생존분석을 이용한 가출 요인 분석

6 팀 주제분석

정은진 이성희 최재영 한대룡 황원영

1. 서론

"밥값, 숙박비 쓰려고"...차 훔치고 가게 턴 세 가출청소년...최근, 가출 청소년은 절도 뿐만 아니라 성폭행, 마약 투여 등 다양한 범죄에 연루되는 기사를 많이 접할 수 있다. 이러한 가출 청소년들은 흔히 '가출 팸'을 결성하여, 가출 카페에서 활동하며 생계를 유지하기 위해 계획적 범죄를 행하기도 한다. 여성가족부 <청소년 유해환경 접촉 종합실태조사>에 따르면, 청소년 가출 경험률이 증가하는 추세임을 확인할 수 있다. 이뿐만 아니라, '서울시 2015년 기준 실태조사'에 따르면, 가출 청소년 5명 중 4명은 재 가출을 경험하는 등, 가출과 귀가의 굴레에서 벗어나지 못한다. 더욱더 많은 청소년이 가출을 감행하며, 귀가 후에도 재가출을 행하는 등 일상적인 생활로의 회귀가 어렵다.첫 가출 예방만이 중요한 것이 아닌, 가출에 대한 근본적인 원인의 탐색이 필요한 시점이다.

청소년 가출에 대한 국내의 선행연구들은 대부분 생태 체계적 관점(eco-system perspective)을 기반으로 개인, 가정, 학교 및 또래 등 다양한 체계에 속한 요인들이 가출에 미치는 영향력을 실증적으로 검증하고 있다(박명숙, 2006; 박영호 . 김태익, 2002; 배문조 . 전귀연, 2002; 정혜경 . 안옥희, 2001; 조학래, 2004; 현은민, 2000). 즉, 청소년 가출은 일시적인 감정에 의해 행해지는 것이 아닌, 개인, 가정, 학교 등 다양한 체계에 속한 요인들에 의해 복합적으로 발생한다. 어떤 요인들이 청소년으로 하여금 가출을 감행하도록 하는 것일까?

청소년들은 복합적인 요인들의 작용으로 가출에 대한 충동을 느끼고 있다. 이러한 가출의 예방을 위해서, 첫 가출이 발생하는 요인들을 분석하고자 했다. 다양한 체계 속 요인들이 가출에 미치는 영향력을 실증적으로 검증해 보고자 '생존 분석을 이용한 청소년들의 가출과 그 첫 가출 시기에 미치는 요인 분석'을 진행해보았다.

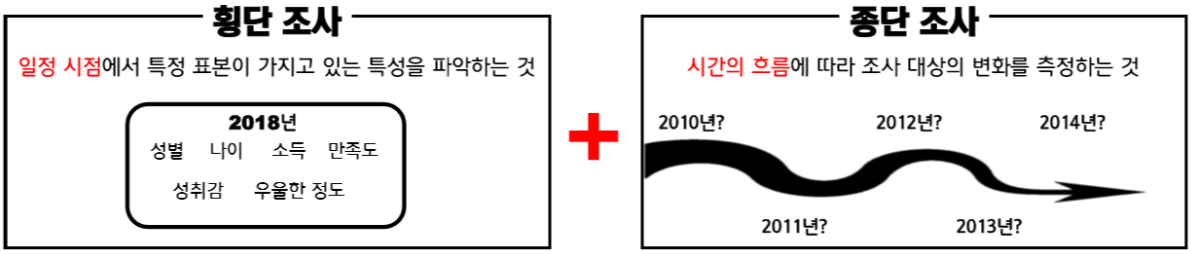
2. 연구대상과 분석 목표

2.1 데이터 소개

한국 아동 청소년 패널 조사(KCYPS)는 중학교 1학년 2,351명을 선정하여 2010년부터 2016년까지 총 7년 동안 매년 추적 조사를 진행하였다. 매년 새로운 학생들을 대상으로 한 것이 아니라 초기 대상자를 추적 조사하여 매년 진행한 것이며 조사원과의 개별 접촉을 통한 면접조사로 진행되었다. 조사 항목은 굉장히 많았으며 그 중 최근 1년 간 가출 경험이 있는가에 대한 질문이 존재하였으며 이를 반응변수로 설정하여 청소년의 가출 요인에 미치는 영향을 분석하고자 하였다.

이 설문조사의 경우 무작위의 학생들을 대상으로 진행한 것이 아니라 16개 시도의 중학교 1학년 학생 수에 비례하여 지역별로 표본 수를 할당하고, '확률비례 통계추출법'에 의거해 조사대상 학교를 추출하였기에 전체 청소년의 특성을 충분히 대표할 수 있을 것이라 판단하였다.

청소년 가출 데이터는 크게 3가지 특징이 존재하였는데 이는 다음과 같다. 첫 번째는 패널데이터라는 것이다. 패널 조사는 횡단조사와 종단조사를 모두 포함하는 개념으로 동일한 주제에 대해 동일한 응답자를 대상으로 반복적으로 조사하는 것을 의미한다. 이런 데이터 특성으로 인해 기존의 횡단 조사 데이터가 파악하지 못하는 동일 대상 내의 여러 변수의 변화와 종단 조사가 파악하지 못하는 동일 시점 내에 변수의 영향 등을 효율적으로 분석이 가능하다.



다음으론 설문조사의 항목이 리커트 척도 항목으로 구성되어 있다는 점이다. 리커트 척도란 어떤 질문에 대하여 예/아니오의 대답이 아니라 긍정/부정의 정도를 측정하는 척도를 의미한다. 이와 같은 데이터를 활용하는 방식은 다양한 방법이 존재하는데 각각의 항목에 대해서 답안을 하나의 범주로 즉 4개의 level을 가진 categorical 변수로 이용을 하거나, 혹은 1,2,3,4의 numeric 변수로 이용하든 여러 방안이 존재한다. 이에 대한 내용은 뒤의 전처리 과정에서 다룰 것이다.

문17) 학생의 학교생활에 대한 질문입니다. 아래 각 항목의 해당 칸에 ○표 해 주십시오.

나는 …	매우 그렇다	그런 편이다	그렇지 않은 편이다	전혀 그렇지 않다
① 학교 수업 시간이 재미있다.	1	2	3	4
② 학교 숙제를 빠뜨리지 않고 한다.	1	2	3	4
③ 수업 시간에 배운 내용을 잘 알고 있다.	1	2	3	4
④ 모르는 것이 있을 때 다른 사람(부모님이나 선생님 또는 친구들)에게 물어본다.	1	2	3	4
⑤ 공부 시간에 딴 짓을 한다.	1	2	3	4

마지막으론, Censored data(중도 절단 데이터)가 존재한다는 것이다. 2012년부터 2017년까지 총 7년의 데이터를 이용하였는데 즉, 이는 2017년에 설문조사 연구가 끝난 것을 의미한다. 이로 인해 2017년까지 가출을 하지 않은 학생들은 그 이후에 가출을 할지 안 할지에 대한 정보가 없는 상태이며 이를 censored date라고 한다. Censored data의 경우 크게 두 가지 종류가 존재하는데 type 1 censoring은 추적관찰이 종료됨으로써 관찰이 불완전해지는 경우를 의미하며 우리의 데이터에서 5년 동안의 조사가 끝난 경우이다. Type 2 censoring은 추적 관찰되는 기간 도중 도중탈락(follow-up loss)되는 경우(random censoring 이라고 한다)를 의미하는데 거주지의 변경이나 대상자의 거부, 사망과 같은 이유로 더 이상 추적조사가 불가능한 경우를 의미한다. 생존분석에서 위의 두 가지의 censoring data를 따로 구분하진 않지만 일반적으로 type 2 censoring의 데이터가 지나치게 많을 경우 분석의 왜

곡을 일으킬 수 있다. 허나, 우리의 경우 type 2 censoring data의 비중은 상당히 적어 이 부분은 따로 고려하지 않았다.

2.2 분석 목표

분석 목표는 크게 두 가지 존재한다. 첫 번째는 설문조사 데이터의 효율적인 활용이다. 설문조사 데이터는 주위에서 쉽게 발견할 수 있을 정도로 굉장히 많이 존재한다. 허나, 설문조사 데이터의 경우 여러 까다로운 점이 존재하였다. 그 중 하나는 리커트 척도이다. 예를 들어, 가족관계에도 굉장히 많은 항목이 존재한다. 이를 각각의 항목 개별적으로 이용한다면 다중공선성 문제가 굉장히 극대화된다. 또 다른 어려움은 NA 값이 굉장히 많이 존재한다는 점이다. 특히나 소득 관련 NA는 이를 삭제할 경우 왜곡된 데이터 편향을 초래할 수 있으므로 무조건적인 삭제는 상당히 위험하였다.

두 번째는 바로 생존분석을 이용한 요인 분석이라는 점이다. 어떻게 보면 청소년의 가출 유무를 통해 binary classification을 진행하는 것이 일반적일 것이다. 허나, 이 경우 가출 시간이라는 시간적인 요소를 전혀 고려하지 못하게 된다. 우리의 경우 시간이라는 추가적인 변수를 이용하고자 하였다. 따라서 뒤의 과정에서 로지스틱 회귀모형과 콕스 비례위험모형을 비교하면서 시간적인 요소를 고려하였을 때 어떤 차이가 있는지 역시 확인할 것이다.

3. 연구과정

3.1 데이터 전처리

3.1.1. 리커트 척도

앞서 설명했듯 해당 주제분석은 마이크로 데이터에서 제공하는 설문 데이터를 사용하였다. 설문 데이터 특성상 리커트 척도로 구성 된 항목들이 많았다. 리커트 척도를 간단히 설명하자면 어떤 질문에 대하여 "긍정/부정(만족/불만족)"의 정도를 측정하는 척도이다. 설문 데이터 내에 수 많은 질문들은 "매우 그렇다(1)/그런 편이다(2)/그렇지 않은 편이다(3)/전혀 그렇지 않다(4)" 중 하나를 택하는 방식으로 구성되었고 이 데이터 구조를 그대로 사용할 경우 몇 가지 문제가 발생하였다. 먼저 설문지 내의 모든 질문들을 각각 변수로 사용할 경우 변수의 수가 필요 이상으로 많아질 것으로 예상되었다. 또한 비슷한 주제 아래에 포함된 여러 질문들의 답변이 비슷한 방향을 띄기 때문에 변수 간의 다중 공선성 문제도 발생하였다. 따라서 리커트 척도로 구성된 답변들을 수치화 하기 위해서 같은 주제에 포함된 여러 질문들의 답을 평균처리하기로 하였다. 예를 들어 부모들의 방임과 관련된 질문들은 답변의 평균이 1에 가까울 수록 부모가 아이를 방임하는 것과 거리가 멀다는 뜻이었고 4에 가까울 수록 부모가 아이를 방임한다고 볼 수 있었다. 이를 통해 방임, 학대, 학습, 교우, 교사, 사회성 이라는 변수를 생성해냈다.

리커트 척도를 수치화 하는 과정에서 답변의 방향성을 통일하는 과정도 필요했다. 몇몇 질문은 1에 가까울수록 긍정적으로 볼 수 있었지만(부모님께서는 내가 잘못하시면 무조건 때리려고 하신다) 그렇지 않은 항목(학교 수업시간이 재미있다)도 있었기에 방향성을 통일하는 작업도 진행되었다.

3.1.2. 범주형 자료의 수치화

설문 데이터 내에는 또한 어떠한 경험의 유무를 묻는 질문도 많았는데 이러한 범주형 문항들을 수치화 하기 위

해서 비슷한 경험의 유무를 묻는 질문은 그 경험을 겪은 횟수로 변환했다. 예를 들어 학교폭력피해경험 유무를 묻는 질문이 많았는데 이를 수치화해서 몇 번의 학교폭력피해경험이 있었는 지로 변환하였다.

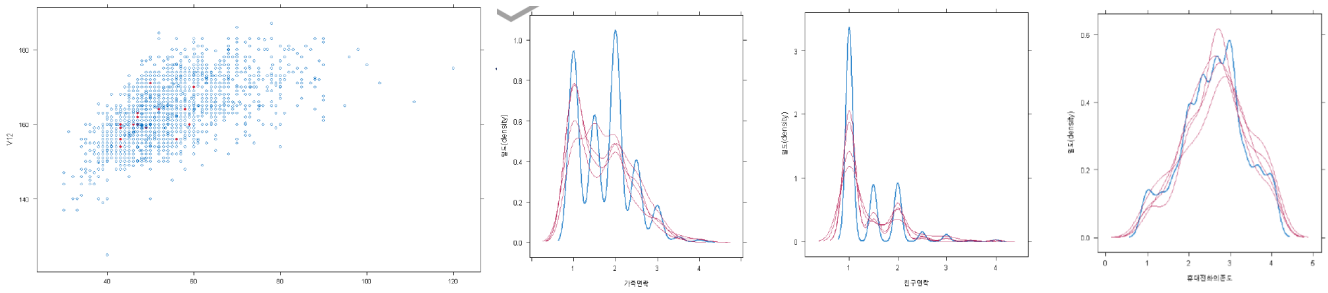
3.1.3. NA값 처리

통합된 데이터 셋은 총 30개의 변수와 2351개의 obs로 이루어져 있었는데 그 안에 수 많은 NA 값이 존재하였다. 이 NA 값들은 패널 데이터의 장점을 최대한 살려서 처리되었고 다음과 같은 방식으로 진행되었다.

먼저 추적조사 특성 상 대체가 불가능한 행은 삭제되었다. 예를 들어 특정 년도에 모든 정보가 기입되어 있지 않은 행이 존재했는데 이와 같은 행은 삭제처리 하였다.

그 다음으로 다른 해의 값으로 대체 가능한 NA 값들은 다른 해의 값을 최대한 활용하여 채워 넣었다. 먼저 연도에 따라 바뀌지 않을 것으로 고려되는 변수(부모님 최종학력, 가족구성, 다문화 가정 여부, 형제자매 유무)는 NA 값이 존재할 경우 다른 해에 기입된 값으로 대체하였다. 연도에 따라 바뀔 것이라고 고려되는 변수(가계소득수입)는 NA값이 존재할 경우 다른 해의 가계소득수입의 평균으로 채워 넣었다.

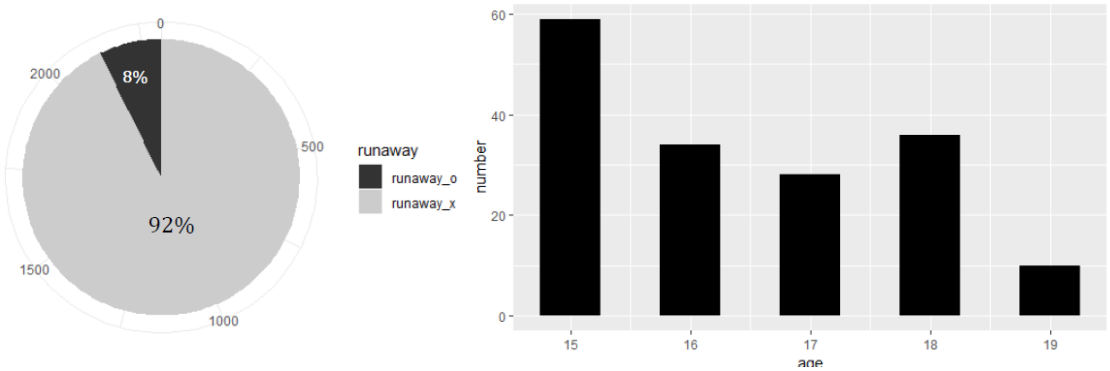
위의 과정을 모두 진행하고도 남은 결측치들은 MICE(Multiple Imputation by Changed Equation)을 활용하여 채워 넣었다. MICE란 다중 대체 방법으로 여러 변수에 걸쳐 존재하는 결측 값들을 관찰 값을 이용하여 예측하는 방식이다. 결측 값의 형태에 따라 채워 넣는 방식이 다양했는데 남은 NA값들이 연속형과 범주형 자료 둘 다 포함하고 있었기에 CART(Classification And Regression Tree) method를 사용하였다. CART method는 각 변수별로 다른 분포를 가정하고 변수별로 다른 대체 모델을 사용할 수 있는 장점을 지녔다.



MICE를 이용해 채워 넣은 결측 값의 분포 시각화
(파란색 : 기존 데이터, 빨간색 : 채워 넣은 결측치 데이터)

3.1.4. 데이터 통합

결측치 처리 과정을 통해 완성된 데이터 셋은 2305개의 obs와 30개의 변수로 이루어졌다. 학생들에게 부여된 고유한 ID 번호로 통합이 진행되었는데 추적조사가 이루어진 데이터였기에 각 ID별로 5년치의 정보가 들어 있었다. 즉, 고유한 ID 번호가 총 5개의 raw(5년치의 정보)를 갖고 있었는데 이후 진행된 생존분석에서는 최초 가출 시점의 정보가 필요하다고 판단되었기에 가출이 발생한 연도의 정보만 사용했다. 만약 해당 학생이 한 번도 가출을 하지 않았을 경우에는 가장 최근 연도의 정보를 사용했다.



최종 데이터 셋의 분포
(왼쪽 : 가출을 경험한 학생과 경험하지 않은 학생, 오른쪽 : 가출한 학생들의 최초 가출 시점 나이)

3.2 생존분석 소개

3.2.1. 생존분석

본 주제분석은 다양한 체계에 속한 요인들이 가출에 미치는 영향을 실증적으로 검증하는 것을 목표로 삼았고 생존분석이라는 분석 방법을 통하여 진행되었다. 생존분석이란 사건이 일어나거나 일어나지 않은 결과와 그러한 사건이 일어날 시점을 예측하고 설명하는 통계 방법이다. 본래 생존분석이란 생존과 사망이 뚜렷한 의료 데이터에서 많이 사용되었으나 최근에는 주식시장에서의 고객 이탈율, 소비자가 핸드폰을 바꾸는 데에 걸리는 시간 등 의료 분야 뿐만이 아닌 다양한 사회 분야에서 사용되는 추세이다.

3.2.2. 생존분석 데이터

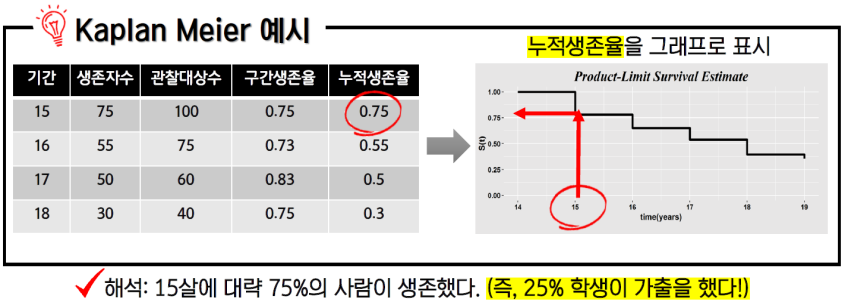
생존분석은 어떠한 결과에 영향을 미친 요인을 분석한다는 점에서 회귀분석과 비슷하지만 반응 변수에 사건 발생 뿐만이 아닌 생존 시간이라는 정보도 포함되어 있다는 점에서 일반적인 회귀분석과의 차별점을 갖는다. 즉, 같은 사건이 발생하더라도 발생 기간이 다르면 추가적인 요인 분석이 가능하고 이를 위해서는 censored data가 필요하다.

Censored data란 절단된 데이터 형태를 의미하는데 예시를 통해 쉽게 이해할 수 있다. 100명의 학생만을 추적 관찰하면서 최초 가출 발생을 탐지한다고 하자. 관찰 기간인 5년이 지난 후 단 10명만이 가출을 했다면 생존 분석을 하는 방식 중 하나는 그 10명만을 분석하는 것이다. 그러나 이는 샘플 사이즈를 매우 감소시키고 selection bias 등 여러 문제를 일으키는데 이를 해결하는 방안으로 90명을 censored 되었다고 간주하여 분석에 포함시킬 수 있다. Censored data는 2가지 종류가 있는데 먼저 추적 관찰이 종료됨으로써 관찰이 불완전해지는 경우가 있다. 본 주제분석에서 5년동안의 조사가 종료되었는데 그 이후의 가출 여부를 확인 불가능한 경우이다. 그리고 추적 관찰되는 기간 중에 도중 탈락되는 경우도 있는데 이는 추적조사가 특정 사유에 의해 끊긴 경우라고 볼 수 있다. 거주지의 변경이나 대상자의 거부 등이 이유가 될 수 있다.

이렇게 처리된 censored data는 시점에 대한 추가 정보를 담을 수 있다. 예를 들어 censored 처리를 하지 않는다면 추적 조사기간 마지막 연도인 2015년에 가출을 한 학생과 가출을 하지 않은 학생은 똑같이 처리가 된다. 왜냐하면 언급했듯 가출을 하지 않은 학생은 2015년의 정보를 사용하기로 했기 때문이다. 하지만 censored 처리를 한다면 두 ID간의 차이를 확인할 수 있다. 똑같이 2015년의 정보를 담고 있는 ID라도 +표기가 붙어있는 경우에는 2015년까지 가출이 발견되지 않았다는 뜻이고 +표기가 없다면 2015년에 가출을 한 학생이라는 뜻이다.

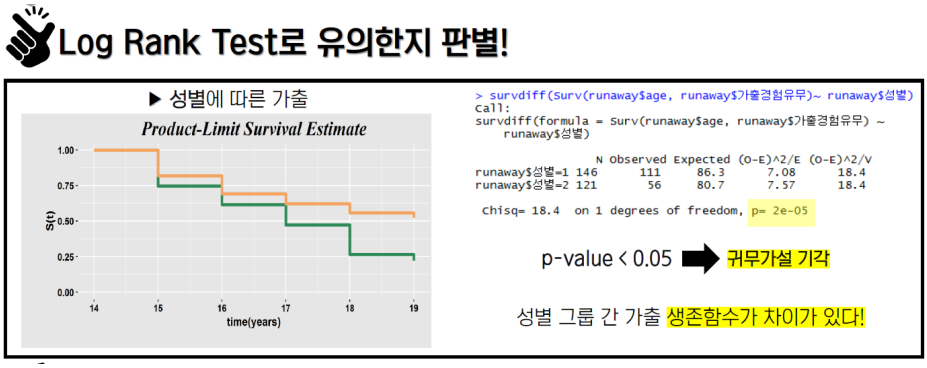
3.2.3. Kaplan Meier & Log Rank Test

생존분석에서 그룹 간의 시간에 따른 생존 분포를 확인하는 데에 사용되는 방법으로는 Kaplan Meier와 Log Rank Test가 있다. Kaplan Meier는 사건이 발생한 시점마다 구간 생존율을 구한 뒤 이의 누적 생존율을 확인하는 방법인데 그래프를 통해 직관적으로 이해할 수 있다.



기간	생존자수	관찰대상수	구간생존율	누적생존율
15	75	100	0.75	0.75
16	55	75	0.73	0.55
17	50	60	0.83	0.5
18	30	40	0.75	0.3

Kaplan Meier의 장점은 이렇듯 그룹 간의 생존 곡선을 시각적으로 확인 할 수 있다는 것인데 이를 수치적으로 판단하기 위해 고안된 방법이 Log Rank Test이다. Log Rank Test의 귀무가설은 집단간의 생존함수의 차이가 없다는 것인데 test를 통해 확인된 p-value가 유의한 경우 이를 기각하고 집단간의 생존함수에 차이가 있다고 판단 할 수 있다.



Kaplan Meier 와 Log Rank Test 둘 다 그룹 간의 차이를 알려주긴 하지만 어떤 변수가 얼마만큼의 영향력을 끼치는 지는 알 수 없다는 단점을 지니고 있다.

3.3 분석 과정 (콕스 비례 위험 모형)

3.3.1. 콕스 비례 위험 모형

생존 분석은 여러 변수의 영향력을 본다는 점에서 다중 회귀분석과 비슷하며, 생존여부를 종속 변수로 한다는 점에서 로지스틱 회귀분석과 비슷하다. 그러나, 회귀분석과 달리 종속 변수를 두 개, 생존여부와 생존시간으로 한다는 점과 절단된 데이터 (censored data)를 이용한다는 점에서 차이가 있다.

"콕스 비례 위험 모형(Cox Propotional Hazard Model)"은 생존과 관련한 다변수의 영향력을 알아보는 생존 분석 방법 중 하나이다. 앞서 소개한 Kaplan-meier, log-rank test가 생존율을 구함으로써 사건이 발생했는지 여부에 집중한 것과 달리, 생존에 영향을 미치는 여러 인자들의 영향력을 알아보는 모형이다. 즉, 여러 인자의 영향력을 고려하여 사건이 관측시점(t)에서 발생할 위험을 구하는 통계 모델이다.

$$h_i(t) = h_0(t)\exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)$$

콕스 비례 위험 모형 수식

다음은, 콕스 비례 위험 모형 수식이다. 일반적인 회귀식과 비슷한 모습인 위 수식은 크게 세가지 부분으로 나누어 보면, 쉽게 이해할 수 있다. H0(t)는 공변량[1](X)의 영향력이 없을 때 해당 개체가 t시점에서 가지고 있는 고유의 위험도이다(모든 공변량이 0일때 exp(0)가 1이 됨을 생각하면 쉽게 이해할 수 있다). 흔히 기저 위험(baseline hazard)라고 부른다. 두 번째 부분은 공변량과 각각의 공변량의 영향력을 나타내는 B이며, 공변량의 생존회귀계수라고 부른다. 마지막은 이 두 부분을 구함으로써 얻어지는 hi(t)로, 해당 i개체가 관측시점(t)에서 가지는 위험이라고 할 수 있다.

3.3.2. 분석 과정

A. Log linear 가정 검정

콕스 비례 위험 모형을 적용해 생존 분석을 하는 데에는 "log linear 가정 검정 – 변수선택 - 비례위험 가정 검정"의 과정이 필요하다. 첫째로, Log linear 가정을 만족시켜야 하는데, 이 모형은 공변량과 잔차가 선형 관계를 가지고 있어야 선형성을 모델에 반영할 수 있기 때문이다. 따라서, 우리의 X변수 32개 각각의 공변량과 마팅게일 잔차 사이의 그래프를 그려 선형성을 확인해 보았다.

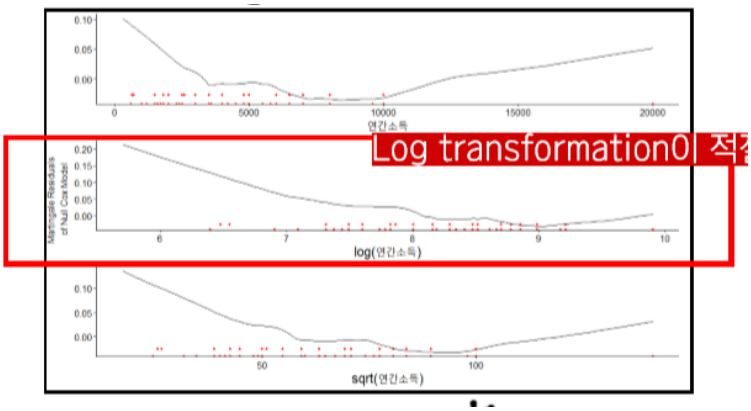


Figure 연간 소득

Figure? 는 공변량 중 하나인 '연간 소득'을 살펴본 것이다. 이 변수 같은 경우는 log 변환을 통하여 선형성을 확보할 수 있었다. 실제로 아래 figure?와 같이 log변환을 통하여 연간소득이 유의미한 변수가 됨을 확인할 수 있다. 이렇게 변환을 통하여 선형성을 확보할 수 있는 변수들은 변환 과정을 진행하였다.

[1] 종속변수에 대하여 독립 변인과 기타 잡음 변인들이 공유하는 변수를 의미

① log transformation 이전,

```
> summary(coxrg1)
call:
coxph(formula = Surv(age, 가출유무) ~ ., data = runaway3)

  n= 2314, number of events= 167

                 coef  exp(coef)   se(coef)      z Pr(>|z|)
성별2        -1.432e+00  2.388e-01  2.389e-01 -5.995 2.04e-09 ***
부최종학력2  -1.038e-01  9.014e-01  2.321e-01 -0.447  0.65467
부최종학력3   3.437e-01  1.410e+00  4.979e-01  0.690  0.49002
모최종학력2  -1.385e-01  8.706e-01  2.366e-01 -0.586  0.55815
모최종학력3  -1.553e+00  2.116e-01  1.061e+00 -1.463  0.14337
부종사상지위1  6.872e-01  1.988e+00  3.341e-01  2.057  0.03972 *
부종사상지위2  7.495e-01  2.116e+00  3.425e-01  2.188  0.02864 *
모종사상지위1  2.066e-01  1.230e+00  1.934e-01  1.069  0.28524
모종사상지위2  6.072e-01  1.835e+00  2.820e-01  2.153  0.03130 *
연간소득      -4.423e-05  1.000e+00  4.240e-05 -1.043  0.29694
보호자건강2  -3.041e-01  7.378e-01  2.333e-01 -1.303  0.19245
보호자건강3  -2.407e-01  7.861e-01  3.281e-01 -0.734  0.46320
보호자건강4  -9.634e-01  3.816e-01  8.446e-01 -1.141  0.25399
키           -2.470e-02  9.756e-01  1.564e-02 -1.579  0.11433
몸무게       -4.260e-02  9.583e-01  1.024e-02 -4.160 3.19e-05 ***
건강상태2    -2.485e-01  7.800e-01  1.938e-01 -1.282  0.19974
건강상태3     1.481e-02  1.015e+00  2.795e-01  0.053  0.95773
건강상태4    -3.661e-01  6.934e-01  8.622e-01 -0.425  0.67107
```

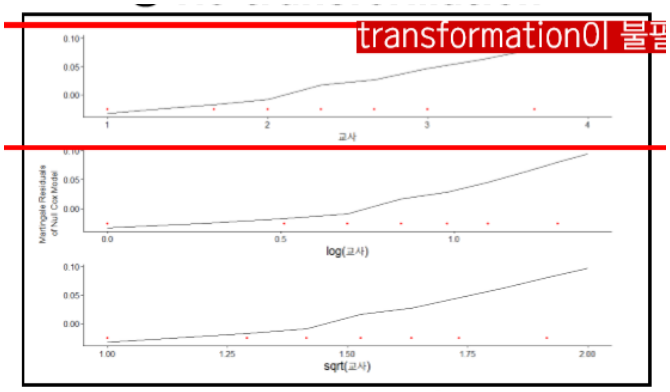
② log transformation 이후,

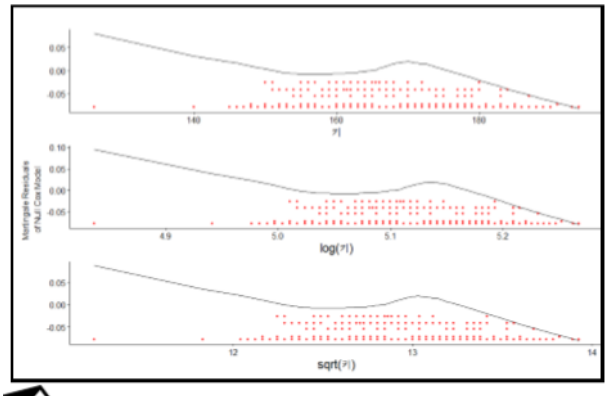
```
> summary(coxrg2)
call:
coxph(formula = Surv(age, 가출유무) ~ ., data = runaway4)

  n= 2314, number of events= 167

                 coef exp(coef)  se(coef)      z Pr(>|z|)
성별2        -0.847633  0.428428  0.198564 -4.269 1.97e-05 ***
부최종학력2   0.023305  1.023578  0.225781  0.103 0.917790
부최종학력3   0.164494  1.178797  0.492105  0.334 0.738179
모최종학력2  -0.082677  0.920649  0.229771 -0.360 0.718980
모최종학력3  -0.689545  0.501804  1.037994 -0.664 0.506495
부종사상지위1  1.075372  2.931082  0.355082  3.029 0.002458 **
부종사상지위2  1.215120  3.370698  0.358477  3.390 0.000700 ***
모종사상지위1  0.068373  1.070764  0.194031  0.352 0.724553
모종사상지위2  0.607915  1.836598  0.272876  2.228 0.025894 *
연간소득      -0.370041  0.690706  0.167529 -2.209 0.027187 *
보호자건강2  -0.354371  0.701615  0.223558 -1.585 0.112934
보호자건강3  -0.156523  0.855111  0.322323 -0.486 0.627243
보호자건강4  -0.268311  0.764670  0.790930 -0.339 0.734433
몸무게       -2.452863  0.086047  0.509242 -4.817 1.46e-06 ***
건강상태2    -0.381370  0.682925  0.188904 -2.019 0.043503 *
건강상태3    -0.082530  0.920784  0.264104 -0.312 0.754667
건강상태4     0.156737  1.169688  0.756914  0.207 0.835952
```

그러나, 아래 figure?와 같이 sqrt나 log변환을 통해서도 선형성을 확보할 수 없는 변수들도 존재하였고, 해당 변수들은 삭제하였다.





모든 변수들의 log linear 가정을 검정한 결과 다음 표?와 같이 정리됐다.

Log 변환	연간소득, 몸무게, 가족여행, damage, 방임, 가족연락, 가출친구, 비행친구, 체험활동
변수 유지	성별, 부 최종학력, 모 최종학력, 부 종사상지위, 모종사상지위, 보호자 건강, 건강상태, 성적만족도, 가족구성, 다문화 가정여부, 형제자매 유무, 전학경험 유무, 학대, 학습, 교사, 교내동아리, 교외동아리
변수 삭제	키, sleep, 교우, 사회성, 친구연락, 휴대전화 의존도

B. 변수선택

가정을 만족하는 26개의 변수들을 가지고 변수 선택(subset selection)을 진행하였다. 변수 선택 과정을 통하여 데이터에 대한 설명력이 뛰어난 동시에 단순한 형태를 가지는 모형을 찾을 수 있다. 여러 방법 중 hybrid selection 방법을 사용해 AIC가 가장 작은 모형을 택했다. 최종적으로 선택된 변수들은 다음 결과창과 같다.

```
> summary(ro.step)
Call:
coxph(formula = Surv(age, 가출유무) ~ 성별 + 부종사상지위 + 모종사상지위 +
    연간소득 + 몸무게 + 성적만족도 + 가족구성 + 가족여행 + damage +
    학대 + 학습 + 교사 + 가족연락 + 가출친구 + 교외동아리 + 체험활동,
    data = runaway4)
```

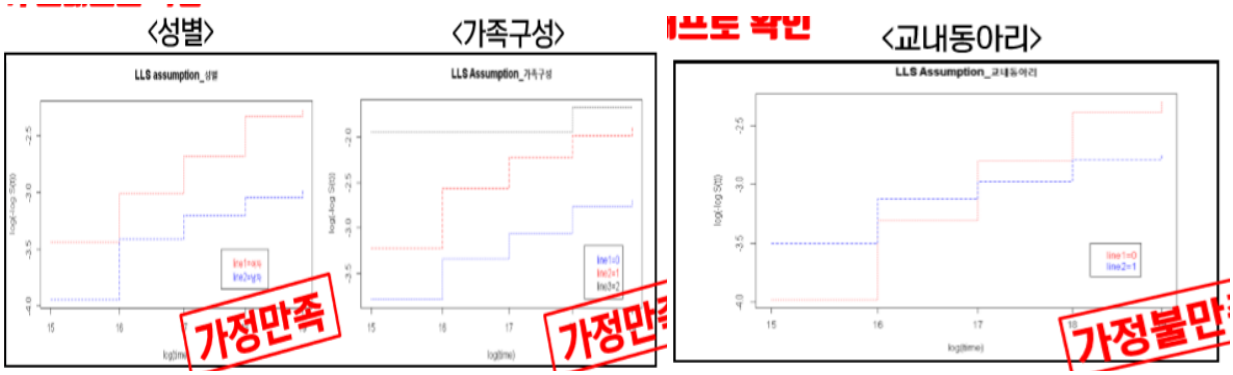
```
> extractAIC(coxrg1)[[2]]
[1] 2200.474
```

```
> extractAIC(ro.step)[[2]]
[1] 2145.494
```

Figure?의 결과를 통해 모형의 AIC가 2200.474에서 2145.494로 줄어드는 것을 확인할 수 있었다.

C. 비례 위험 가정 검정

최종으로 선택한 변수들로 비례 위험 가정을 만족시키는지를 검정했다. 비례 위험 가정은 위험비[2]가 시간에 의존하지 않고 항상 일정(비례)하다는 가정이다. 예를 들어, 16살에 담배를 피는 학생이 피지 않는 학생보다 가출할 위험비가 2배라면, 18살에 담배를 피는 학생이 피지 않는 학생보다 가출할 위험비 역시 2배여야 한다. 비례 위험 가정은 각 변수 별로, 그래프를 그렸을 때, 두 그래프가 시간에 관계없이 일정한 간격을 두고 있어야 만족시킨다고 할 수 있다.



Figure? 그래프의 성별 변수에서 남성을 나타내는 파란선과, 여성을 나타내는 빨간선이 겹치지 않고 시간에 따라 일정한 간격을 유지하는 것을 볼 수 있다. 반면, figure?의 교내동아리 참여 유무를 나타내는 빨간선과 파란선이 특정 시점에서 교차되는 것을 통해, 일정하지 않음을 확인할 수 있다. 위와 같은 방법으로 확인한 결과 교내동아리를 제외한 다른 범주형 변수들은 모두 가정을 만족함을 확인할 수 있었다.

그래프 외에도, 통계적 검정을 통하여 비례 위험 가정을 만족시키는지를 통계적으로 확인할 수 있었다. 공변량의 효과를 나타내는 B(t)가 시간에 따라 변하지 않아야 하기 때문에, B에 대하여 다음과 같은 수식에서 θ가 0이 된다면, 가정을 만족시킨다고 할 수 있다.

$$\text{시간 가변 회귀 계수} : \beta(t) = \beta + \theta g(t) \qquad H_0 : \theta = 0$$

[2] Hazard ratio. hi(t)/h0(t)

Figure? 는 검정 결과이다. 대부분의 변수들은 귀무가설을 기각 시키지 못하기 때문에, B(t)가 시점(t)과 상관없이 일정한 값을 가짐을 확인할 수 있었다. 그러나 변수 "교내 동아리"만 p-value가 0.05이하로 귀무가설을 기각 시킴으로써, 비례 위험 가정을 불만족함을 다시 한번 확인할 수 있었다.

```
> cox.zph(ro.step)
                  rho     chisq        p
성별2          -0.07778  1.21867 2.70e-01
부종사상지위1 -0.04054  0.34578 5.57e-01
부종사상지위2  0.00630  0.00772 9.30e-01
모종사상지위1 -0.10394  1.87937 1.70e-01
모종사상지위2  0.02023  0.06966 7.92e-01
연간소득       0.09650  1.76725 1.84e-01
몸무게         0.08766  1.49336 2.22e-01
성적만족도2   -0.02947  0.14447 7.04e-01
성적만족도3    0.00311  0.00163 9.68e-01
성적만족도4   -0.01858  0.06102 8.05e-01
가족구성1      0.04722  0.49014 4.84e-01
가족구성2     -0.02054  0.08366 7.72e-01
가족여행      -0.18362  6.48265 1.09e-02
damage        -0.00732  0.01134 9.15e-01
학대           0.14063  3.84125 5.00e-02
학습           0.14557  3.93640 4.73e-02
교사          -0.11449  2.23809 1.35e-01
가족연락      -0.07006  0.83126 3.62e-01
가출친구       0.06950  0.99128 3.19e-01
교내동아리2   -0.30295 18.52837 1.67e-05
교외동아리2   -0.04890  0.41739 5.18e-01
체험활동      -0.10399  2.24206 1.34e-01
```

```
> summary(ro.step)
call:
coxph(formula = Surv(age, 가출유무) ~ 성별 + 부종사상지위 + 모종사상지위 +
    연간소득 + 몸무게 + 성적만족도 + 가족구성 + 가족여행 + damage +
    학대 + 학습 + 교사 + 가족연락 + 가출친구 + 교외동아리 + 비행친구수 +
    체험활동, data = runaway4)

  n= 2314, number of events= 167

                  coef exp(coef) se(coef)      z Pr(>|z|)
성별2          -0.92235   0.39758  0.19801 -4.658 3.19e-06 ***
부종사상지위1  0.98467   2.67692  0.33899  2.905 0.003676 **
부종사상지위2  1.08663   2.96428  0.34183  3.179 0.001479 **
모종사상지위1  0.09170   1.09604  0.19196  0.478 0.632847
모종사상지위2  0.64496   1.90591  0.26997  2.389 0.016894 *
연간소득      -0.44339   0.64186  0.15144 -2.928 0.003413 **
몸무게        -2.35795   0.09461  0.48570 -4.855 1.21e-06 ***
성적만족도2   -0.12859   0.87933  0.37766 -0.340 0.733485
성적만족도3    0.20089   1.22249  0.37706  0.533 0.594192
성적만족도4    1.25283   3.50022  0.38487  3.255 0.001133 **
가족구성1      0.63309   1.88341  0.27264  2.322 0.020230 *
가족구성2      1.42783   4.16964  0.51641  2.765 0.005694 **
가족여행       0.49084   1.63369  0.12341  3.977 6.97e-05 ***
damage         0.96824   2.63329  0.24386  3.970 7.17e-05 ***
학대           0.63506   1.88714  0.10822  5.868 4.41e-09 ***
학습           0.26107   1.29832  0.13533  1.929 0.053722 .
교사           0.27357   1.31465  0.13353  2.049 0.040491 *
가족연락       0.85560   2.35278  0.22746  3.762 0.000169 ***
가출친구       1.03921   2.82698  0.14350  7.242 4.42e-13 ***
교외동아리2   -0.50635   0.60269  0.28117 -1.801 0.071727 .
비행친구수    -0.25446   0.77533  0.11148 -2.283 0.022458 *
체험활동       0.63201   1.88139  0.13723  4.605 4.12e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

우리의 최종 모델 결과는 figure?와 같으며, 모델 유의성을 확인한 결과, p-value가 0.05 이하로 충분히 유의한 모델이라는 결정을 내릴 수 있었다.

Likelihood ratio test	446.9 on 22 df	p=< 2e-16
Wald test	556.6 on 22 df	p=< 2e-16
Score (logrank) test	1009 on 22 df	p=< 2e-16

4. 결론 및 한계, 의의

4.1 결론

우리는 '생존 분석'을 통해, 청소년들의 시기에 첫 가출 발생 요인을 분석해보았다. 청소년들의 첫 가출 발생 시점에 어떤 요인들이 복합적으로 작용하는 지를 검증하여, 첫 가출에 대한 효과적인 개입 방안을 제시할 수 있다. 앞서 설명했듯, 우리는 유의미한 변수들을 개인 요소, 가족 요소, 학교 요소로 나누었고 이러한 요소에 따라 가출을 예방할 수 있는 방안을 제시할 수 있었다.

청소년의 첫 가출 요인은 개인요소에서 부종사상지위가 높을 경우(즉, 근로를 하는 경우) 가출할 확률이 높았고, 가정 요소의 경우, 학대 정도가 심할수록 가출할 확률이 증가하였다. 마지막으로 학교 요소의 경우, 비행 피해가 많을수록 가출할 확률이 증가하는 등, 다양한 요인에 따른 가출 확률의 증감을 살펴볼 수 있었다. 이뿐만 아니라, 관련 정도가 큰 변수 등, 가출에 더욱 유의미하게 나타나는 요인들을 파악할 수 있었다.

요인분석을 통해 청소년의 첫 가출에는 학대에 대한 노출이 많을수록, 가족과 연락 빈도가 낮을수록, 가출 친구의 수가 많을수록 등 주변사람과의 관계에 관한 항목이 큰 요인이 유의미하게 작용하였다. 그 외에도, 부종사상지위 또는 가족구성 같이 가정환경에 대한 요소가 열악한 경우에 유의미하게 작용함을 알 수 있었다.

청소년 가출은 일시적이고 우발적으로 발생하는 것이 아니며, 이러한 가출 청소년은 각종 범죄에 쉽게 노출된다. 앞선 분석을 통해 도출된 유의미한 요인들을 살펴, 청소년들의 첫 가출을 예방할 수 있도록 프로그램과 청소년 문화 정착이 필요하다.

4.2 한계 및 의의

첫번째 한계로는, 가정을 만족하지 못하는 변수를 제외시켜야 했다는 점이다. 생존 분석을 진행하기 위해서는 다양한 가정을 만족시켜야만 했다. 그래서 만족시키지 못한 변수들의 경우에는 분석에서 사용하기가 어려웠다. 교우관계 변수의 경우, 선형성을 만족시키지 못하는 등, 가출과 유의미한 관계가 있어 보이는 변수를 제외시켜야 했다.

두번째 한계로는, 가출 요인에 대한 명확한 선후 관계 파악이 불가능하였다. 이는 설문 데이터의 한계로, 설문 기점에 따라 가출의 여부가 달라지기에 정확한 요인의 선후를 파악하기 어려웠다.

이에 반해, 우리는 설문데이터를 최대한 이용하도록 노력하였다. 앞선 전처리과정에서 보였듯, 다양한 N/A값 처리, 5개년 정보의 정리, 리커트 척도의 활용 등 주어진 설문데이터를 전적으로 이용하고자 노력하였다.

또한, 일방적으로 요인분석에서 사용되는 로지스틱 회귀분석과는 차별점을 두어, 콕스 회귀를 시행하여 요인분석을 진행하였다. 이에 따라 우리는 가출에 대한 "시기"를 고려할 수 있었다. 단순한 요인분석을 지나, 시기를 고려하여 유의미한 변수를 더 생성할 수 있었다.

마지막으로, 우리 분석은 사회적인 문제를 수치적으로 분석했다. 사회적 문제를 단순한 요인 분석으로 진행시키기보다는 데이터를 활용하여 수치적인 근거를 들어 가출 시기에 대한 요인을 분석해보았다.