우리 아이가 가출했어요!

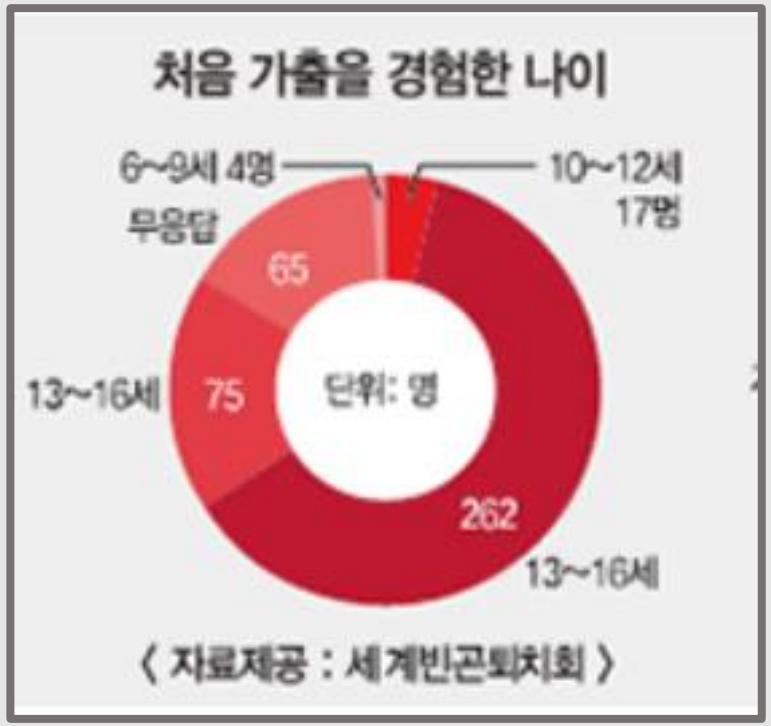
(생존분석을 이용한 가출 요인 분석)

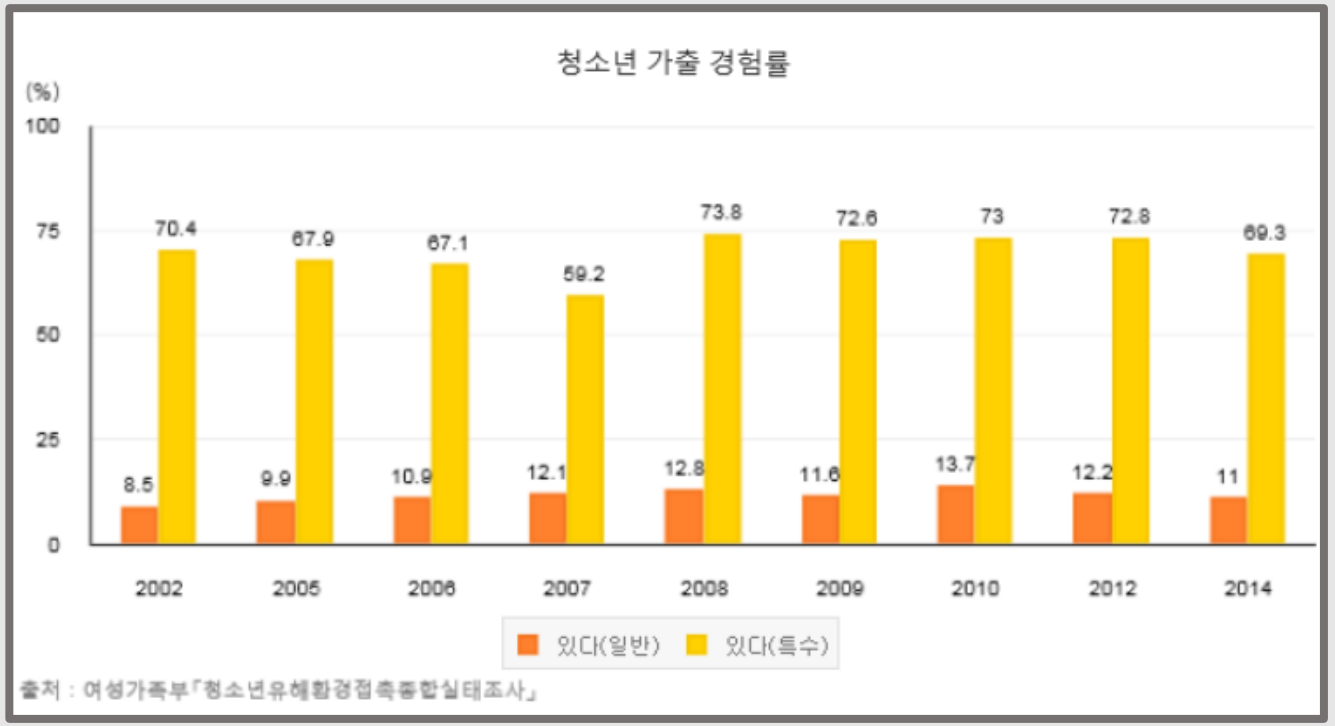
회귀분석팀 정은진, 한대룡, 이성희, 황원영, 최재영

목 차

01
주제 선정

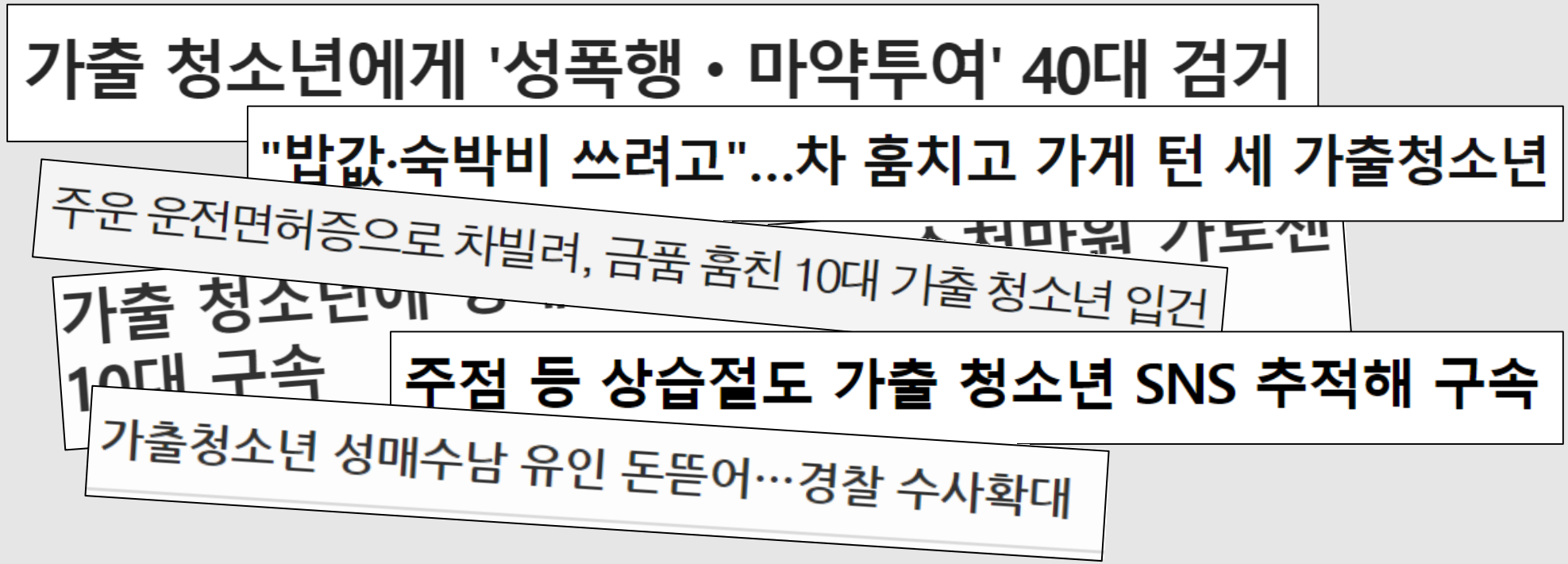
1 주제 선정 배경





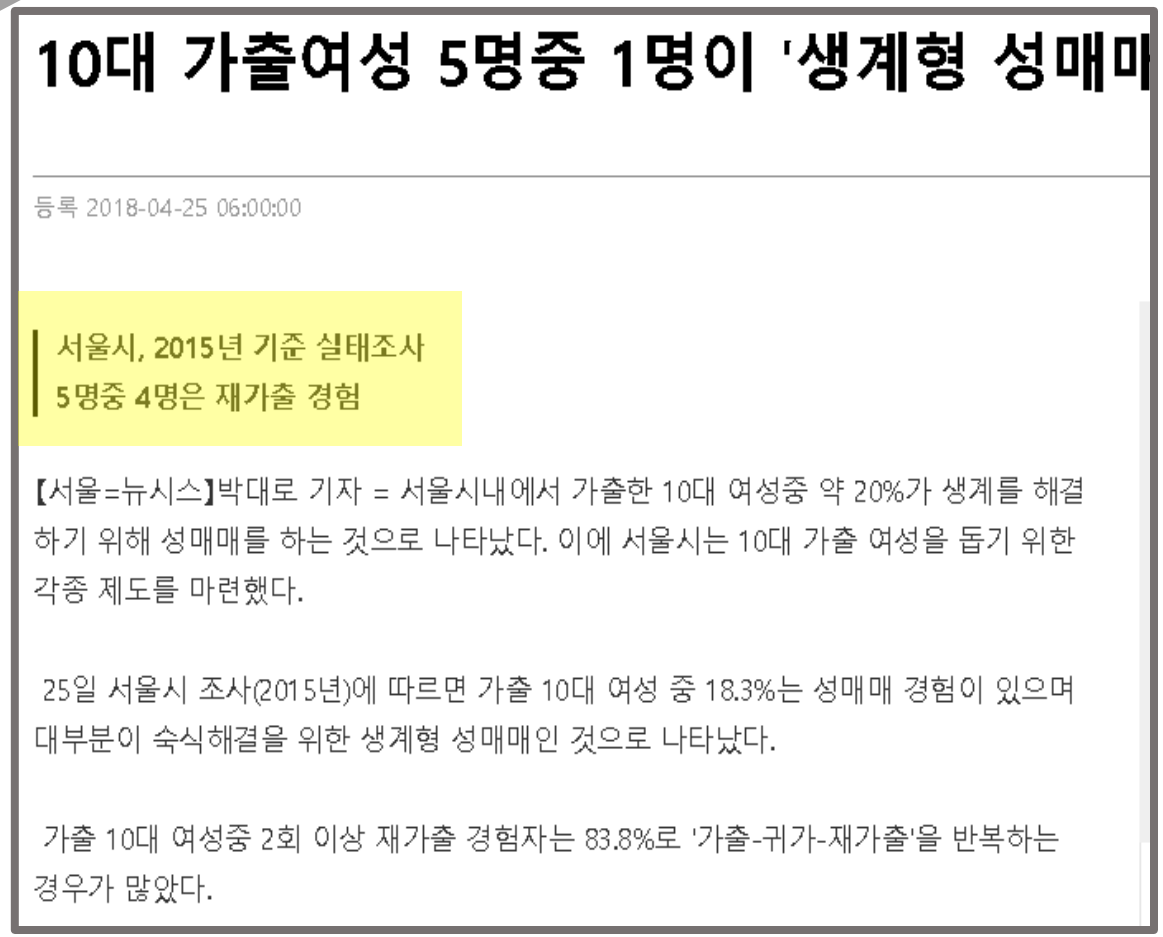
- ✓ 청소년 가출 경험률이 증가하는 추세
- ✓ 첫 가출을 경험하는 나이는 13-16세(중학생) 비율

1 주제 선정 배경



각종 범죄에 쉽게 노출되는 ‘가출 청소년’

1 주제 선정 배경



10대 가출여성 5명중 1명이 '생계형 성매매

등록 2018-04-25 06:00:00

서울시, 2015년 기준 실태조사
5명중 4명은 재가출 경험

【서울=뉴시스】박대로 기자 = 서울시내에서 가출한 10대 여성중 약 20%가 생계를 해결하기 위해 성매매를 하는 것으로 나타났다. 이에 서울시는 10대 가출 여성을 돕기 위한 각종 제도를 마련했다.

25일 서울시 조사(2015년)에 따르면 가출 10대 여성 중 18.3%는 성매매 경험이 있으며 대부분이 숙식해결을 위한 생계형 성매매인 것으로 나타났다.

가출 10대 여성중 2회 이상 재가출 경험자는 83.8%로 '가출-귀가-재가출'을 반복하는 경우가 많았다.

<서울시, 2015년 기준 실태조사>

- 5명 중 4명은 재가출 경험
- 가출 10대 여성중 2회 이상 재가출 경험자는 83.8%

"가출 - 귀가 - 재가출"의 굴레

〈표 16〉 첫 가출 이유

구분	응답자 수(명)	비율	케이스 비율
가족 간 불화, 폭력, 폭언	139	32.4%	63.8%
자유롭게 살고 싶어서	115	26.8%	52.8%
친구와 놀고 싶어서	74	17.2%	33.9%
학교 다니기 싫어서	56	13.1%	25.7%
집안형편이 어려워서	28	6.5%	12.8%
기타	13	3.0%	6.0%
성정체성 고민 때문에	4	0.9%	1.8%
합계	429	100%	196.8%

* 복수응답 허용

가출 여자청소년 공간 이용 및 폭력 피해 실태

여자청소년 212명을 대상으로 설문조사

청소년의 가출은 청소년의 특성 상
일시적이고 우발적으로 발생하는 문제가 아니라
개인, 가정, 학교 및 또래 등
다양한 생태체계에 속한 요인들이 복합적으로
작용하여 발생한다.

(박명숙, 2006; 박영호. 김태익, 2002; 배문조. 전귀연)

개인, 가정, 학교 및 또래 등 다양한 체계에 속한 요인들이 가출에 미치는 영향력을 실증적으로 검증할 필요가 있음

1주차

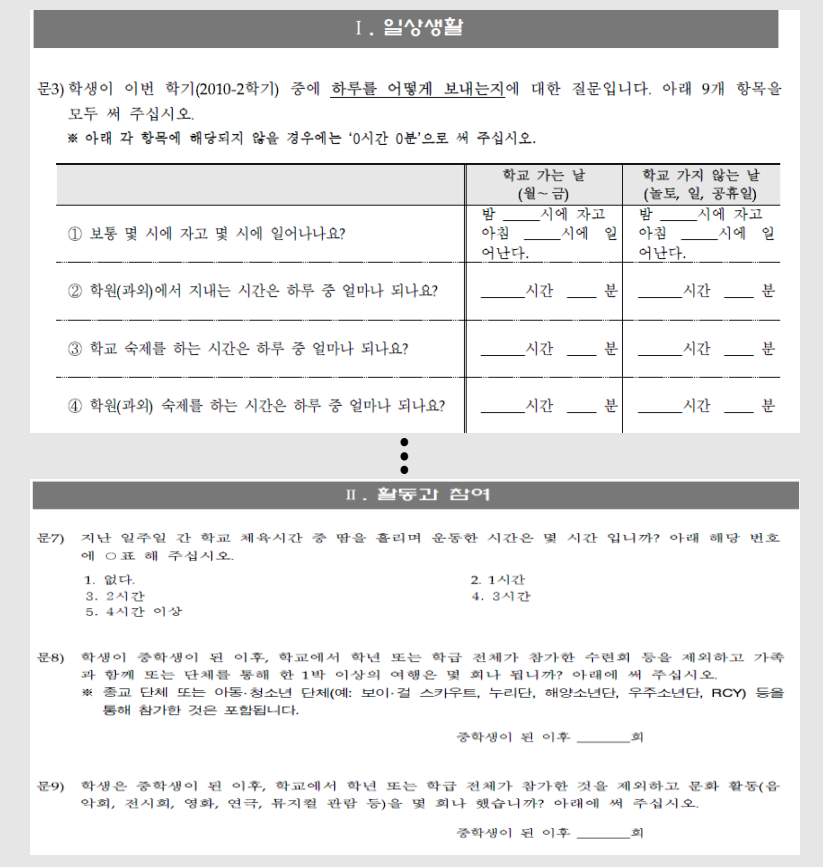
설문조사 데이터를 탐색해보고,
생존 분석에 대해 알아보자.

2주차

청소년들의 가출과 가출 시기에
미치는 요인을 분석해보자.

2 설문 데이터 소개

‘마이크로 데이터’ 한국 아동청소년 패널조사 > 중1 [2011-2015]



Ⅰ. 일상생활

문3) 학생이 이번 학기(2010-2학기) 중에 하루를 어떻게 보내는지에 대한 질문입니다. 아래 9개 항목을 모두 써 주십시오.
※ 아래 각 항목에 해당되지 않을 경우에는 ‘0시간 0분’으로 써 주십시오.

	학교 가는 날 (월~금)	학교 가지 않는 날 (놀토, 일, 공휴일)
① 보통 몇 시에 자고 몇 시에 일어나나요?	밤 ____시에 자고 아침 ____시에 일어난다.	밤 ____시에 자고 아침 ____시에 일어난다.
② 학원(과외)에서 지내는 시간은 하루 중 얼마나 되나요?	____시간 ___ 분	____시간 ___ 분
③ 학교 숙제를 하는 시간은 하루 중 얼마나 되나요?	____시간 ___ 분	____시간 ___ 분
④ 학원(과외) 숙제를 하는 시간은 하루 중 얼마나 되나요?	____시간 ___ 분	____시간 ___ 분

Ⅱ. 활동과 참여

문7) 지난 일주일 간 학교 체육시간 중 땀을 흘리며 운동한 시간은 몇 시간 입니까? 아래 해당 번호에 ○표 해 주십시오.

1. 없다.
2. 1시간
3. 2시간
4. 3시간
5. 4시간 이상

문8) 학생이 중학생이 된 이후, 학교에서 학년 또는 학급 전체가 참가한 수련회 등을 제외하고 가족과 함께 또는 단체를 통해 한 1박 이상의 여행은 몇 회나 됩니까? 아래에 써 주십시오.
※ 종교 단체 또는 아동·청소년 단체(예: 보이·걸 스카우트, 누리단, 해양소년단, 우주소년단, RCY) 등을 통해 참가한 것은 포함됩니다.

중학생이 된 이후 _____회

문9) 학생은 중학생이 된 이후, 학교에서 학년 또는 학급 전체가 참가한 것을 제외하고 문화 활동(음악회, 전시회, 영화, 연극, 뮤지컬 관람 등)을 몇 회나 했습니까? 아래에 써 주십시오.

중학생이 된 이후 _____회

- ✓ 한국 아동 · 청소년 패널조사(KCYPS)에서 실시한 조사
- ✓ 조사원과의 개별 접촉을 통한 면접조사
- ✓ 중1(2,351명)을 선정하여 7년에 걸쳐 매년 추적조사

(‘가출’ 문항이 없는 2010년, 2016년은 제외)

제1차 (2010)	제2차 (2011)	제3차 (2012)	제4차 (2013)	제5차 (2014)	제6차 (2015)	제7차 (2016)
중1	중2	중3	고1	고2	고3	대1

2011-2015 수집

2 설문 데이터 소개: 표본 추출

'마이크로 데이터' 한국 아동청소년 패널조사 > 중1 [2011-2015]

구분		중1 패널	
		표본학교수	원표본(명)
전체		78	2,351
서울특별시		8	234
광역시	부산광역시	5	133
	대구광역시	4	155
	인천광역시	5	163
	광주광역시	4	95
	대전광역시	4	107
	울산광역시	4	115
시군부	경기도	10	346
	강원도	4	114
	충청북도	4	126
	충청남도	4	94
	전라북도	4	115
	전라남도	4	115
	경상북도	5	159
	경상남도	5	145
	제주도	4	135

- 16개 시도 중학교 1학년 학생 수에 비례하여 지역별로 표본 수 할당
- 조사대상 학교는 '확률비례 통계추출법'에 의거해 추출하여 '시도 및 도시규모별'로 조사대상 학교 선정
- 조사대상 학년이 최소한 2개 학급 이상인지, 학생 수가 50명인지 등을 확인한 후, 무작위 선정

*확률비례 통계추출법 : 모집단을 구성하고 있는 집락의 규모가 심하게 차이가 날 경우 각 집락을 불균등 확률로 뽑는 추출방법

3 설문 데이터 소개: 설문지 구성

① 가족 관계와 같은 객관적인 질문

항목	부	모	보호자 (부모님이 안 계신 경우)
1. 출생 연도	년	년	년
2. 교육 수준	1. 중졸 이하 2. 고졸 3. 전문대 졸 4. 대졸 5. 대학원 졸	1. 중졸 이하 2. 고졸 3. 전문대 졸 4. 대졸 5. 대학원 졸	1. 중졸 이하 2. 고졸 3. 전문대 졸 4. 대졸 5. 대학원 졸
3. 근로 여부	1. 일을 하고 있다. 2. 일을 하고 있지 않다.	1. 일을 하고 있다. 2. 일을 하고 있지 않다.	1. 일을 하고 있다. 2. 일을 하고 있지 않다.

② 주관적인 감정을 묻는 질문

나는 …	매우 그렇다	그런 편이다	그렇지 않은 편이다	전혀 그렇지 않다
⑪ 우리 반 아이들과 잘 어울린다.	1	2	3	4
⑫ 친구와 다투었을 때 먼저 사과한다.	1	2	3	4
⑬ 내 짝이 교과서나 준비물을 안 가져왔을 때 함께 보거나 빌려 준다.	1	2	3	4
⑭ 친구가 하는 일을 방해한다.	1	2	3	4
⑮ 놀이나 모둠활동을 할 때 친구들이 내 말을 잘 따라 준다.	1	2	3	4

③ 개인의 경험과 관련된 질문

	피해경험 여부		피해경험이 있다면, 그 횟수를 써 주십시오.
① 심한 놀림이나 조롱당하기	1.있다	2.없다	지난 1년 동안 _____ 회
② 집단따돌림(왕따)당하기	1.있다	2.없다	지난 1년 동안 _____ 회
③ 심하게 맞기(폭행)	1.있다	2.없다	지난 1년 동안 _____ 회
④ 협박당하기	1.있다	2.없다	지난 1년 동안 _____ 회
⑤ 돈이나 물건 뺏기기(삥뜯기기)	1.있다	2.없다	지난 1년 동안 _____ 회
⑥ 성폭행이나 성희롱	1.있다	2.없다	지난 1년 동안 _____ 회

3 설문 데이터 소개: 패널 조사

횡단조사? 종단조사? 패널조사?

횡단 조사

일정 시점에서 특정 표본이 가지고 있는 특성을 파악하는 것

2018년

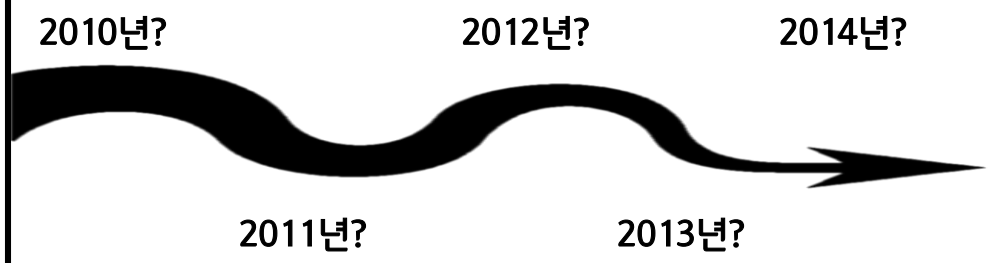
성별 나이 소득 만족도

성취감 우울한 정도

+

종단 조사

시간의 흐름에 따라 조사 대상의 변화를 측정하는 것



패널 조사

동일한 주제와 동일한 응답자에 대해 반복적으로 조사하는 것

시계열자료와 횡단면자료를 하나로 합쳐 놓은 자료

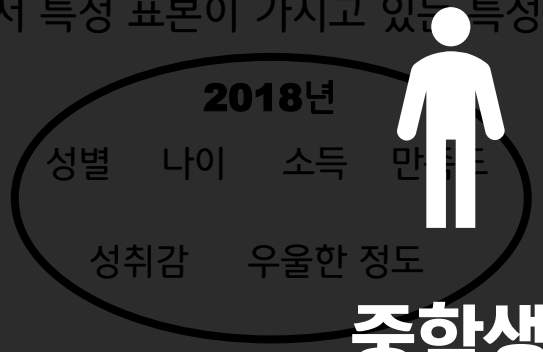
3 설무데이터 소개: 패널 조사

우리 설문 데이터는?

#패널 조사

횡단 조사

일정 시점에서 특정 표본이 가지고 있는 특성을 파악하는 것



중학생

종단 조사

시간의 흐름에 따라 조사 대상의 변화를 측정하는 것

고등학생

패널 조사

종단적 측면 : 2011년 ~ 2015년 (5년)

동일한 주제와 동일한 응답자에 대해 반복적으로 조사하는 방법

횡단적 측면 : 일상생활 / 활동과 참여 / 학습 및 학교생활 /사회정서 / 부모 및 친구관계 / 지역사회와 공동체

3 설문 데이터 소개: 리커트 척도

리커트 척도란? 어떤 질문에 대하여 "긍정/부정(만족/불만족)"의 정도를 측정하는 척도

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

3 설문 데이터 소개

우리 설문 데이터는?

#리커트 척도

"의 정도를 측정하는 척도

[문18] 귀하가 부모님(부모님이 안 계신 경우에는 보호자)을 어떻게 생각하는지에 대한 질문입니다. 다음 각 문항에 대하여 자신에게 해당되는 항목에 응답해 주십시오.

부모님(보호자)께서는 …	매우 그렇다	편이다	않은 편이다	전혀 그렇지 않다
1) 내가 일과 후에 어디에 가는지 알고 계신다	1	2	3	4
2) 내가 시간을 어떻게 보내는지 알고 계신다	1	2	3	4
3) 나의 의견을 존중해 주신다	1	2	3	4

✓리커트 척도로 구성

02
데이터 전처리

리커트 척도는 어떻게 활용해야 할까?

문26) 학생이 친구들을 어떻게 생각하는지에 대한 질문입니다. 아래 각 항목의 해당 칸에 ○표 해 주십시오.

	매우 그렇다	그런 편이다	그렇지 않은 편이다	전혀 그렇지 않다
① 내 친구들은 나와 이야기를 나눌 때 내 생각을 존중해 준다.	1	2	3	4
② 내 친구들은 내가 말하는 것에 귀를 기울인다.	1	2	3	4
③ 나는 내 친구들에게 내 고민과 문제에 대해 이야기한다.	1	2	3	4
④ 내 친구들은 나를 잘 이해해 준다.	1	2	3	4
⑤ 나는 속마음을 털어놓고 싶을 때 친구들에게 말할 수 있다.	1	2	3	4

· 모든 항목을 각각의 변수로 사용할 경우 **다중공선성** 발생!

· 리커트 척도는 흔히 질문들의 **합산 값** or **평균**의 차이로 통계적 검증

▶우리 설문 데이터는? **#평균**

① 리커트 척도 문항의 수치화

부모님의 양육방식 - 방임

질문 문항 1 ① ② ③ ④
(다른 일보다 나를 더 중요하게 생각하신다)
질문 문항 2 ① ② ③ ④
(내가 학교에서 어떻게 생활하는지 관심을 갖고 물어보신다)
질문 문항 3 ① ② ③ ④
(내 몸이나, 옷, 이불 등이 깨끗하도록 항상 신경 쓰신다)
⋮

→

부모님의 양육방식 - 방임

학생 1	3
학생 2	2.5

방임, 학대, 학습, 교우, 교사, 사회성 변수 생성!

리커트 척도 문항 질문의 방향성 통일

부모님의 양육방식 - 학대

내가 잘못하면 부모님(보호자)께서는 무조건 때리려고 하신다

매우 그렇다	그런 편이다	그렇지 않은 편이다	전혀 그렇지 않다
1	2	3	4

학교적응 - 학습활동

학교 수업 시간이 재미있다

매우 그렇다	그런 편이다	그렇지 않은 편이다	전혀 그렇지 않다
1	2	3	4

모델링 과정에서 변수 해석의 편리성을 위해 **방향성**을 통일

② 범주형 문항의 수치화

연간 피해경험 유무
피해경험 종류1 유무 ① ② (심한 놀림이나 조롱 당하기)
피해경험 종류2 유무 ① ② (집단 따돌림(왕따)당하기)
피해경험 종류3 유무 ① ② (심하게 맞기(폭행))
⋮

→

연간 피해경험 유무	
학생 1	2회
학생 2	0회

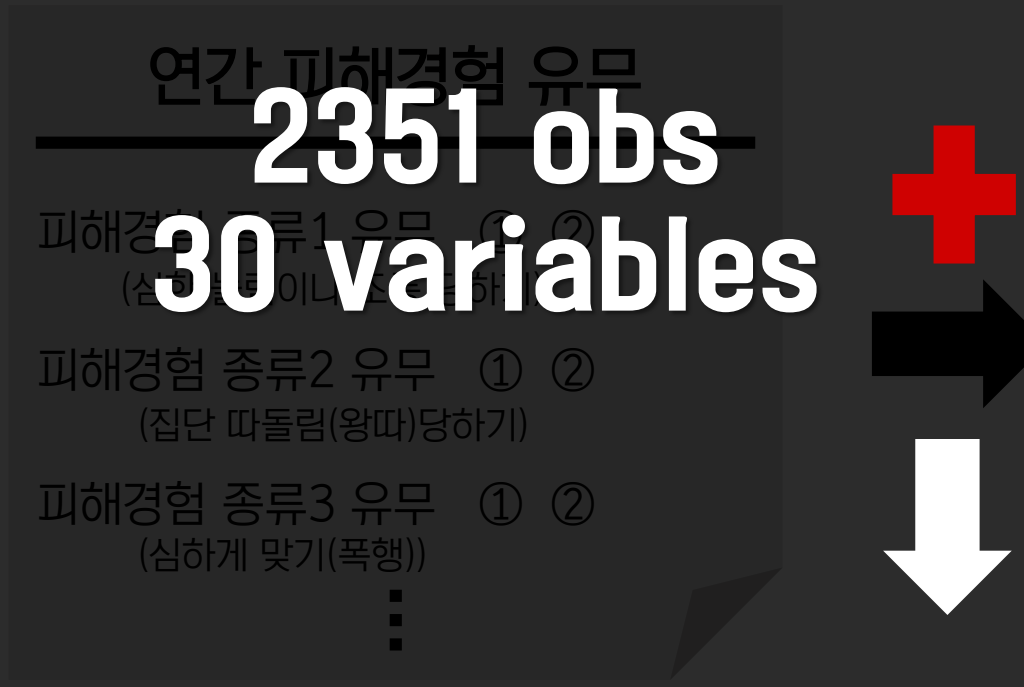
피해종류횟수변수 생성!

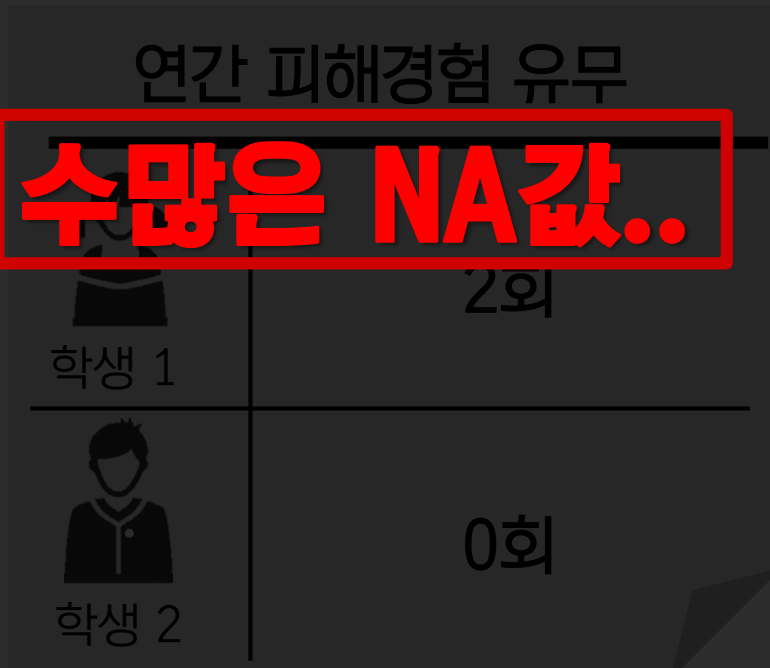
통합된 우리 설문 데이터는?

2351 obs
30 variables

+

수많은 NA값..





패널데이터의 장점을 살려 N/A 값을 최대한 살려보자!

2 N/A값 처리

NA값 처리방식

- 추적조사 특성 상 대체 불가한 행 삭제
- 전후 년도 데이터 활용
  - 전년도 데이터 기입
  - 해당 이외 년도 평균값 기입
  - 해당 이외 년도 값 랜덤하게 기입
- MICE

2 N/A값 처리

NA값 처리방식 ① 행 삭제

```
> rawone[is.na(rawone$V2),]
         V1 V2 V3 V4 V5 V6 V7 V8 V9   V10 V11 V12 V13 V14 V15 V16 V17 V18 V19 V20 V21 V22 V23 V24 V25 V26 V27 V28 V29 V30 V31
93   108132 NA NA NA NA NA NA NA NA    NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
144  117120 NA NA NA NA NA NA NA NA    NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
170  117428 NA NA NA NA NA NA NA NA    NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
202  118414 NA NA NA NA NA NA NA NA    NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
285  128516 NA NA NA NA NA NA NA NA    NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
303  128603 NA NA NA NA NA NA NA NA    NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
345  135912 NA NA NA NA NA NA NA NA    NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
```

특정 연도에 모든 정보가 기입되어 있지 않은 행

➡ 추적조사 특성상 대체 불가한 행이므로 삭제

2 N/A값 처리

NA값 처리방식 ② 다른 해의 값으로 대체

IDcode	2011_edu	2012_edu	2013_edu	2014_edu	2015_edu
107730	4	4	4	4	4
107731	4	4	NA	NA	4
107732	2	2	2	2	2
107733	4	4	4	4	4
107734	4	4	4	4	4
107735	4	4	4	4	4

IDcode	2011_edu	2012_edu	2013_edu	2014_edu	2015_edu
107730	4	4	4	4	4
107731	4	4	4	4	4
107732	2	2	2	2	2
107733	4	4	4	4	4
107734	4	4	4	4	4
107735	4	4	4	4	4

Ex) 부모님 최종학력, 가족구성, 다문화가정 여부, 형제자매 유무

연도에 따라 바뀌지 않는 변수 → 전년도에 기입 된 값으로 대체

2 N/A값 처리

NA값 처리방식 ② 다른 해의 값으로 대체

IDcode	2011_income	2012_income	2013_income	2014_income	2015_income
117406	6000	6000	6100	6500	7000
117409	7000	NA	10000	10000	10000
117410	NA	9000	8000	8500	8500
117412	7000	7000	7000	7500	8000
117413	4000	5000	5000	5000	5000
117414	3600	3600	4000	4000	4000
117415	2300	2000	2000	1800	1800

IDcode	2011_income	2012_income	2013_income	2014_income	2015_income
117406	6000	6000	6100	6500	7000
117409	7000	9250	10000	10000	10000
117410	8500	9000	8000	8500	8500
117412	7000	7000	7000	7500	8000
117413	4000	5000	5000	5000	5000
117414	3600	3600	4000	4000	4000
117415	2300	2000	2000	1800	1800

Ex) 가계소득수입

연도에 따라 바뀌는 값 → **평균 값으로 대체!**

2 N/A값 처리

NA값 처리방식 ② 다른 해의 값으로 대체

IDcode	2011_health	2012_health	2013_health	2014_health	2015_health
108230	2	2	2	2	2
108231	2	3	3	NA	2
108232	2	2	2	2	2
108233	3	2	4	3	3
108234	2	2	2	2	2

IDcode	2011_health	2012_health	2013_health	2014_health	2015_health
108230	2	2	2	2	2
108231	2	3	3	3	2
108232	2	2	2	2	2
108233	3	2	4	3	3
108234	2	2	2	2	2

Ex) 부모님 건강상태

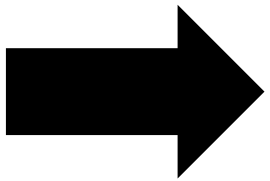
연도에 따라 바뀌는 값 → **랜덤한 값으로 대체!**

N/A값 처리 결과

N/A값 처리 후, N/A값이 몇 개 남았을까?

```
> colSums(is.na(rawone))
```

최종학력(부)	256
최종학력(모)	228
종사상지위(부)	260
종사상지위(모)	232
가구 연간 소득	140
보호자건강상태	83
가족구성	82
다문화가정	82
형제자매	82



N/A 값을 처리한 new 데이터

```
> colSums(is.na(one))
```

최종학력(부)	39
최종학력(모)	30
종사상지위(부)	42
종사상지위(모)	30
가구 연간 소득	28
보호자건강상태	14
가족구성	7
다문화가정	7
형제자매	7

N/A값 처리 결과


```
> colSums(is.na(rawone))
```

최종학력(부)	256
최종학력(모)	228
종사상지위(부)	260
종사상지위(모)	232
가구 연간 소득	140
보호자건강상태	83
가족구성	82
다문화가정	82
형제자매	82

남은 결측값을 처리해보자!

```
> colSums(is.na(one))
```

최종학력(부)	39
최종학력(모)	30
종사상지위(부)	42
종사상지위(모)	30
가구 연간 소득	28
보호자건강상태	14
가족구성	7
다문화가정	7
형제자매	7

4 NEW N/A값 처리 방법

결측값의 종류가 다양하다고..?

▶결측값(N/A) 종류

① **MCAR**(Missing Completely at Random)

② **MAR**(Missing at Random)

③ **MNAR**(Missing Not at Random)

4 NEW N/A값 처리 방법

MCAR (Missing Completely at Random)

결측이 랜덤으로 발생함

해당 변수나 다른 변수에 영향 받지 않음

MAR (Missing at Random)

결측 여부가 다른 변수와 연관 있음

Ex) 소득수준이 낮은 아이들이 시험점수 무응답

MNAR (Missing Not at Random)

결측 여부가 해당 변수의 값에 의해 결정됨

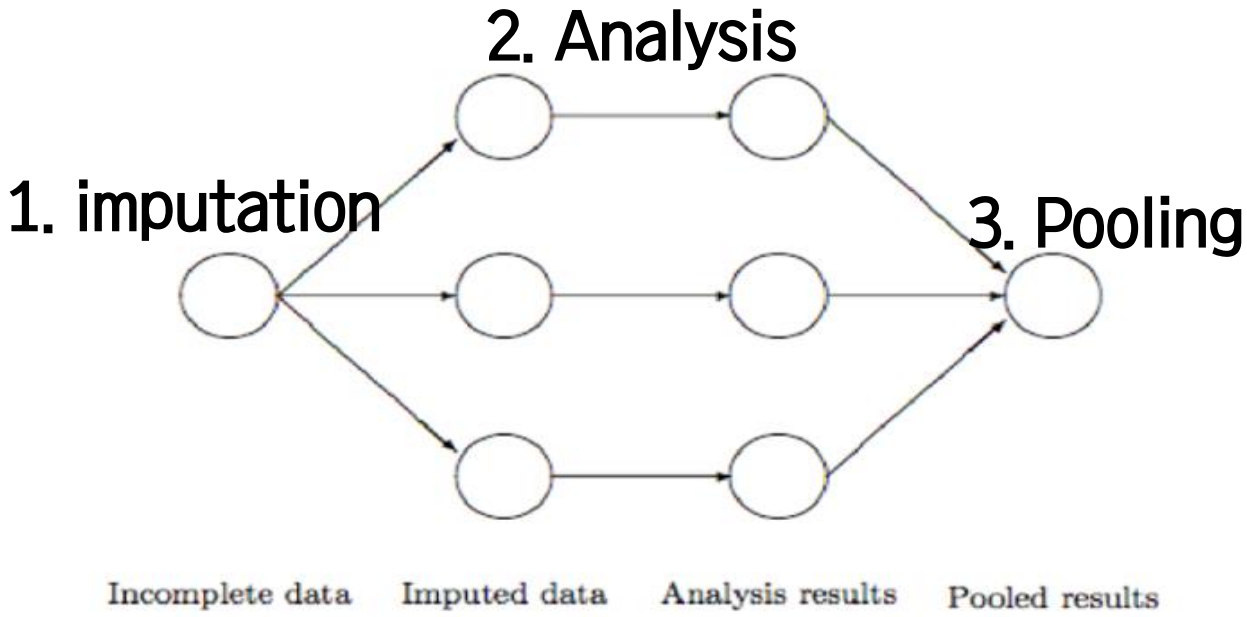
Ex) 소득수준이 낮은 사람들이 소득수준 무응답

MAR 과 MNAR은 제외하고 분석하면 분석결과가 편향 될 수 있음

NEW N/A값 처리 방법

MICE(Multiple Imputation by Changed Equation)

다중 대체 방법: 여러 변수에 걸쳐 존재하는 결측값을 관찰값을 이용하여 예측한다.



STEP1. 여러 개의 결측치 대치 세트(m) 생성

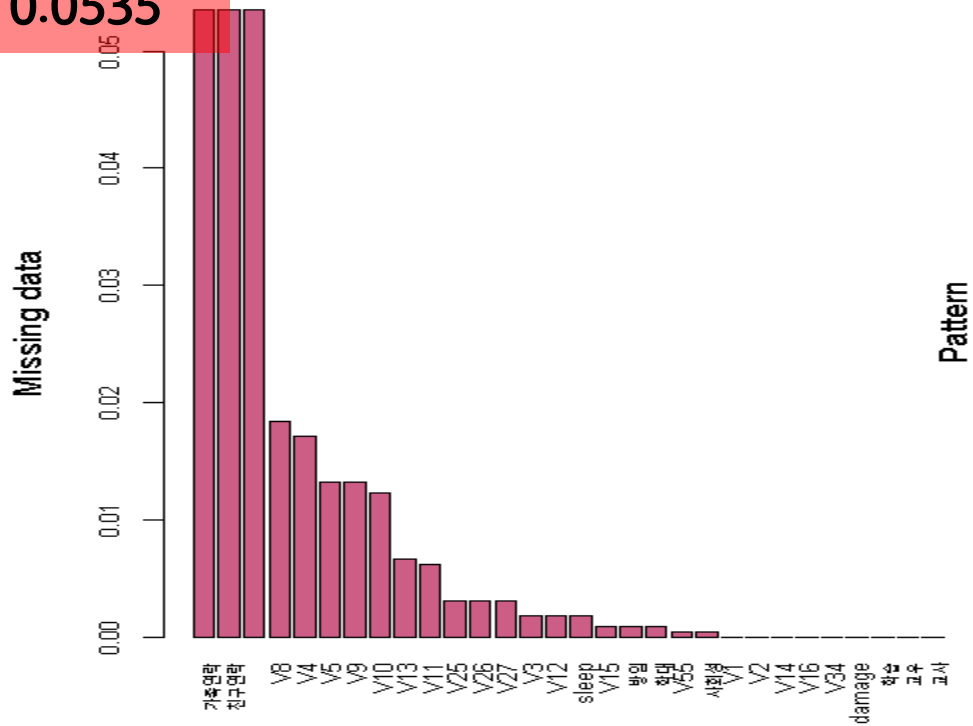
STEP2. with함수로 통계모델링

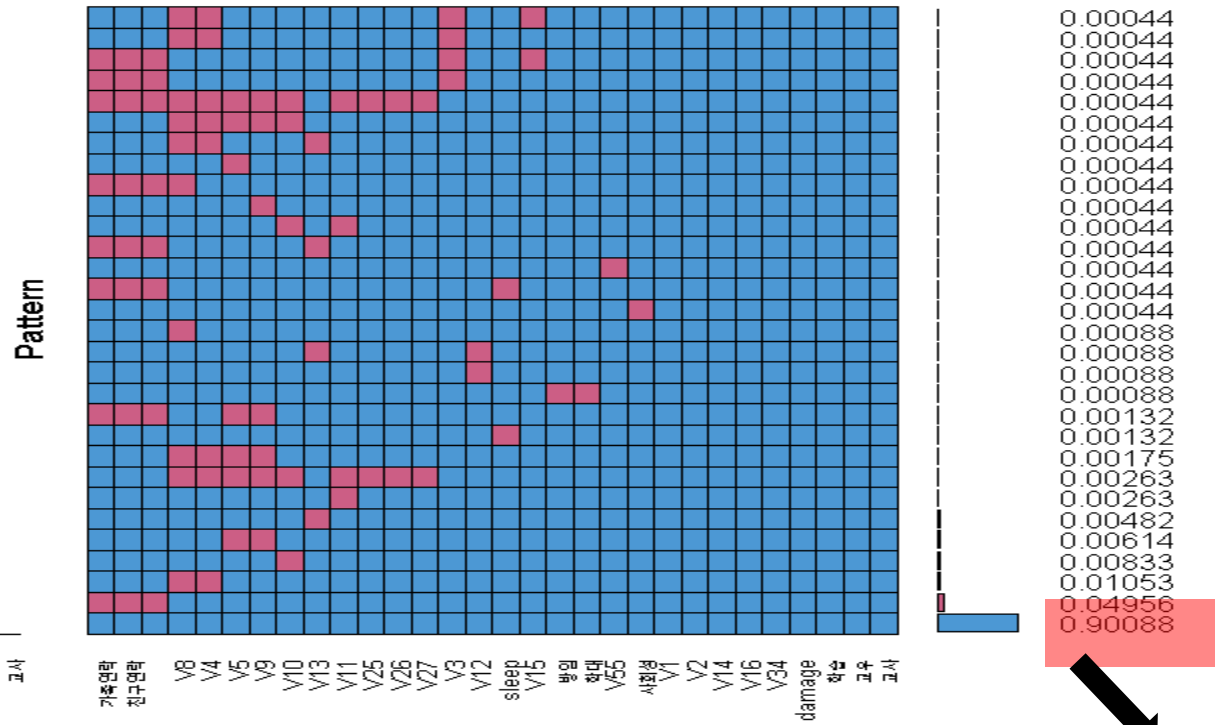
STEP3. Pool함수로 m개의 대치세트 평균해서 결과 도출!

NEW N/A값 처리 방법

결측 데이터 사전 탐색 : MICE는 데이터 시각화 하기에 좋은 function이 많다.

0.0535



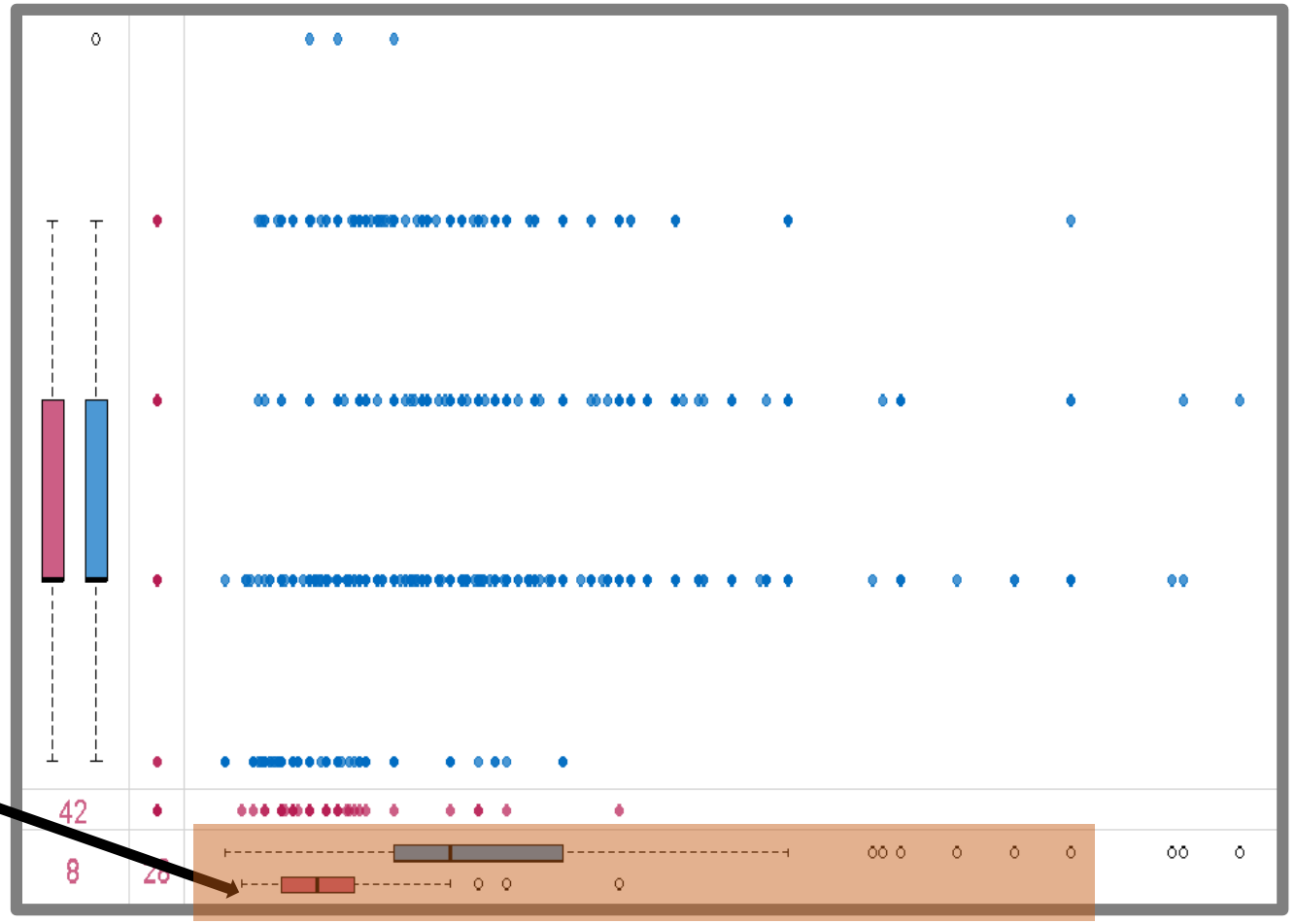


누락된 값이 없는 Dataset이 90%!

0.90088

NEW N/A값 처리 방법

파란 BOXPLOT : 다른 관찰값이 있을 때
빨간 BOXPLOT : 다른 관찰값이 없을 때



연간소득이 낮은 집단에서
"아버지의 종사상 지위"에 대한
결측값이 주로 나타나고 있네!

NEW N/A값 처리 방법

결측 값을 채워 보자!

어떤 방식으로 채워 넣을까?

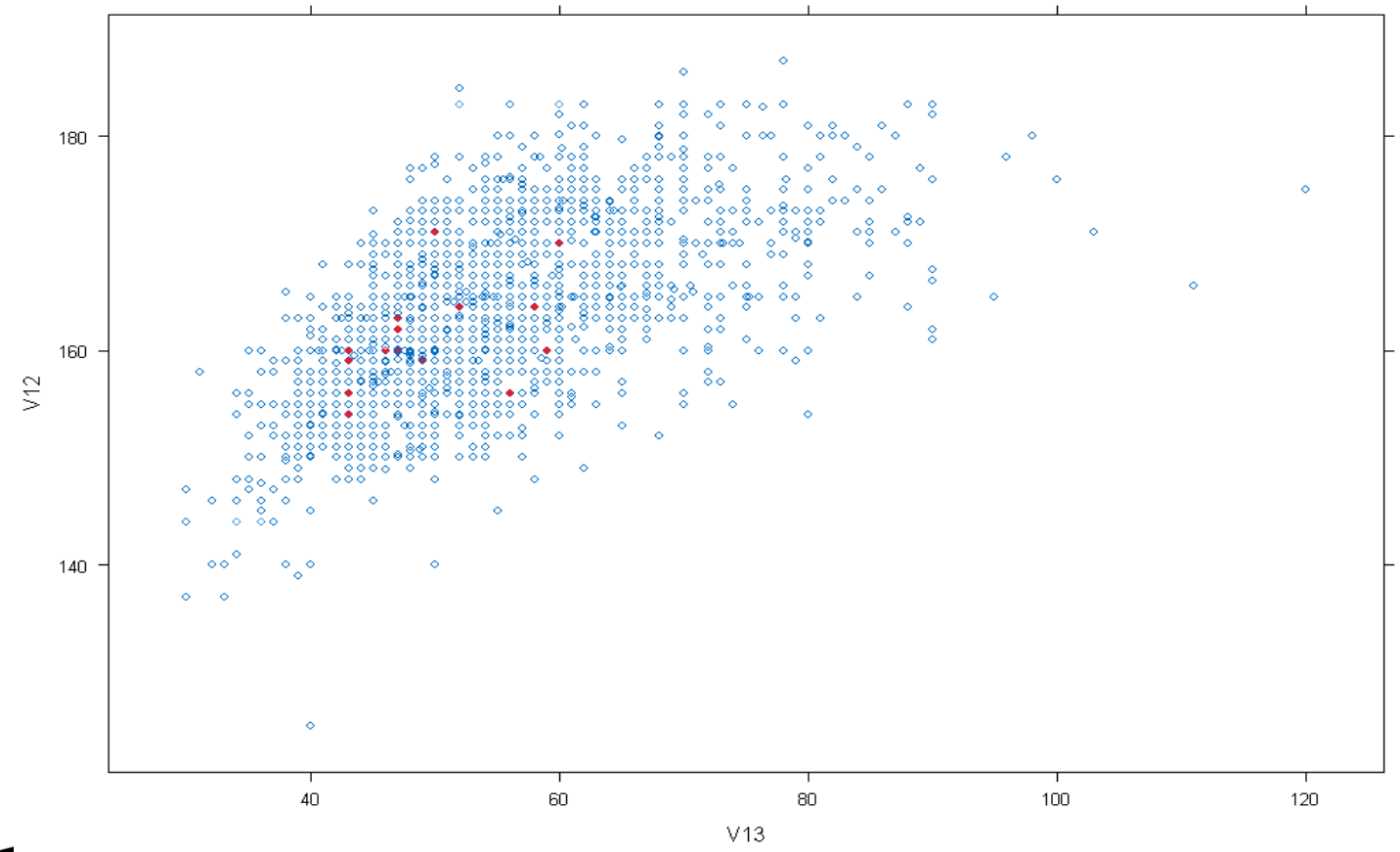
```
methods(mice)
 1] mice.impute.2l.bin       mice.impute.2l.lmer      mice.impute.2l.norm
 4] mice.impute.2l.pan       mice.impute.2lonly.mean  mice.impute.2lonly.norm
 7] mice.impute.2lonly.pmm   mice.impute.cart         mice.impute.jomoImpute
10] mice.impute.lda          mice.impute.logreg       mice.impute.logreg.boot
13] mice.impute.mean         mice.impute.midastouch   mice.impute.norm
16] mice.impute.norm.boot    mice.impute.norm.nob     mice.impute.norm.predict
```

method	
PMM (Predictive Mean Matching)	숫자 변수
logreg (logistic regression)	이진 변수
Polyreg (Bayesian polytomous regression)	인자 변수 (level >=2)
CART (classification And Regresson Tree)	연속형, 범주형 둘다 가능

* 각 변수별로 다른 분포 가정
* 변수별로 다른 대체 모델 사용 가능
* 범주형 변수도 사용 가능

NEW N/A값 처리 방법

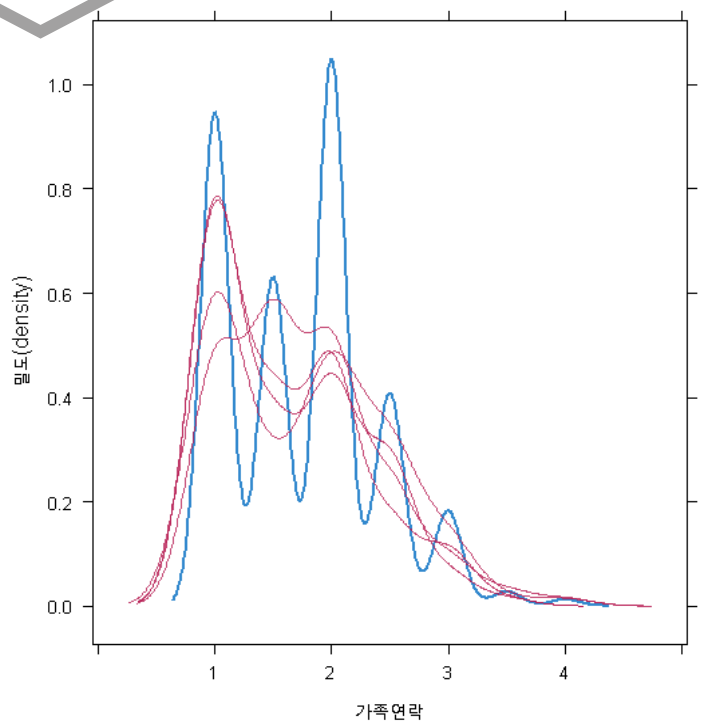

```
# step2. imputed vs observed ( 두 변수 사이)
xyplot(sample_mice ,V12~V13,par.settings = list(superpose.symbol = list(pch = 10, cex = 0.7)))
```



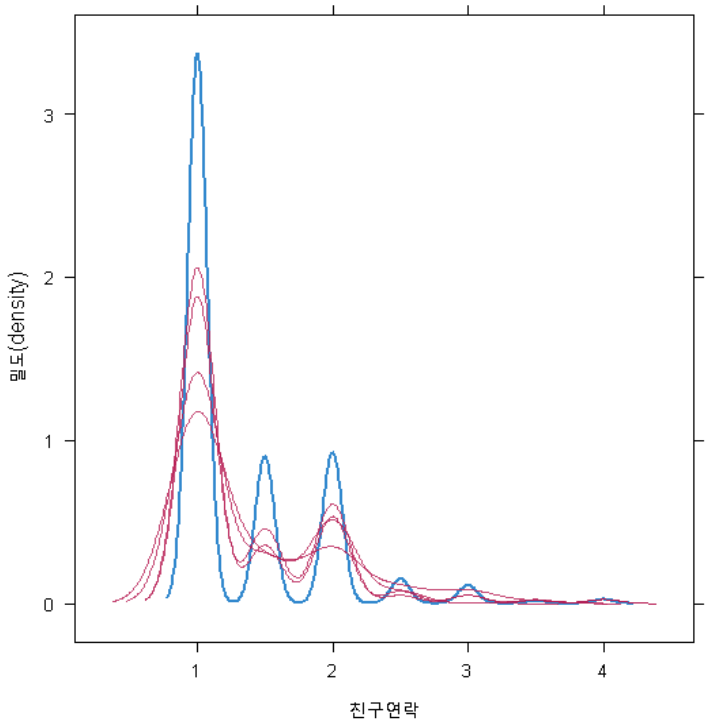
실제값 vs 대체된 값 비교

키(V12)와 몸무게(V13)
두 변수 사이의 관계를
알아보자.

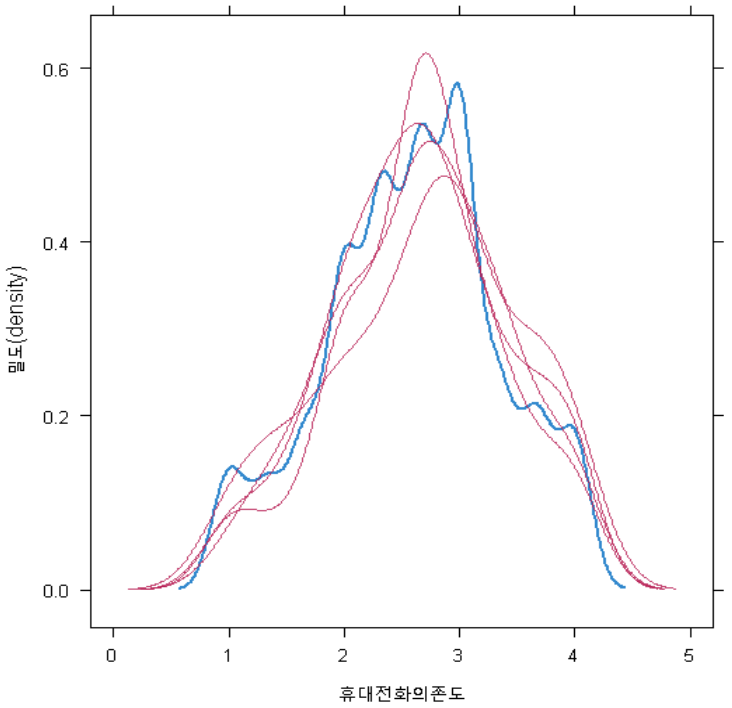
NEW N/A값 처리 방법



가족 연락



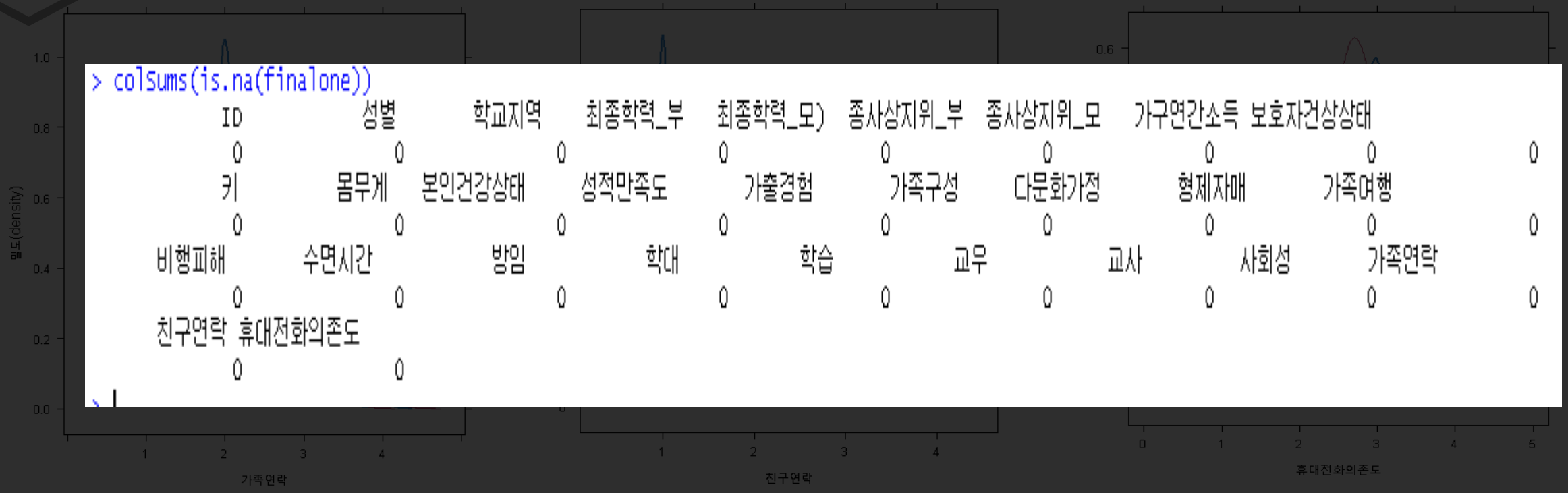
친구 연락



휴대전화의존도

실제값의 분포 vs 대체된 값의 분포

NEW N/A값 처리 방법




```
> colSums(is.na(finalone))
          ID         성별     학교지역    최종학력_부    최종학력_모)  종사상지위_부  종사상지위_모  가구연간소득 보호자건상상태
           0            0            0            0            0            0            0            0            0
          키       몸무게  본인건강상태     성적만족도       가출경험       가족구성     다문화가정       형제자매       가족여행
           0            0            0            0            0            0            0            0            0
  비행피해     수면시간         방임          학대          학습           교우           교사         사회성       가족연락
           0            0            0            0            0            0            0            0            0
 친구연락 휴대전화의존도
           0            0
>
```

실제 값의 분포
예측 값의 분포

```
# step2. distribution (한 변수)
densityplot(sample_mice , ~가족연락)
densityplot(sample_mice , ~친구연락)
densityplot(sample_mice , ~휴대전화의존도)
```

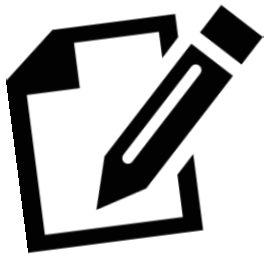
N/A가 없다!

5 데이터 통합

비행 연간 행동 경험-**최초 가출 발생 해** 기준 설문조사 수집

2014년 가출

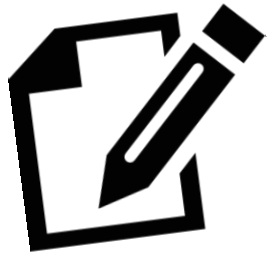
2011년
2012년
2013년
2014년
2015년



2014년 설문조사만 수집

2011년 가출

2011년
2012년
2013년
2014년
2015년



2011년 설문조사만 수집

우리 데이터는?

비행 연간 행동 경험-가출 발생 해 기준 설문조사 수집

1=가출 2=가출을 하지 않음

IDcode	2011_runaway	2012_runaway	2013_runaway	2014_runaway	2015_runaway
107821	1	2	2	2	2
107822	2	2	2	2	2
107824	2	1	2	2	2
107825	2	2	2	2	2
107826	2	2	2	2	2
107827	2	2	2	2	2

2011년 가출

- 107821 : 가출을 한 2011년도 자료 사용
- 107824 : 가출을 한 2012년도 자료 사용
- 가출 경험 없는 학생은 2015년도 자료 사용

2011년 설문조사만 수집

5. 데이터 통합

우리 데이터는?

비행 연간 행동 경험-가출 발생 해 기준 설문조사 수집

IDcode		교우	교사	사회성	가족연락	친구연락	휴대전화의존도	age
14210		1.666667	2.000000	2.000000	1.0	1.0	2.000000	16
14323		1.333333	2.000000	3.333333	2.5	1.5	3.000000	18
14325		1.000000	1.000000	2.333333	1.0	1.0	2.666667	19
14327	.....	1.666667	2.666667	2.000000	2.0	1.0	1.333333	16
14413		2.000000	2.000000	1.666667	2.0	2.0	2.666667	16
14521		1.666667	2.000000	2.333333	2.0	1.0	1.000000	16
14601		2.000000	2.000000	2.666667	1.5	2.5	2.666667	18

▶14210학생: 16세에 첫 가출을 했기에, 이 학생의 16세 당시 설문 조사 내용을 이용

첫 번째 가출이 발생한 해의 정보로 통합-(age가 다르게 통합)

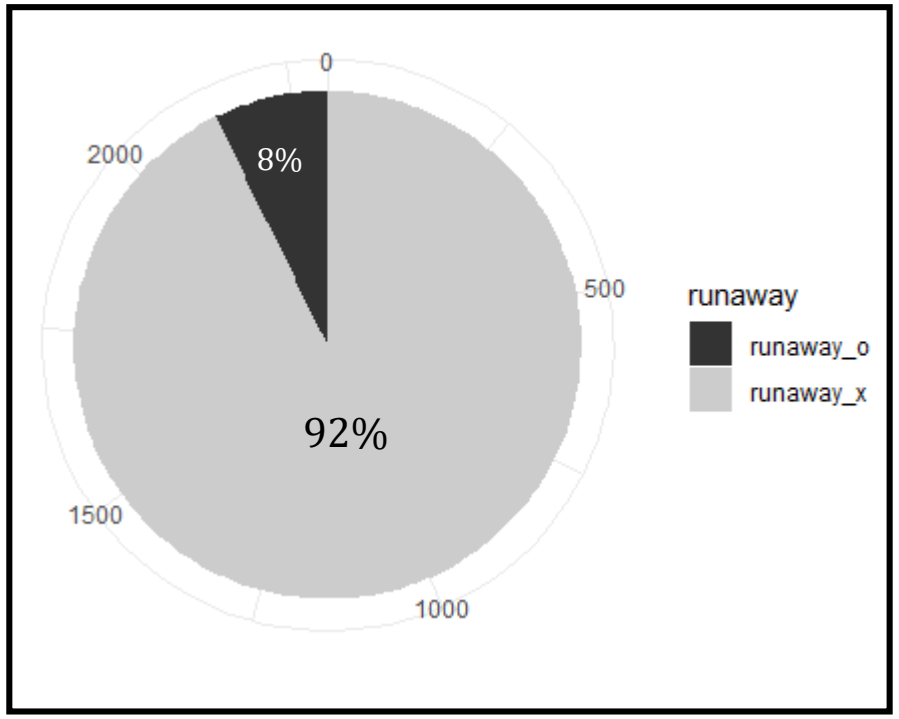
2011년
2012년
2013년
2014년
2015년

2011년 가출

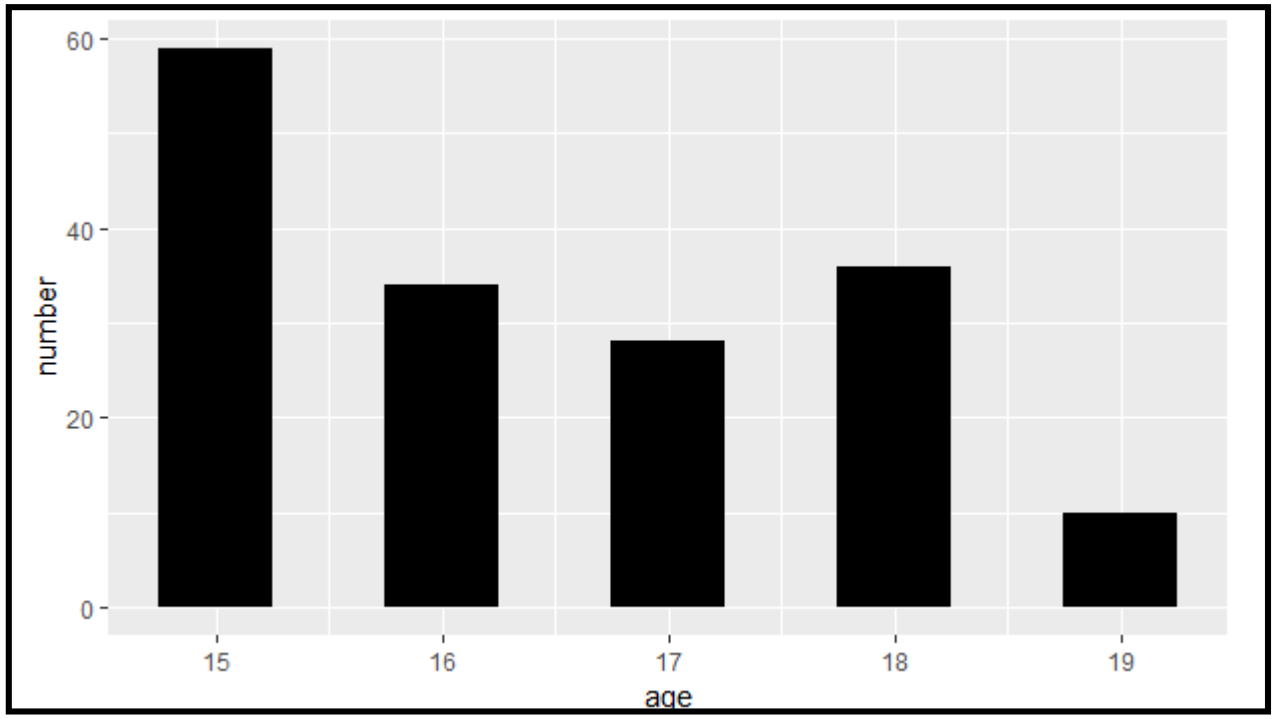
**왜 첫번째 가출의 해의 정보로 통합했을까?
생존분석은 어떤 현상이 발생하기 까지에 걸린 시간으로 첫번째 가출의 해 정보를 이용

5 데이터 통합

통합 데이터 한눈에 보기!



전체 학생 중 8%가 가출 경험 有



첫 가출의 나이 분포

6 최종 변수 소개

▶2011-2015년 데이터를 V1(학생 ID)을 기준으로 데이터 통합

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	ID	성별	지역	최종학력_	최종학력_	종사상지우	종사상지우	가구연간소	보호자건강	키	몸무게	건강상태평	성적만족도	가출경험우	가족구성	다문화	형제자매	전학경험우
2	14210	2	10	1	1	1	0	9000	2	162	43	4	2	1	1	2	2	2
3	14323	1	10	1	1	2	3	16000	1	170	67	1	3	1	1	2	2	3
4	14325	1	10	1	1	1	1	2500	3	170	68	1	2	1	1	2	2	3
5	14327	1	10	1	1	2	4	5500	1	175	72	1	2	1	1	2	2	2
6	14413	1	10	2	1	1	0	10000	2	172	67	2	2	1	1	2	2	5
7	14521	2	10	1	1	1	1	5000	2	158	49	3	4	1	4	2	2	0
8	14601	1	10	1	1	3	3	5000	2	177	54	2	2	1	1	2	2	1
9	14609	1	10	1	1	1	1	7000	2	171	58	2	3	1	1	2	2	0
10	14721	2	10	1	1	1	1	10000	2	167	62	1	4	1	1	2	2	0
11	14722	2	10	1	1	2	1	5000	2	151	49	2	4	1	1	2	2	2
12	14808	2	10	1	1	1	1	1000	2	164.5	43	3	3	1	1	2	2	0
13	14817	2	10	1	1	1	3	1500	2	157	60	2	2	1	1	2	2	5
14	14942	1	10	1	1	1	0	4000	2	169	52	2	3	1	1	2	2	1
15	23801	1	20	1	1	0	0	720	1	169	60	1	2	1	2	2	2	3
16	23806	1	20	1	1	1	1	4000	1	169	54	1	3	1	1	2	2	4
17	23810	1	20	1	1	0	0	960	2	172	72	3	3	1	5	2	2	1
18	23901	1	20	1	1	1	0	5500	2	171	58	2	2	1	1	2	2	0
19	23917	2	20	1	1	1	0	2000	1	159	60	1	2	1	2	2	2	0

Obs : 2305
Variable : 30

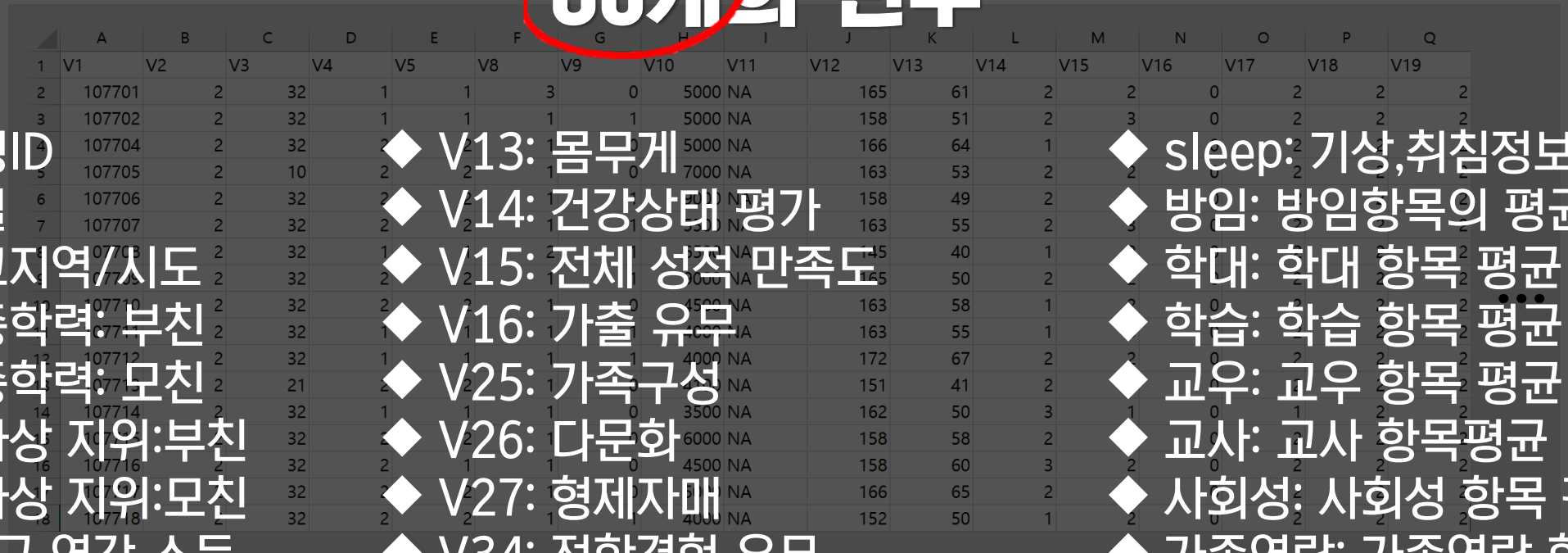
→ tidy data!

6 최종 변수 소개

2305개의 OBS

▶2011-2015년 데이터를 V1(학생ID)을 기준으로 데이터 통합

30개의 변수



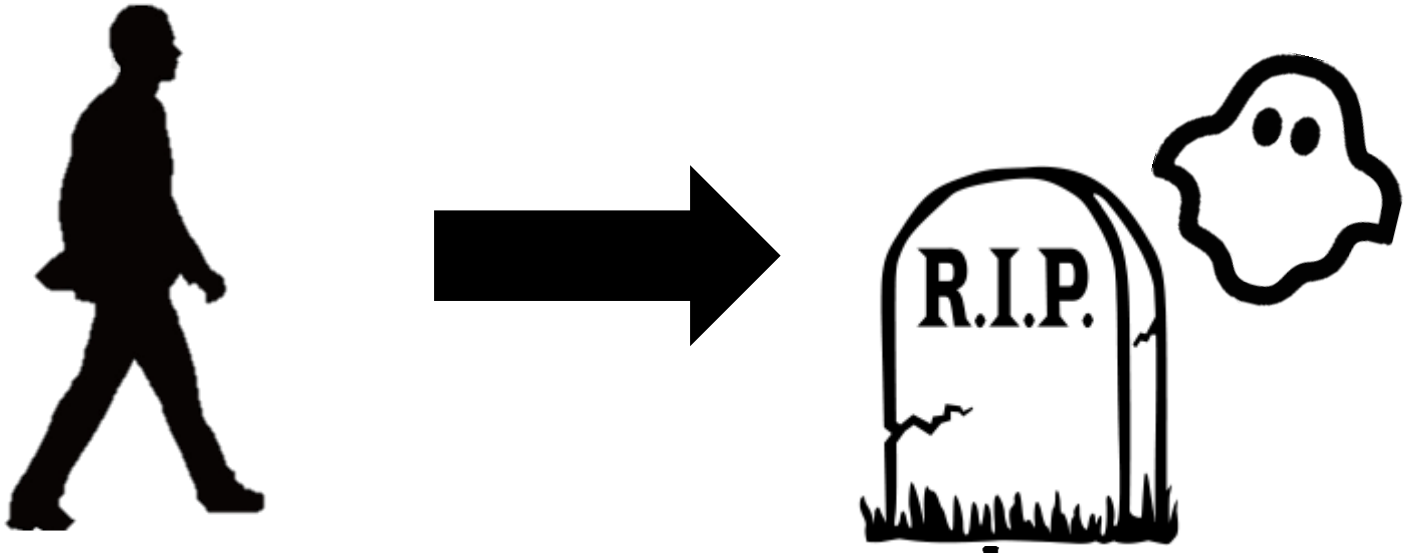
- V1: 학생ID
- V2: 성별
- V3: 학교지역/시도
- V4: 최종학력: 부친
- V5: 최종학력: 모친
- V8: 종사상 지위:부친
- V9: 종사상 지위:모친
- V10: 가구 연간 소득
- V11: 보호자 건강상태 평가
- V12: 키
- V13: 몸무게
- V14: 건강상태 평가
- V15: 전체 성적 만족도
- V16: 가출 유무
- V25: 가족구성
- V26: 다문화
- V27: 형제자매
- V34: 전학경험 유무
- V55: 가족여행횟수
- damage: 비행:연간피해경험 유무
- sleep: 기상,취침정보
- 방임: 방임항목의 평균
- 학대: 학대 항목 평균
- 학습: 학습 항목 평균
- 교우: 교우 항목 평균
- 교사: 교사 항목평균
- 사회성: 사회성 항목 평균
- 가족연락: 가족연락 항목 평균
- 친구연락: 친구연락 항목 평균
- 휴대전화의존도: 의존도 항목 평균

03
생존분석

생존분석(survival analysis)이란?

:사건이 일어나거나 일어나지 않는 결과와 그러한 사건이 일어날 시점을 예측하고 설명하는 통계방법

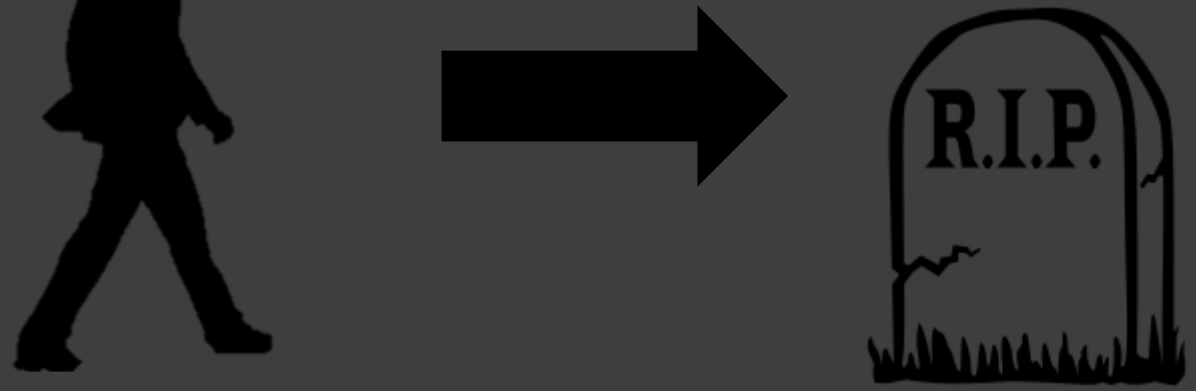
✓생존분석의 이름은 의학계에서 생존과 사망을 다룬 것에서 유래



생존분석(survival analysis)이란?

- 사건이 일어나거나 일어나지 않는 결과와 그러한 사건이 일어날 시점을 예측하고 설명하는 통계방법
- 생존분석의 이름은 의학계에서 사망과 생존을 다룬 것에서 유래

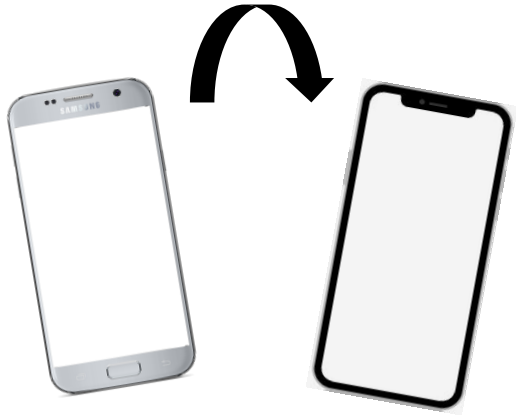
그렇다면 생존분석은 생존과 사망을 다룬 데이터만 가능할까?
(생존과 사망이 뚜렷한 의료 데이터만 쓸 수 있는 건가?)



2 생존 분석 예시

주식시장에서 고객 이탈율

취업결정요인 분석



소비자가 핸드폰을 바꾸는데 걸리는 시간

재활 완료까지의 시간

의료계뿐만 아니라 상당히 **다양한 분야에서 사용되는 기법**이다!

3 회귀분석 vs 생존분석

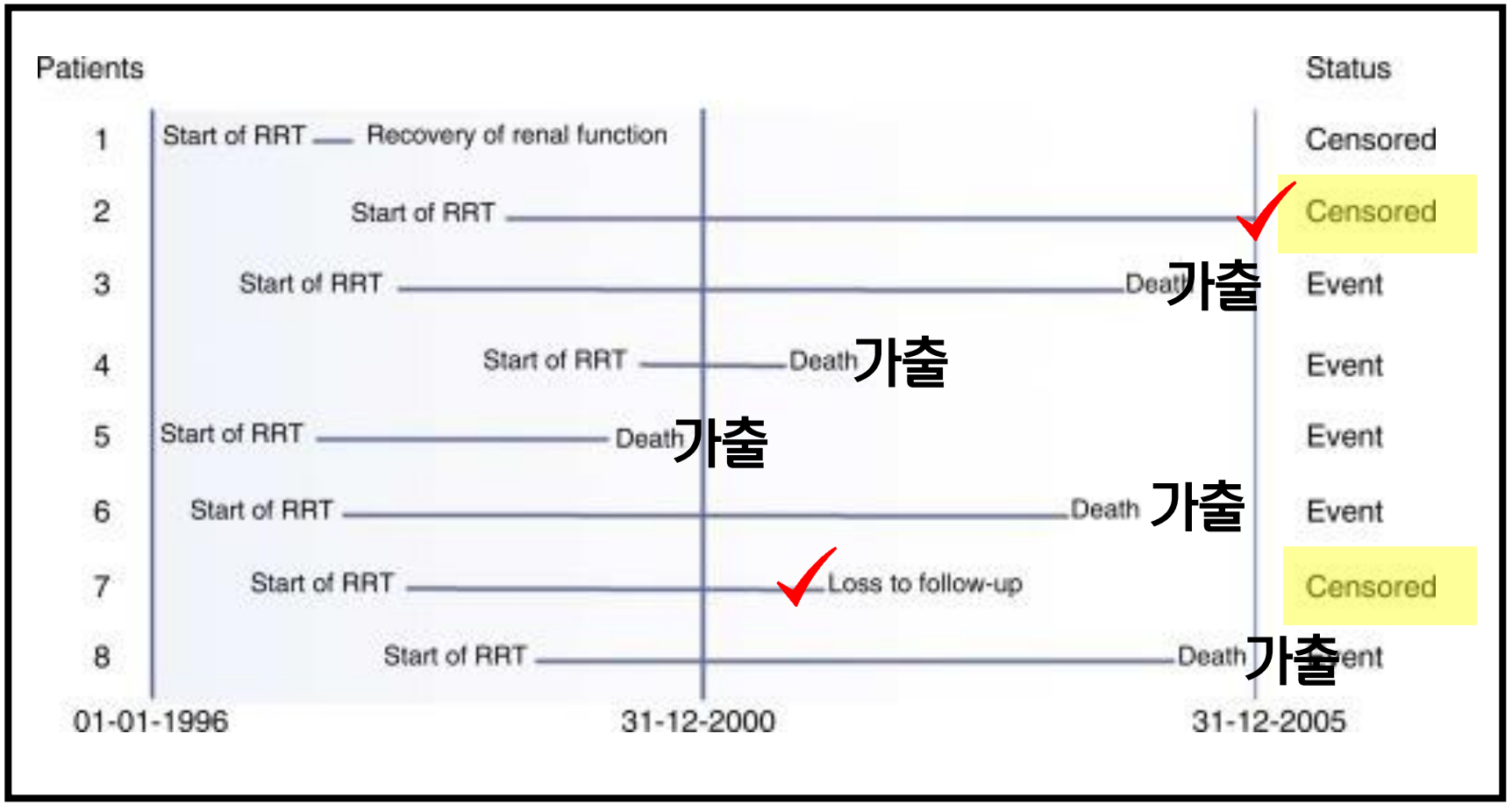
✓회귀분석 반응변수: 사건 발생

✓생존분석 반응변수: 사건 발생+생존시간

같은 사건이 발생하더라도 발생 기간이 다르면 추가적인 요인 분석 가능

BUT 생존분석을 위해선 **Censored Data**가 필요함!

3 생존분석의 censored data?



Censored data(절단된 데이터?)

Type 1 censoring

✓ 추적관찰이 종료됨으로써 관찰이 불완전해지는 경우

➡ *우리의 데이터에서 5년 동안의 조사가 끝난 경우*

Type 2 censoring

✓ 추적 관찰되는 기간 중에 도중탈락(follow-up loss)되는 경우
(random censoring이라고도 한다.)

✓ 왜 도중탈락 될까? 거주지의 변경, 대상자의 거부, 사망 등

➡ *우리의 데이터에서 추적조사가 끊긴 경우*

4 Censored data 예시

2000명의 학생 중, 5년이 지난 시점에서 **1800명이 가출 경험이 없다.**

(생존분석을 진행하기 전, 반응변수에 차별점을 두어야한다.)

가출유무	나이
O	15
O	17
X	19
X	16

→

생존분석
15
17
19+
16+

19+의 경우, type1 censoring;
조사가 진행되는 중 가출이 없었다.

16+의 경우, type2 censoring;
조사가 진행되는 중 도중 탈락되어
가출 유무를 알 수 없다.

▶Kaplan Meier 계산 방법

사건(사망)이 발생한 시점마다 구간생존율을 계산

$$P(t) = \frac{t\text{시점의 생존자수}}{t\text{시점의 관찰대상수}}$$

➡

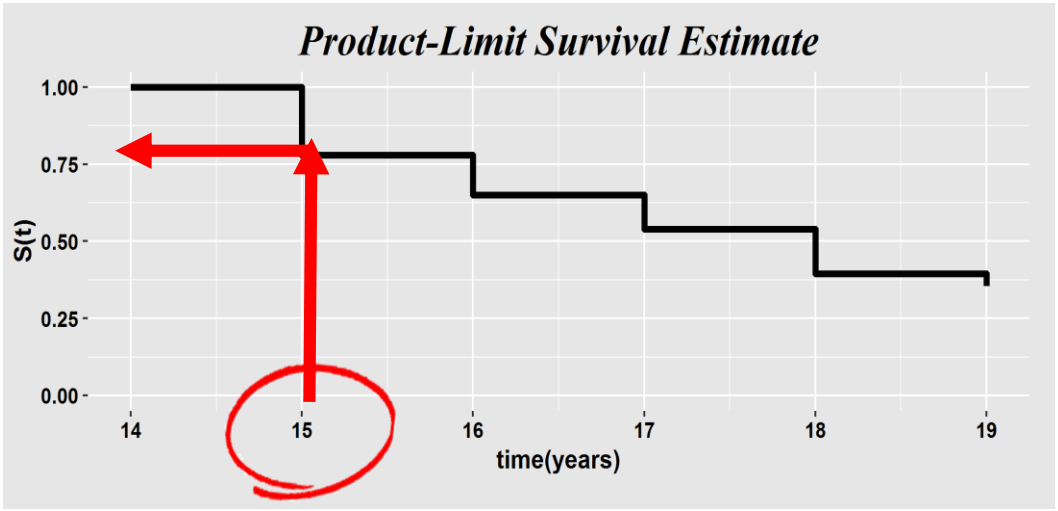
이들의 누적 생존율을 추정

$$S(t) = S(t-1) \times P(t)$$

Kaplan Meier 예시

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

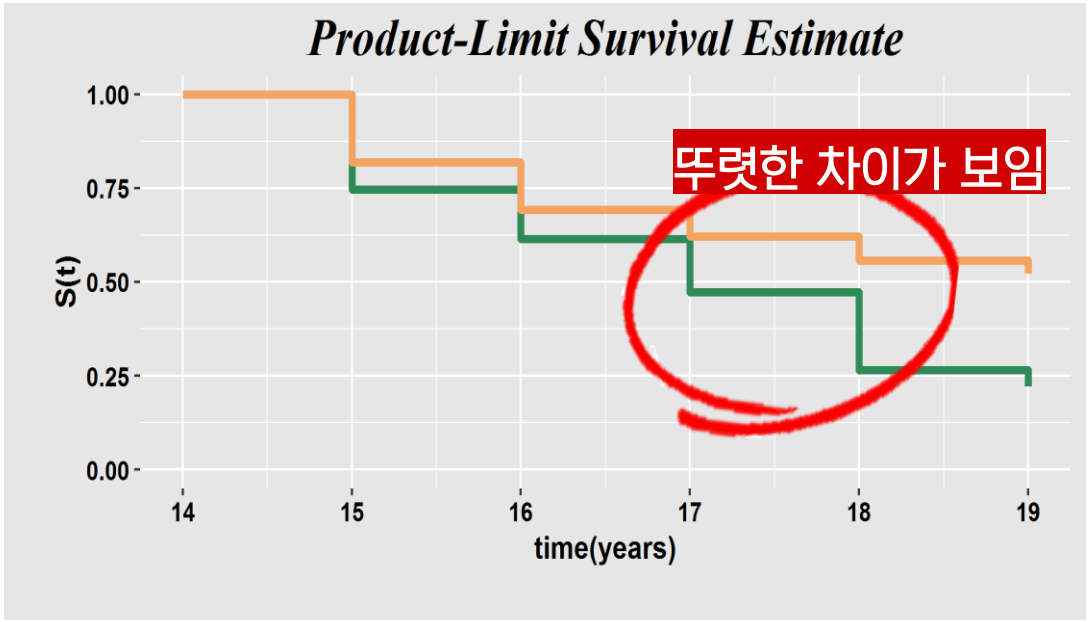
누적생존율을 그래프로 표시



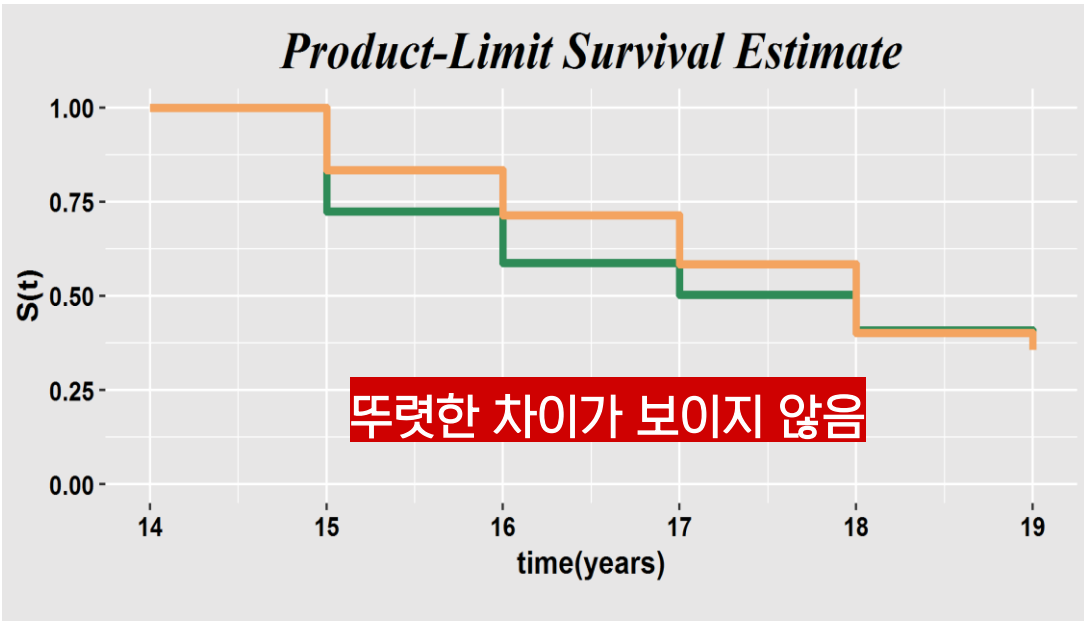
✓ 해석: 15살에 대략 75%의 사람이 생존했다. (즉, 25% 학생이 가출을 했다!)

Kaplan Meier 장점 그룹 간 생존곡선 비교 가능!

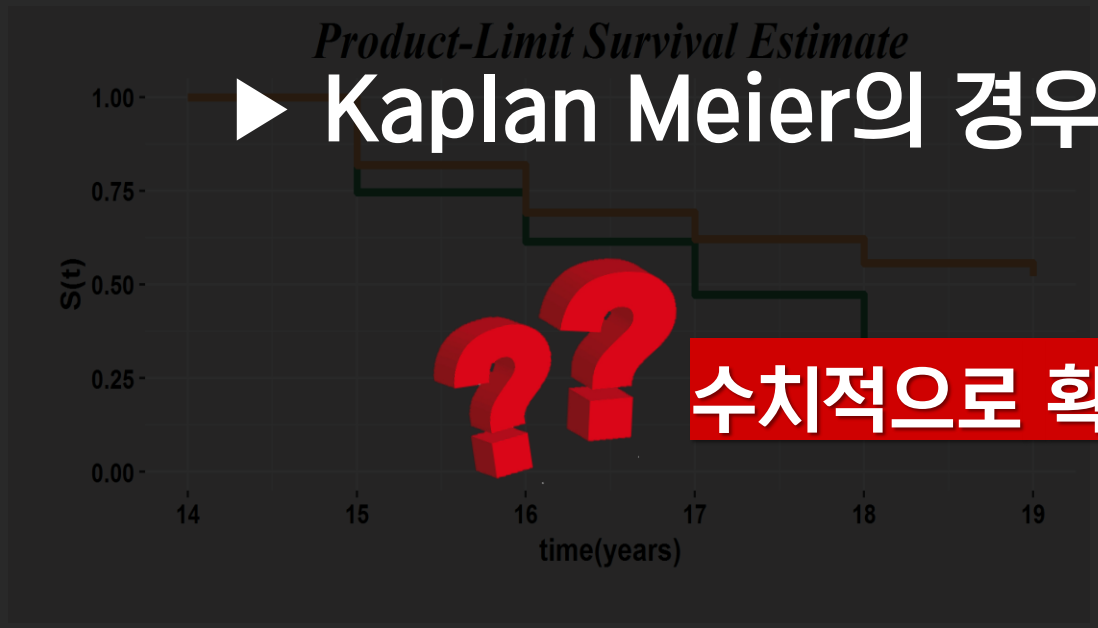
▶ 성별에 따른 가출



▶ 키(height)에 따른 가출



▶ 성별에 따른 가출



▶ 키(height)에 따른 가출

▶ Kaplan Meier의 경우, 수치적으로 확인 불가능하다

수치적으로 확인하는 방법은 없을까?

6 생존분석-Log Rank Test

▶Log Rank Test 방법

:집단 별 생존 함수의 차이가 유의한지를 검증하는 테스트

H_0 : 집단에 따른 생존함수 차이가 없다.

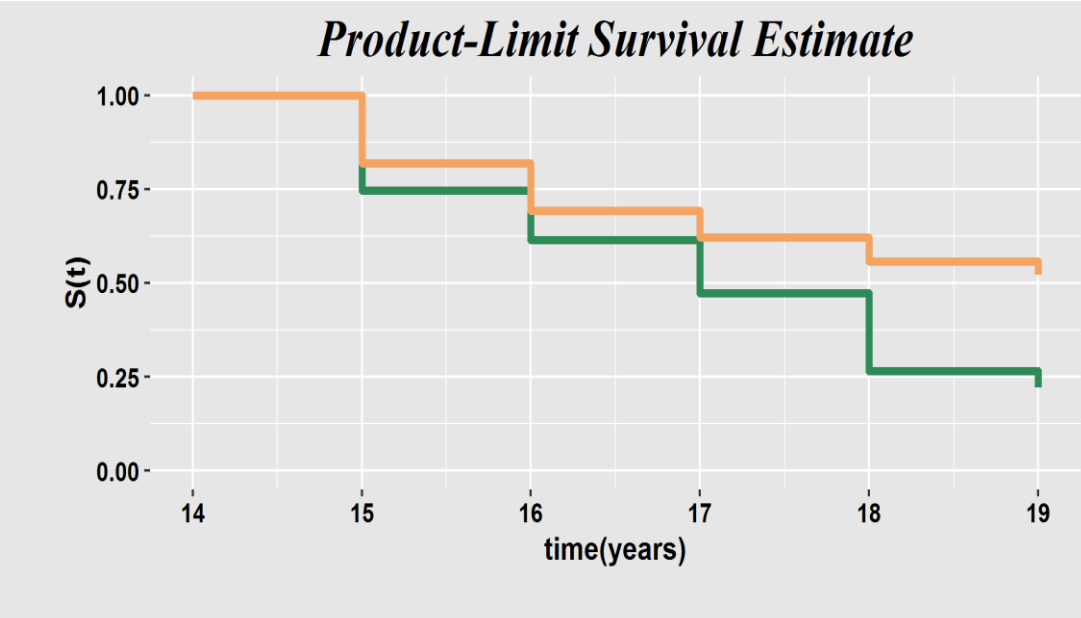
H_1 : 집단에 따른 생존함수 차이가 있다.

*자세한 원리는 부록 참조!

6 생존분석-Log Rank Test

Log Rank Test로 유의한지 판별!

▶ 성별에 따른 가출




```
> survdiff(Surv(runaway$age, runaway$가출경험유무)~ runaway$성별)
Call:
survdiff(formula = Surv(runaway$age, runaway$가출경험유무) ~
    runaway$성별)

                   N Observed Expected (O-E)^2/E (O-E)^2/V
runaway$성별=1 146       111     86.3      7.08      18.4
runaway$성별=2 121        56     80.7      7.57      18.4

 Chisq= 18.4  on 1 degrees of freedom, p= 2e-05
```

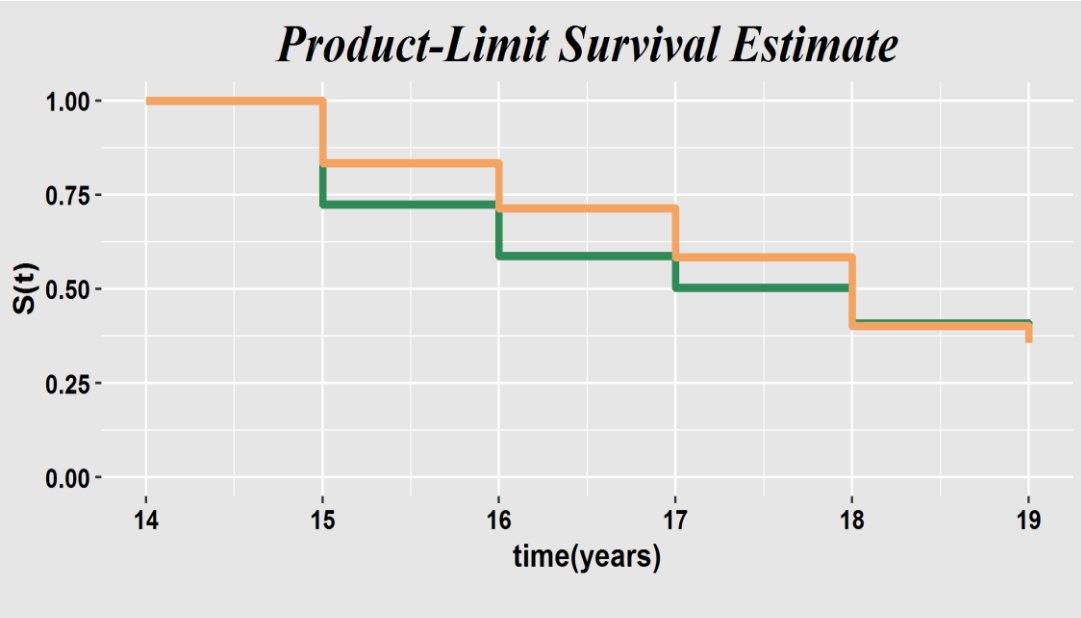
p-value < 0.05 ➡ **귀무가설 기각**

성별 그룹 간 가출 **생존함수가 차이가 있다!**

6 생존분석-Log Rank Test

Log Rank Test로 유의한지 판별!

▶ 키(height)에 따른 가출



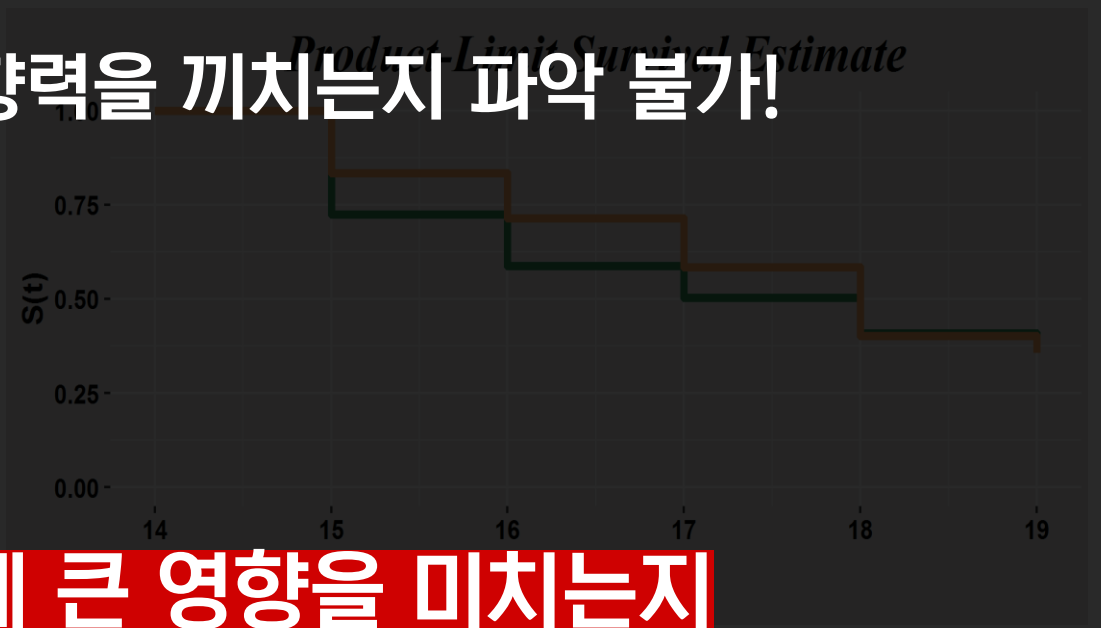

```
> survdiff(Surv(runaway$age, runaway$가출경험유무)~ runaway$키>median(runaway$키))
Call:
survdiff(formula = Surv(runaway$age, runaway$가출경험유무) ~
    runaway$키 > median(runaway$키))

                                        N Observed Expected (O-E)^2/E
runaway$키 > median(runaway$키)=FALSE 134       82     79.1    0.1070
runaway$키 > median(runaway$키)=TRUE  133       85     87.9    0.0963
                                      (O-E)^2/V
runaway$키 > median(runaway$키)=FALSE     0.253
runaway$키 > median(runaway$키)=TRUE      0.253

 Chisq= 0.3  on 1 degrees of freedom, p= 0.6
```

p-value> 0.05 ➡ 귀무가설 기각X

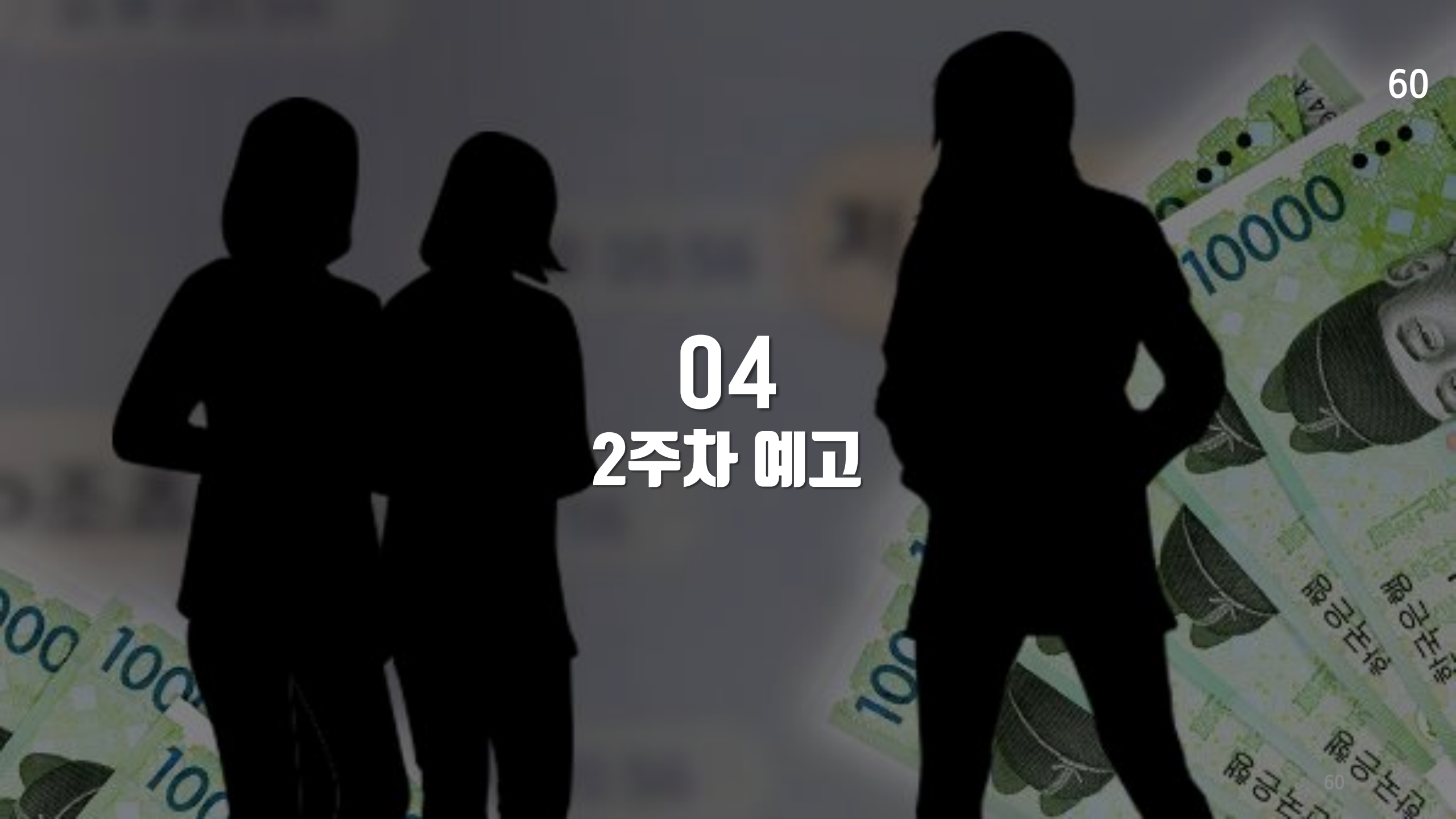
키가 크고 작은 그룹 간 가출 생존함수가 차이가 없다!

▶ **But** 그룹 간 차이를 알려줄 뿐 **모수에 대한 추정 불가**

▶ 어떤 변수가 어느 정도의 영향력을 끼치는지 파악 불가!



어떤 변수가 가출에 큰 영향을 미치는지
어떻게 알 수 있을까?

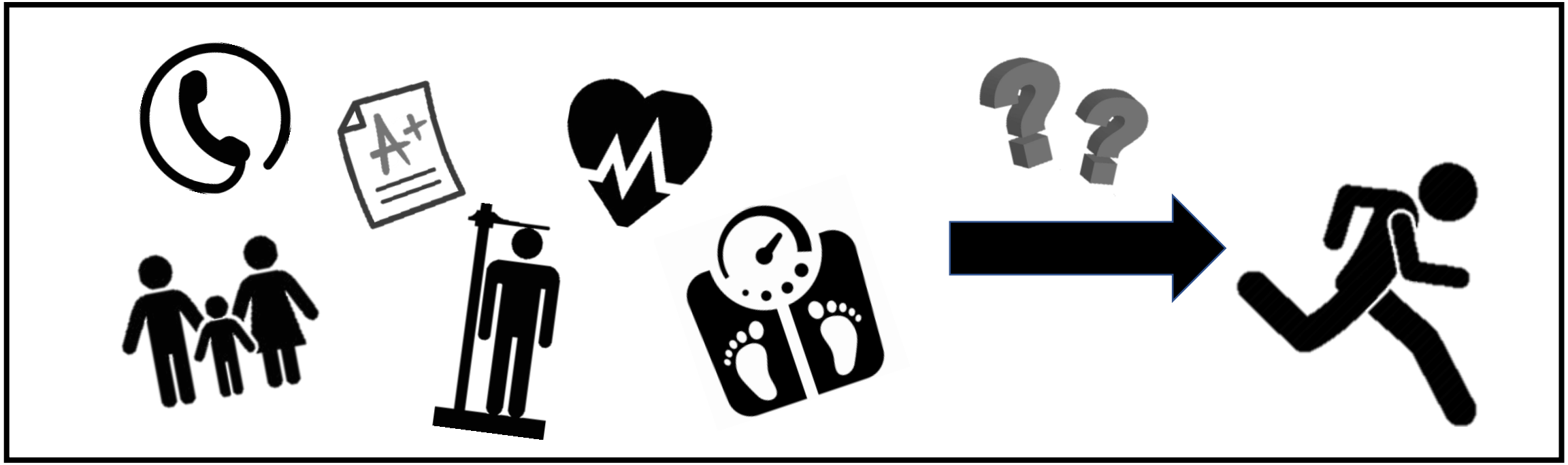


04
2주차 예고

2주차 예고

Cox Propotional Hazard Model(콕스 비례 위험 모형)

:생존과 관련한 다양한 변수의 영향력을 알아보는 분석 방법



#2주차! *이를 바탕으로 청소년 가출의 요인들에 대해 파악해보자!*

THANK YOU!

*부록

Log-rank test

k개의 집단의 생존곡선을 비교

$H_o : S_1(t) = S_2(t) = \cdots = S_k(t)$

$H_1 : not\ H_o$

자유도가 $k-1$인 카이제곱 분포를 따름

**예시

관찰기간 (개월)	관찰대상수		표준치료법 사망수		새 치료법 사망수	
	표준치료법	새 치료법	관찰빈도	기대빈도	관찰빈도	기대빈도
2	13	13	1	0.500	-	0.500
3	12	13	1	0.480	-	0.520
5	11	13	1	0.458	-	0.542
8	10	13	1	0.435	-	0.565
11	9	13	1	0.818	1	1.182
12	8	12	-	0.400	1	0.600
14	7	10	1	0.412	-	0.588
15	6	9	-	0.800	2	1.200
18	5	7	-	0.417	1	0.583
21	5	6	1	0.455	-	0.545
계			7	5.175	5	6.825

$1 \times \frac{13}{13+13} = 0.5$

$1 \times \frac{13}{13+13} = 0.5$

$2 \times \frac{6}{6+9} = 0.8$

$2 \times \frac{9}{6+9} = 1.2$

$$\chi^2 = \sum \frac{(\text{관찰사망수 합} - \text{기대사망수 합})^2}{\text{기대사망수 합}} = \frac{(7-5.175)^2}{5.175} + \frac{(5-6.825)^2}{6.825} = 1.132 < 3.84$$

$\rightarrow p > 0.05$