할 것은 다이어트 약의 사용이다. 두 번째 잠재 층에서는 다이어트 약을 3(많이 사용)이라고 응답할 확률이 약 7%로 나타난 반면 나머지 두 잠재 층이 거의 0%에 가까운 응답 확률을 보여준다. 다이어트 약 문항에 대한 응답은 두 번째 잠재 층의 특징을 잘 나타내고 있다. 이를 통해 두 번째 잠재 층은 '다이어트 약 사용 집단'이라고 해석할 수 있다. 세 번째 잠재 층은 담배, 술 문항에서는 2(약간 사용)의 응답 확률이 가장 높았고 마리화나 문항에서는 3(많이 사용)의 응답 확률이 가장 높았다. 다이어트 약 문항에서는 1(사용 안함)의 응답 확률이 가장 높았다. 이를 통해 세 번째 잠재 층은 '마리화나 사용 집단'이라고 해석할 수 있다.

- 약물 사용 잠재 층 회귀 모형의 3가지 잠재 층 해석 -

잠재 층 1 : 약물을 비사용 집단
잠재 층 2 : 다이어트 약 집단
잠재 층 3 : 마리화나 집단

[1)] 진한 글씨는 각 잠재 층에서 문항별로 가장 높은 확률의 응답을 나타냄

		잠재 층		
		1	2	3
모수 γ^*의 추정치		0.636	0.201	0.1638
문항 / 응답		모수 ρ의 추정치		
담배	1	**0.937** [1)]	0.474	0.338
	2	0.055	**0.494**	**0.521**
	3	0.007	0.032	0.141
술	1	**0.713**	0.123	0.077
	2	0.272	**0.820**	**0.615**
	3	0.015	0.057	0.308
마리화나	1	**0.983**	**0.623**	0.248
	2	0.017	0.377	0.258
	3	0.000	0.000	**0.493**
다이어트 약	1	**0.976**	**0.847**	**0.924**
	2	0.019	0.079	0.076
	3	0.004	0.074	0.000

Table 12. 잠재 층이 3개인 약물 사용 잠재 층 회귀 모형에서 모수 γ^*, ρ의 추정 결과

모수 γ^*의 추정 결과로는 각 잠재 층의 분포(prevalance)를 추정할 수 있다. γ^*의 추정치를 살펴보면 첫 번째 잠재 층의 비율이 가장 많고, 다음으로는 두 번째 잠재 층이, 마지막으로는 세 번째 잠재 층의 비율이 가장 적게 나타났다. 이를 통해 대부분 청소년의 약물 사용 현황은 양호하다고 판단된다. 하지만 다이어트 약 집단과 마리화나 집단의 비율이 각 16%와 20%로 적지 않은 비율로 나타났다. 따라서 해당 두 집단에 대한 약물 사용 지도가 필요할 것으로 보인다.

다음으로는 모형 속 4개의 공변량에 대한 회귀계수 β를 추정하고 이를 통해 잠재 층 별로 오즈 비를 계산하고자 한다. 약물 사용 잠재 층 회귀 모형에서는 첫 번째 잠재 층을 base line으로 선택했다.

Table 13.을 통해 삶의 만족도 잠재 층 회귀 모형을 해석하면 다음과 같다. 먼저 두 번째 잠재 층(‘다이어트 약 집단’)과 첫 번째 잠재 층(‘약물 비사용 집단’)의 오즈비이다. 여성일 경우 남성에 비해 ‘약물 비사용 집단’보다 ‘다이어트 약 집단’에 속할 확률이 높다. 흑인이 경우 백인에 비해 ‘다이어트 약 집단’보다 ‘약물 비사용 집단’에 속할 확률이 높다. 종교가 중요한 경우 중요하지 않은 경우에 비해 ‘다이어트 약 집단’보다 ‘약물 비사용 집단’에 속할 확률이 높다. 다음으로는 세 번째 잠재 층(‘마리화나 집단’)과 첫 번째 잠재 층(‘약물 비사용 집단’)의 오즈비이다. 여성의 경우 남성에 비해 ‘마리화나 집단’보다 ‘약물 비사용 집단’에 속할 확률이 높다. 흑인의 경우 백인에 비해 ‘마리화나 집단’보다 ‘약물 비사용 집단’에 속할 확률이 높다. 정치적 성향이 보수적인 경우 그렇지 않은 경우에 비해 ‘마리화나 집단’보다 ‘약물 비사용 집단’에 속할 확률이 높다. 종교가 중요한 경우 그렇지 않은 경우에 비해 ‘마리화나 집단’보다 ‘약물 비사용 집단’에 속할 확률이 높다.

[1] * 표시는 유의수준 0.05에서 유의한 변수를 나타냄

2/1	β 추정치	오즈비	p-value[1]
Intercept	-0.826	0.438	0.037 *
성별(여성)	0.583	1.792	0.072
인종(흑인)	-1.14	0.320	0.011 *
정치적 성향(보수)	0.035	1.036	0.912
정치적 성향(중립)	-0.373	0.689	0.226
정치적 성향(진보)	-0.213	0.808	0.501
종교의 중요도(중요함)	-0.787	0.455	0.002 *
3/1	β 추정치	오즈비	p-value
Intercept	-0.387	0.679	0.116
성별(여성)	-1.034	0.356	< 0.001 *
인종(흑인)	-0.878	0.415	0.012 *
정치적 성향(보수)	-1.055	0.348	0.009 *
정치적 성향(중립)	-0.416	0.660	0.100
정치적 성향(진보)	0.212	1.236	0.381
종교의 중요도(중요함)	-0.407	0.666	0.049 *

Table 13. 잠재 층이 4개인 약물 사용 잠재 층 회귀 모형의 잠재 층 별 오즈 비 (base line 잠재 층 : 첫 번째 잠재 층)

이러한 모형 해석을 통해 다음과 같은 시사점을 발견할 수 있다. 여성의 경우 남성보다 담배, 술, 마리화나의 약물 사용은 덜하지만 다이어트 약 사용량이 많은 경향이 있다. 흑인이거나 종교를 중요하게 생각할 경우 전반적 약물을 덜 사용하는 경향이 있다.

2-4 잠재 층 효과를 제어한 공변량 사이의 연관성

2-3 잠재 층 회귀 모형에서 우리는 4개의 잠재 층을 가진 삶의 만족도 잠재 층 회귀 모형과 3개의 잠재 층을 가진 약물 사용 회귀 모형을 적합했다. 각 잠재 층 회귀 모형은 사후확률 θ_{il}를 사용해 1,443명의 관측치에 대해 적절한 잠재 층을 예측한다. 두 잠재 층 회귀 모형으로 예측된 관측치의 잠재 층 비율은 Table 14.와 Table 15.에서 확인할 수 있다. 여기서 예측 비율은 2-3절에서 언급한 γ^*의 추정치와는 다른 개념으로. 모수에 대한 추정치가 아닌 실제 데이터에 대해 예측된 잠채 층 비율을 의미한다.

잠재 층	1	2	3	4
예측 비율	0.243	0.140	0.334	0.283

Table 14. 삶의 만족도 잠재 층 회귀 모형의 잠재 층 별 예측 비율

잠재 층	1	2	3
예측 비율	0.697	0.168	0.135

Table 15. 약물 사용 잠재 층 회귀 모형의 잠재 층 별 예측 비율

각 관측치에 대한 잠재 층이 주어졌을 때 우리는 코크란-맨텔-핸젤 검정을 사용해 잠재 층의 효과를 제어한 공변량의 연관성을 살펴볼 수 있다.

변수 조합	자유도	Q_{CMH}	p-value
(성별, 인종)	1	1.315	0.252
(성별, 정치적 성향)	3	9.036	0.0288
(성별, 종교의 중요도)	1	25.857	< 0.001
(인종, 정치적 성향)	3	54.255	< 0.001
(인종, 종교의 중요도)	1	96.213	< 0.001
(정치적 성향, 종교의 중요도)	3	28.537	< 0.001

Table 16. 삶의 만족도 잠재 층 효과를 제어 후 성별, 인종, 정치적 성향, 종교의 중요도 4개 변수에 대한 코크란-맨텔-핸젤 검정

변수 조합	자유도	Q_{CMH}	p-value
(성별, 인종)	1	0.092	0.762
(성별, 정치적 성향)	3	19.269	< 0.001
(성별, 종교의 중요도)	1	20.426	< 0.001
(인종, 정치적 성향)	3	3.832	0.280
(인종, 종교의 중요도)	1	28.603	< 0.001
(정치적 성향, 종교의 중요도)	3	55.675	< 0.001

Table 17. 약물 사용 잠재 층 효과를 제어 후 성별, 인종, 정치적 성향, 종교의 중요도 4개 변수에 대한 코크란-맨텔-핸젤 검정

Table 16.과 Table 17.를 통해 코크란-맨텔-핸젤 검정 결과를 살펴볼 수 있다. 약물 사용 잠재 층 효과를 제어한 코크란-맨텔-핸젤 검정은 앞서 2-2-1절에서 시행한 카이제곱 독립성 검정과 같은 결과를 보여주었다. 하지만 삶의 만족도 잠재 층 효과를 제어한 코크란-맨텔-핸젤 검정에서는 2-2-1절의 카이제곱 독립성 검정과 달리 인종과 정치적 성향 사이에 유의한 연관성이 있다는 결과가 나왔다. 이는 2-3-1절의 삶의 만족도 잠재 층 회귀 모형을 해석하면서 인종과 정치적 성향 사이의 연관성이 의심되었던 사실과 부합하는 결과이다.

2-5. 잠재 층 사이의 연관성

카이제곱 독립성 검정을 통해 삶의 만족도 잠재 층과 약물 사용 잠재 층의 연관성도 살펴볼 수 있다. 이를 통해 청소년의 삶의 만족도와 약물 사용 사이의 연관성을 파악할 수 있고 이는 청소년 지도 방향 설정에도 도움이 될 것이다. 카이제곱 독립성 검정을 위해 Table 18.과 같은 분할표를 생성했다.

		약물 사용 잠재 층		
		1	2	3
삶의 만족도 잠재 층	1	239	77	35
	2	127	37	38
	3	353	75	54
	4	287	53	68

Table 18. 두 잠재 층의 예측값에 대한 분할표

두 잠재 층에 대한 카이제곱 독립성 검정 결과, 카이제곱 검정 통계량 X^2 = 24.504 (자유도 = 6)으로 나와 p-value가 0.001보다 작았다. 따라서 삶의 만족도 잠재 층과 약물 사용 잠재 층 사이에는 유의한 연관성이 있음을 알 수 있다.

두 잠재 층 사이에 구체적으로 어떠한 연관성이 있는지 살펴보기 위해 카이제곱 독립성 검정 후 각 cell에 대한 표준화 잔차를 살펴보았다. 표준화 잔차는 Table 19.에서 확인할 수 있다.

		약물 사용 잠재 층		
		1	2	3
삶의 만족도 잠재 층	1	-0.762	2.978	-2.231
	2	-2.283	0.634	2.375
	3	2.061	-0.872	-1.818
	4	0.326	-2.413	2.200

Table 19. 카이제곱 독립성 검정 후 계산된 표준화 잔차

카이제곱 독립성 검정 후 계산된 표준화 잔차는 각 cell에 대해 귀무가설(두 변수가 서로 독립이다)이 얼마나 부적합한지를 나타낸다. 통상적으로 표준화 잔차의 절댓값이 2 또는 3 이상일 때 해당 cell에 귀무가설이 적합하지 않다고 판단한다. 또한 표준화 잔차가 양수라면 두 변수가 독립이라는 가정보다 더 적은 관측치가, 반대로 음수라면 독립 가정보다 더 많은 관측치가 관찰되었다는 의미이다. Table 19.에 나타난 표준화 잔차를 통해 두 잠재 층 사이의 연관성을 해석해보도록 하겠다. 의미있는 결과로 보이는 것은 '정부 불만족 집단'은 독립 가정에 비해 '약물 비사용 집단'에서 적은 관측치가 관찰되었고 '마리화나 집단'에서 많은 관측치가 관찰되었다. 반대로 '만족 집단'은 독립 가정에 비해 '약물 비사용 집단'에서 많은 관측치가 관찰되었고 '마리화나 집단'에서 적은 관측치가 관찰되었다. 이를 통해 삶의 만족도가 낮은 집단이 마리화나와 같은 강도 높은 약물을 사용하는 경향이 있다는 결론을 내릴 수 있다.

3. 결론

본 연구는 적절한 청소년 지도 방향을 설정하기 위해 2004년에 진행된 MTF 연구의 데이터를 사용해 청소년의 삶의 만족도와 약물 사용 행태에 대해 분석했다. 또한 청소년의 성별, 인종, 정치적 성향, 종교의 중요도를 함께 살펴봄으로써 청소년의 특성을 파악하고 삶의 만족도 및 약물 사용 행태와 어떠한 연관성이 있는지 분석했다. 이러한 분석을 위해 카이제곱 독립성 검정, 로그-선형 모형, 잠재 층 회귀 모형, 코크란-맨텔-핸젤 검정을 사용했다.

데이터 분석 결과는 다음과 같다. 청소년의 성별, 인종, 정치적 성향, 종교의 중요도 사이의 연관성을 알아보기 위한 카이제곱 독립성 검정에서 성별과 인종, 그리고 인종과 정치적 성향을 제외한 나머지 변수 조합에서는 유의한 연관성이 나타났다. 더불어 같은 변수를 사용해 로그-선형 모형을 적합하고 모형 적합도 검정을 실시한 결과 (성별*정치적 성향*종교의 중요도, 인종*종교의 중요도) 모형이 데이터에 가장 적합한 것으로 나타났다. 해당 로그-선형 모형에는 성별과 인종, 인종과 정치적 성향의 교호작용이 존재하지 않아 카이제곱 검정 결과와 부합하는 모형이었다. 해당 모형을 통해서 청소년의 성별, 인종, 정치적 성향, 종교의 중요도 사이의 연관성을 구체적으로 살펴볼 수 있었다. 남성은 정치적 성향이 보수적일 경향이 있었고 백인은 종교를 중요하게 생각하지 않는 경향이 있었다. 또한 다른 정치적 성향에 비해 보수적 정치 성향을 가진 사람이 종교를 중요하게 생각하는 경향이 있었다.

잠재 층 회귀 모형을 통해서 청소년의 삶의 만족도와 약물 사용 행태를 파악할 수 있었다. 삶의 만족도 회귀 모형은 4개의 잠재 층을 가진 모형이 가장 적절했다. 4개의 잠재 층을 가진 삶의 만족도 회귀 모형을 살펴본 결과 청소년은 삶의 만족도 측면에서 '친구를 좋아하는 평균 집단', '정부 불만족 집단', '만족 집단', 그리고 '대체로 만족, 정부 불만족 집단'의 4가지 집단(잠재 층)으로 나눌 수 있었다. 잠재 층의 분포는 '만족 집단'과 '대체로 만족, 정부 불만족 집단'이 약 30%로 가장 많게 나타났으며 '정부 불만족 집단'이 약 15%로 가장 적게 나타났다. 공변량으로 사용된 성별, 인종, 정치적 성향, 종교의 중요도와 잠재 층 사이의 연관성을 살펴본 결과 여성은 남성에 비해 삶의 만족도가 평균에 가까운 경향이 있었고 종교를 중요하게 여기는 것이 삶의 만족도를 높이는 경향이 있었다. 흑인의 경우 정부에 불만족하는 경향이 있다. 보수 정당이 집권했던 당시 사회를 반영하듯 정치 성향이 보수적일수록 정부 만족도가 높고 반대로 정치 성향이 진보적일수록 정부 만족도가 낮았다. 삶의 만족도 잠재 층 회귀 모형에서는 앞선 카이제곱 검정 및 로그-선형 모형 결과와 달리 인종과 정치적 성향 사이의 연관성이 의심되었다.

약물 사용 잠재 층 회귀 모형의 경우 3개의 잠재 층을 가진 모형이 가장 적절했다. 3개의 잠재 층을 가진 약물 사용 회귀 모형을 살펴본 결과 청소년은 약물 사용 측면에서 '약물 비사용 집단', '다이어트 약 집단', '마리화나 집단'의 3가지 집단(잠재 층)으로 나눌 수 있었다. 잠재 층의 분포는 '약물 비사용 집단'이 약 60%로 가장 많았으며 '다이어트 약 집단'과 '마리화나 집단'의 비율은 크게 차이나지 않았다. 공변량과 잠재 층 사이의 연관성을 살펴본 결과 여성의 경우 남성보다 담배, 술, 마리화나의 약물 사용은 덜하지만 다이어트 약 사용량이 많은 경향이 있었다. 흑인이거나 종교를 중요하게 생각할 경우 전반적 약물을 덜 사용하는 경향이 있었다.

코크란-맨텔-핸젤 검정을 통해 잠재 층 효과를 제어한 청소년의 성별, 인종, 정치적 성향,

종교의 중요도의 연관성을 파악할 수 있었다. 약물 사용 잠재 층의 효과를 제어한 코크란-맨텔-핸젤 검정 결과는 앞서 진행한 카이제곱 독립성 결과와 같았지만 삶의 만족도 잠재 층의 효과를 제어한 코크란-맨텔-핸젤 검정에서는 인종과 정치적 성향이 유의한 연관성을 가지는 것으로 나타났다.

카이제곱 독립성 검정 및 표준화 잔차를 통해 삶의 만족도 잠재 층과 약물 사용 잠재 층 사이의 연관성을 살펴볼 수 있었다. 그 결과 삶의 만족도 잠재 층과 약물 사용 잠재 층 사이에는 유의한 연관성이 나타났으며, 삶의 만족도와 약물 사용은 반비례하는 경향을 확인할 수 있었다.

이러한 분석 결과를 반영해 본 연구에서는 다음과 같은 청소년 지도 방향을 제시하고자 한다. 삶의 만족도 측면에서는 전반적으로 삶의 만족도가 떨어지는 '정부 불만족' 집단에 대한 지도가 필요하다. 먼저 정부에 대한 불만족은 당시 정권에 의한 일시적 경향일 수 있으므로 추후 추가적 설문을 통해 자세히 확인해야 할 것이다. 또한 '정부 불만족 집단'은 네 집단 중 친구 만족도가 가장 낮게 나타났으므로 가정과 학교에서 그들의 교우관계에 관심을 가지고 지도해야 할 것이다. 다음으로 약물 사용 측면에서는 약 40%를 차지하는 '다이어트 약 집단'과 '마리화나 집단'에 대한 지도가 필요하다. 다이어트는 비만인 경우에만 올바른 식단 조절과 운동으로 할 수 있도록 가정과 학교에서의 교육이 필요하다. 특히 여자 청소년들에게 이러한 지도가 필요할 것이다. 또한 마리화나 흡연의 위험성과 발생할 수 있는 부작용에 대한 교육을 통해 '마리화나 집단'의 마리화나 사용을 줄일 수 있도록 해야 할 것이다. 특히 백인 청소년을 대상으로 이러한 교육 및 지도가 필요하다. '다이어트 약 집단'과 '마리화나 집단'은 '약물 비사용 집단'에 비해 담배·술 사용량도 높으므로 전반적 약물 사용을 줄일 수 있도록 가정과 학교에서의 관심이 필요할 것이다. 종교를 중요하게 여길수록 약물 사용이 줄어드는 경향을 보였으므로 종교 시설과의 협업을 통해 청소년 약물 중독 예방 캠페인을 시행하는 것도 좋은 방법이다. 삶의 만족도와 약물 사용은 반비례하는 경향이 나타났으므로 약물 사용이 많은 청소년을 대상으로 삶의 만족도를 높일 수 있도록 지도하는 것을 추천한다. 이러한 지도 방향을 통해 청소년들이 건강한 사회 구성원으로 자랄 수 있을 것으로 기대된다.

---

4. 참고문헌

1. 이재원, 박미라, 유한나. 생명과학연구를 위한 통계적 방법(2005), 자유아카데미
2. Agresti, A. (1990). Categorical data analysis. New York [u.a.]: Wiley.
3. Linda M. Collins, Stephanie T. Lanza, Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences(2009), Wiley.
4. Drew A. Linzer, Jeffrey B. Lewis (2011). poLCA: An R Package for Polytomous Variable Latent Class Analysis. Journal of Statistical Software, 42(10), 1-29. URL http://www.jstatsoft.org/v42/i10/.

5. 분석 코드 (R)

```
library(foreign)
library(poLCA)

#### 1. 데이터 전처리 + 리코딩 ####
data_sat <- read.dta("04264-0002-Data.dta")
var <- paste0('V', c(1646:1648, 1650, 1653,
                     1102, 1216, 1254, 1712,
                     1150, 1151, 1167, 1170))
data_sat_1 <- data_sat[,var]
colnames(data_sat_1) <- c(paste0('q', 1:5), paste0('d', 1:4),
                          'gender', 'race', 'political', 'religion')
summary(data_sat_1)
summary(data_sat_1$political)

# 결측치 제거 : covariates의 결측치는 제거해야 함
na_obs <- which(apply(data_sat_1[,10:13], 1, function(x) any(x == 'MISSING')))
length(na_obs) # 1120개의 결측치
data_sat_2 <- data_sat_1[-na_obs,]

str(data_sat_2) # 최종적으로 1443개의 관측치

# 5개의 설문 문항 전처리 : 1~7의 답변, 1 <- 불만족 / 만족 -> 7
# 1-2 : 불만족, 3-5 : 중간, 6-7 : 만족으로 리코딩
# missing value : -9 -> NA로 변경
for(i in 1:5){
  q <- data_sat_2[,i]
  not.sat <- which(q %in% 1:2); data_sat_2[not.sat, i] <- 1
  mid <- which(q %in% 3:5); data_sat_2[mid, i] <- 2
  sat <- which(q %in% 6:7); data_sat_2[sat, i] <- 3
  missing <- which(q == -9); data_sat_2[missing, i] <- NA
}

for(i in 6:9){
  data_sat_2[,i] <- as.numeric(data_sat_2[,i])
  d <- data_sat_2[,i]
  none <- which(d == 2); data_sat_2[none, i] <- 1
  often <- which(d %in% 3:5); data_sat_2[often, i] <- 2
  lot <- which(d %in% 6:8); data_sat_2[lot, i] <- 3
  missing <- which(d == 1); data_sat_2[missing, i] <- NA
}

summary(data_sat_2)

# 4개의 covariates 리코딩
# gender
data_sat_2$gender <- factor(as.numeric(data_sat_2$gender),
                            levels = 2:3, labels = c('male', 'female'))

# race
```

```
data_sat_2$race <- factor(as.numeric(data_sat_2$race),
                          levels = 2:3, labels = c('white', 'black'))

# political belief
cons <- which(as.numeric(data_sat_2$political) %in% 2:3)
mod <- which(as.numeric(data_sat_2$political) == 4)
lib <- which(as.numeric(data_sat_2$political) %in% 5:7)
none <- which(as.numeric(data_sat_2$political) == 8)

poli <- rep(0, nrow(data_sat_2))
poli[cons] <- 'CONS'
poli[mod] <- 'MOD'
poli[lib] <- 'LIB'
poli[none] <- 'NONE'
data_sat_2$political <- factor(poli, levels = c('NONE', 'CONS', 'MOD', 'LIB'))

# religion importance
not_imp <- which(as.numeric(data_sat_2$religion) %in% 2:3)
imp <- which(as.numeric(data_sat_2$religion) %in% 4:5)

rel <- rep(0, nrow(data_sat_2))
rel[not_imp] <- 'not_imp'
rel[imp] <- 'imp'
data_sat_2$religion <- factor(rel, levels = c('not_imp', 'imp'))

summary(data_sat_2)
data_sat <- data_sat_2

# response Table
lapply(lapply(apply(data_sat, 2, table, useNA = 'ifany'), '/', nrow(data_sat)), round, 3)


#### 2. 카이제곱 독립성 검정 ####
chi.sq.test <- list()
comb <- combn(4, 2)
for(i in 1:ncol(comb)){
  comb.var <- comb[,i]+9
  data <- data_sat[,comb.var]

  chi.sq.test[[i]] <- chisq.test(table(data), correct = FALSE)
  names(chi.sq.test)[i] <- paste(colnames(data)[1], '&', colnames(data)[2])
}

chi.sq.test

#### 3. 로그-선형 모형 ####

loglin.data <- matrix(0, 32, 5); loglin.data <- as.data.frame(loglin.data)
colnames(loglin.data) <- c(colnames(data_sat)[10:13], 'count')
```

```
iter <- 1
for(i in 1:2){for(j in 1:2){for(k in 1:4){for(l in 1:2){
  x1 <- as.numeric(data_sat$gender) == i; loglin.data[iter, 1] <- levels(data_sat$gender)[i]
  x2 <- as.numeric(data_sat$race) == j; loglin.data[iter, 2] <- levels(data_sat$race)[j]
  x3 <- as.numeric(data_sat$political) == k; loglin.data[iter, 3] <- levels(data_sat$political)[k]
  x4 <- as.numeric(data_sat$religion) == l; loglin.data[iter, 4] <- levels(data_sat$religion)[l]

  loglin.data[iter, 5] <- sum(x1 & x2 & x3 & x4)
  iter <- iter+1
}}}}

sum(loglin.data$count) # observation 수 확인

fit_full_0 <- glm(count ~ (gender + race + political + religion)^4, data = loglin.data,
                  family = poisson())

fit_full <- glm(count ~ gender + race + political + religion +
                  gender*political + gender*religion + race*political +
                  race*religion + political*religion +
                  race*political*religion + gender*political*religion, data = loglin.data,
                family = poisson())

a0 <- anova(fit_full, fit_full_0)
1-pchisq(a0$Deviance[2], a0$Df[2]) # p-value = 0.87, reduced model 선택

summary(fit_full) # 3차 교호작용 항이 유의하지 않아보이므로 reduced model 적합 후 비교

# gender:political:religion만 제거한 reduced model 적합
fit_1 <- glm(count ~ gender + race + political + religion +
                  gender*political + gender*religion + race*political +
                  race*religion + political*religion +
                  race*political*religion, data = loglin.data, family = poisson())

a1 <- anova(fit_1, fit_full)
1-pchisq(a1$Deviance[2], a1$Df[2]) # p-value < 0.05


# race:political:religion만 제거한 reduced model 적합
fit_2 <- glm(count ~ gender + race + political + religion +
               gender*political + gender*religion + race*political +
               race*religion + political*religion +
               gender*political*religion, data = loglin.data, family = poisson())


a2 <- anova(fit_2, fit_full)
1-pchisq(a2$Deviance[2], a2$Df[2]) # p-value = 0.08, reduced model 선택

summary(fit_2)

# 3차 교호작용 항 모두 제거
fit_3 <- glm(count ~ gender + race + political + religion +
```

```
                gender*political + gender*religion + race*political +
                race*religion + political*religion, data = loglin.data,
              family = poisson())

a3 <- anova(fit_3, fit_2)
1-pchisq(a3$Deviance[2], a3$Df[2]) # p-value < 0.05

# race:political 교호작용 항 제거
fit_4 <- glm(count ~ gender + race + political + religion +
                gender*political + gender*religion + race*religion +
                political*religion + gender*political*religion, data = loglin.data,
              family = poisson())

a4 <- anova(fit_4, fit_2)
1-pchisq(a4$Deviance[2], a4$Df[2]) # p-value = 0.12, reduced model 선택

summary(fit_4)

# 3차 교호작용 다시 제거
fit_5 <- glm(count ~ gender + race + political + religion +
                gender*political + gender*religion + race*religion +
                political*religion, data = loglin.data,
              family = poisson())

a5 <- anova(fit_5, fit_4)
1-pchisq(a5$Deviance[2], a5$Df[2]) # p-value < 0.05

# fit_4를 최종 모델로 선택
summary(fit_4)

#### 4. LCA with covariates ####
# satisfaction model fitting
f <- cbind(q1, q2, q3, q4, q5)~1
idf_check <- list('nclass = 2' = NULL,
                  'nclass = 3' = NULL, 'nclass = 4' = NULL,
                  'nclass = 5' = NULL, 'nclass = 6' = NULL)
model_selection <- matrix(0, 5, 8)
colnames(model_selection) <- c('nclass','npar','G^2', 'df', 'p-value',
                               'AIC','BIC','likelihood')

for(i in 1:5){
  set.seed(2021021352)
  model_selection[i,1] <- i+1
  model <- poLCA(f, data = data_sat[,1:5], nclass = i+1, na.rm = F, verbose = F, nrep = 50)
  model_selection[i,2] <- model$npar
  model_selection[i,3] <- model$Gsq
  model_selection[i,4] <- model$resid.df
  model_selection[i,6] <- model$aic
  model_selection[i,7] <- model$bic
  model_selection[i,8] <- model$llik
```

```
  boot_gsquare <- rep(0, 100)
  for(j in 1:100){
    set.seed(j)
    iclass <- sample(1:(i+1), nrow(data_sat), replace = T, prob = model$P)
    response <- matrix(0, nrow(data_sat), 5)

    for (k in 1:nrow(data_sat)){
      for (l in 1:5){
        response[k, l] <- sample(1:3, 1, prob = model$probs[[l]][iclass[k],])
      }
    }
    boot_data <- as.data.frame(response)
    colnames(boot_data) <- paste0('q', 1:5)
    boot_model <- poLCA(f, data = boot_data, nclass = i+1,
                        na.rm = F, verbose = F)
    boot_gsquare[j] <- boot_model$Gsq
  }
  model_selection[i,5] <- mean(boot_gsquare >= model$Gsq)

  # identification check
  idf_check[[i]] <- table(round(model$attempts, 1))
}

model_selection
idf_check
sapply(lapply(idf_check, max), '/', 50)

windows(10, 10)
par(mfrow = c(2, 3))
par(mar = c(3, 5, 2 ,2))
for(i in 1:5){
  barplot(idf_check[[i]], horiz = T, las = 1,
          main = paste('frequeancy of maximum log-likelihood (# of class = ', i+1, ')', sep = ''))
}

set.seed(2021021352)
f.1 <- cbind(q1, q2, q3, q4, q5) ~ gender+race+political+religion
model.sat <- poLCA(f.1, data = data_sat, nclass = 4,
                   na.rm = F, verbose = F, nrep = 200, maxiter = 10000)
model.sat

round(exp(model.sat$coeff),3)

# drug model fitting
f <- cbind(d1, d2, d3, d4)~1
idf_check <- list('nclass = 2' = NULL,
                  'nclass = 3' = NULL, 'nclass = 4' = NULL,
                  'nclass = 5' = NULL, 'nclass = 6' = NULL)
model_selection <- matrix(0, 5, 8)
colnames(model_selection) <- c('nclass','npar','G^2', 'df', 'p-value',
                               'AIC','BIC','likelihood')
```

```
for(i in 1:5){
  set.seed(2021021352)
  model_selection[i,1] <- i+1
  model <- poLCA(f, data = data_sat[,6:9], nclass = i+1, na.rm = F, verbose = F, nrep = 50)
  model_selection[i,2] <- model$npar
  model_selection[i,3] <- model$Gsq
  model_selection[i,4] <- model$resid.df
  model_selection[i,6] <- model$aic
  model_selection[i,7] <- model$bic
  model_selection[i,8] <- model$llik

  boot_gsquare <- rep(0, 100)
  for(j in 1:100){
    set.seed(j)
    iclass <- sample(1:(i+1), nrow(data_sat), replace = T, prob = model$P)
    response <- matrix(0, nrow(data_sat), 4)

    for (k in 1:nrow(data_sat)){
      for (l in 1:4){
        response[k, l] <- sample(1:3, 1, prob = model$probs[[l]][iclass[k],])
      }
    }
    boot_data <- as.data.frame(response)
    colnames(boot_data) <- paste0('d', 1:4)
    boot_model <- poLCA(f, data = boot_data, nclass = i+1,
                        na.rm = F, verbose = F)
    boot_gsquare[j] <- boot_model$Gsq
  }
  model_selection[i,5] <- mean(boot_gsquare >= model$Gsq)

  # identification check
  idf_check[[i]] <- table(round(model$attempts, 1))
}

model_selection
idf_check
sapply(lapply(idf_check, max), '/', 50)

windows(10, 10)
par(mfrow = c(2, 3))
par(mar = c(3, 5, 2 ,2))
for(i in 1:5){
  barplot(idf_check[[i]], horiz = T, las = 1,
          main = paste('frequeancy of maximum log-likelihood (# of class = ', i+1, ')', sep = ''))
}

set.seed(2021021352)
f.1 <- cbind(d1, d2, d3, d4) ~ gender+race+political+religion
model.drug <- poLCA(f.1, data = data_sat, nclass = 3,
                    na.rm = F, verbose = F, nrep = 200, maxiter = 10000)
```

```
probs.start.new <- poLCA.reorder(model.drug$probs.start, c(2, 1, 3))

set.seed(2021021352) # baseline 변경을 위해 모형 다시 적합
model.drug.1 <- poLCA(f.1, data = data_sat, nclass = 3,
                      na.rm = F, verbose = F, nrep = 200, maxiter = 10000,
                      probs.start=probs.start.new)
model.drug.1
round(exp(model.drug.1$coeff),3)

#### 5. CMH 검정 ####
data_sat$sat.class <- model.sat$predclass
data_sat$drug.class <- model.drug.1$predclass

CMH.test.sat <- list()
comb <- combn(4, 2)
for(i in 1:ncol(comb)){
  comb.var <- c(comb[,i]+9, 14)
  data <- data_sat[,comb.var]

  CMH.test.sat[[i]] <- mantelhaen.test(table(data), correct = FALSE)
  names(CMH.test.sat)[i] <- paste(colnames(data)[1], '&', colnames(data)[2])
}

CMH.test.sat


CMH.test.drug <- list()
comb <- combn(4, 2)
for(i in 1:ncol(comb)){
  comb.var <- c(comb[,i]+9, 15)
  data <- data_sat[,comb.var]

  CMH.test.drug[[i]] <- mantelhaen.test(table(data), correct = FALSE)
  names(CMH.test.drug)[i] <- paste(colnames(data)[1], '&', colnames(data)[2])
}

CMH.test.drug


#### 6. 삶의 만족도와 약물 사용의 관계 ####
class.ind <- chisq.test(table(data_sat[,c('sat.class', 'drug.class')]), correct = FALSE)
class.ind
round(class.ind$stdres, 3)
```