Part 3

차이를 예측하는 통계 모형

표본(Sample)과 모집단(Population)

모집단의 부분인 표본

- **모집단**
  - 관심있는 관측치를 모두 모아 놓은 것
- **표본**
  - 모집단 중 우리가 데이터로 가지고 있는 일부 관측치
- **예제) 투표와 출구조사**
  - 전체 투표자의 투표 결과를 일부 출구조사 참가 투표자의 데이터로 유추
  - 모집단 : 전체 투표자
  - 표본 : 전체 투표자 중 출구조사에 응답한 투표자

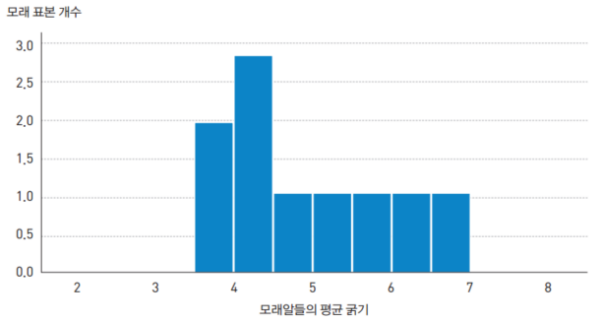
통계량(Statistic)과 분포(Distribution)

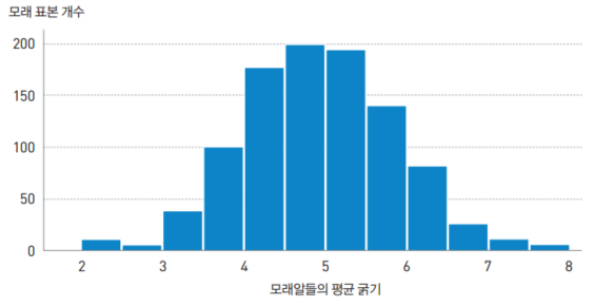
통계량

- 하나의 표본(데이터)에서 계산된 숫자 값

분포

- 여러 개의 표본에서 계산된 여러 개의 통계량이 만들어 낸 숫자 패턴
- 예제) 모래알의 굵기
  - 모래 한 줌의 평균 굵기 $3cm$: 1개 통계량
  - 모래 한 줌 씩 10번 측정한 평균 굵기 : 10개 통계량
  - 모래 한 줌 씩 1,000번 측정한 평균 굵기 : 1,000개 통계량

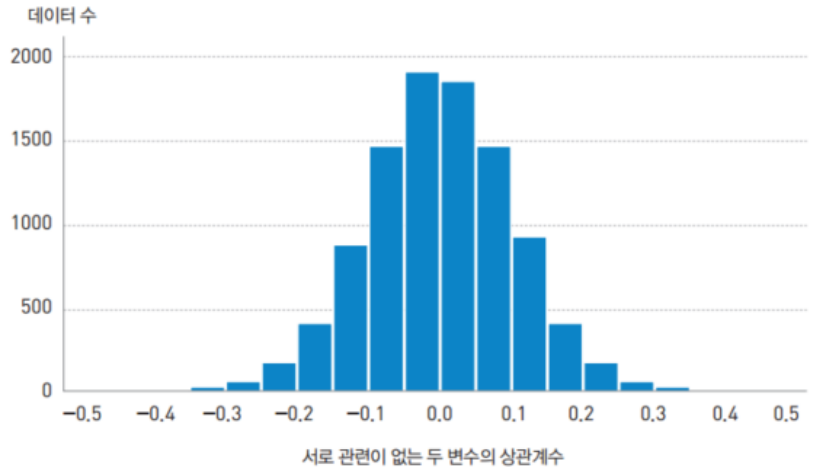




의미 없는 데이터로 만든 분포

자연스러운 확률

- 서로 전혀 관련이 없는 두 연속형 변수의 상관계수
  - 관련이 없다고 해서 항상 상관계수가 0은 아님
  - 우연히 높아 보이는 상관계수가 계산될 가능성도 있음
  - 그러나 대부분 0 근처의 상관계수를 가짐
- 예제) 랜덤으로 1,078개 관측치를 가지게 만든 두 변수의 상관계수



분포를 활용한 판단

의미 없는 분포와 데이터의 비교

- 랜덤 혹은 차이가 없음을 가정하고 만든 의미 없는 데이터로 통계량을 계산하고 분포를 생성
- 우리가 가진 데이터로부터 계산된 통계량을 이 분포에 넣어 상대적인 위치를 측정
- 예제) 아빠 키-아들 키의 상관계수 0.5의 상대적 위치를 측정
  - 앞에서 확인한 서로 관련없는 두 연속형 변수의 상관계수 분포를 활용
  - 서로 관련이 없는 두 연속형 변수에서 0.5라는 상관계수가 나오는 것은 거의 불가능
  - 따라서 아빠 키와 아들 키는 관련이 있다고 결론

p-값의 계산

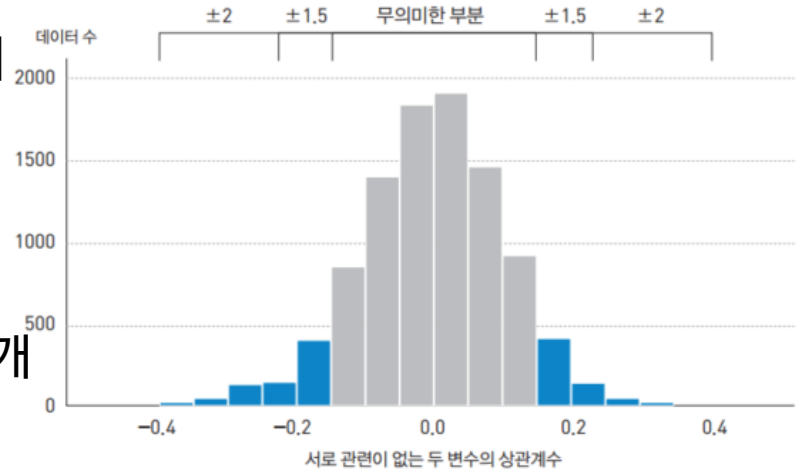
p-값(p-value)

- 데이터에서 계산된 통계량의 기준이 되는 분포 속에서의 상대적 위치(확률)
- 0에 가까울 수록 유의미한 차이라고 판단
- 1에 가까울수록 충분히 우연히 나올 수 있는 흔한 일이라고 판단
- 예제) 1,078개의 관측치를 가진 두 변수의 상관계수
  - 상관계수가 0.3 일 때 :
    - 10,000개의 의미 없는 상관계수 중에서 ± 0.3보다 큰 것은 22개

      → p-값= $\frac{22}{10{,}000} = 0.0022 = 0.22\%$

  - 상관계수가 0.2 일 때 :
    - 10,000개의 의미 없는 상관계수 중에서 ± 0.2보다 큰 것은 484개

      → p-값 = $\frac{484}{10{,}000} = 0.0484 = 4.84\%$



유의수준(Significant level)

유의수준

- 계산된 p-값으로 “통계적으로 유의미한 지”를 판단하는 기준 값
- 일반적으로 5%를 활용
- p-값이 유의수준 5%보다 작으면 통계적으로 유의미한 것으로 판단
- 예제) $1{,}078$개의 관측치를 가진 두 변수의 상관계수가 0.3일 때
  - 계산된 p-값이 0.22%로 유의수준 5%보다 작음
  - 따라서 두 변수의 상관관계는 유의미하다고 판단

통계검정(Statistical Test)

- 주어진 가설에 대해 그 유의성을 데이터로부터 계산된 p-값과 유의수준으로 판단하는 과정

[대표적인 통계 검정]

t-값과 t-검정(t-Test)

t-검정

- 평균에 대한 검정. 일반적으로 평균이 0인지 아닌지를 판단할 때 활용
- 데이터에서 t-값을 계산하고 그 상대적인 위치인 p-값으로 판단
- 예제)
  - 대한민국 성인 남성의 평균 키는 $170cm$보다 유의미하게 클까?

t-값

$$t-\text{값} = \frac{(\text{표본평균} - \text{기준값})}{\text{표본평균의 표준편차}}$$

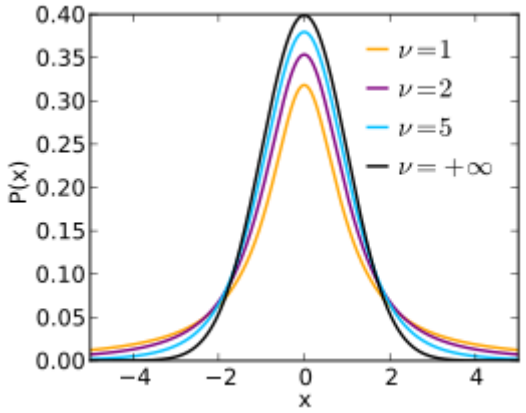
- 데이터에서 계산된 표본평균과 표준편차, 그리고 기준 값으로 계산
- 예제) 성인 남성 16명으로 부터 계산된 키의 평균이 $174cm$, 표준편차가 $2cm$일 때
  - 성인 남성의 평균 키가 $170cm$보다 큰 지를 판단하기 위한 t-값은

$$\frac{(\text{표본평균} - \text{기준값})}{\text{표본평균의 표준편차}} = \frac{(174cm - 170cm)}{2cm} = 2$$

t-분포

t-분포의 필요성

- 계산된 t-값의 상대적 위치인 p-값을 계산해야 통계적 유의성을 확인가능
- 데이터의 관측치 수에 따라 달라지는 t-분포에 계산된 t-값을 넣어 p-값을 확인
  - t-분포는 자유도 $\nu = n - 1$에 따라 형태가 조금씩 달라짐
- 예제) 성인 남성 16명으로 부터 계산된 키의 평균이 $174cm$, 표준편차가 $2cm$일 때
  - t-값 $= 2$ (앞에서 계산)
  - p-값 $= 0.03197 = 3.2\%$
    (t-값을 자유도가 15인 t-분포에 넣어 상대적 위치를 계산)
  - p-값이 일반적인 유의수준 5%보다 작으므로
    성인 남성의 키는 $170cm$보다 유의미하게 크다고 할 수 있음



교차표와 독립

교차표

- 두 범주형 변수의 관계를 수준조합에 따른 관측치 수로 표현한 표
- 행 백분율, 열 백분율을 계산하면 서로 다른 변수의 수준 간의 관계를 확인 가능
- 독립
  - 두 변수의 수준 간에 의미 있는 관계가 없을 때

상품	20대	30대	합계
A			30
B			40
C			30
합계	50	50	100

[행/열 합계만 아는 교차표]

상품	20대	30대	합계
A	15	15	30
B	20	20	40
C	15	15	30
합계	50	50	100

[독립을 가정한 교차표]

상품	20대	30대	합계
A	30	0	30
B	20	20	40
C	0	30	30
합계	50	50	100

[차이가 있는 교차표]

카이제곱분포를 활용한 독립성 검정

독립성 검정

- 두 범주형 변수의 관계를 요약한 교차표에 대한 검정
- 두 변수의 수준들 간에 유의미한 관계가 있는지 혹은 두 변수가 독립인지를 확인

카이제곱값(χ^2-값)의 계산 과정

- 실제 교차표와 독립을 가정한 교차표의 차이 계산(Part1 참고)
- 각 차이 값을 제곱하고 독립을 가정한 교차표의 값으로 나눔
- 이후 모든 값을 더해서 카이제곱값을 확인
- 예제) 올림픽 메달 데이터에서의 카이제곱값 계산(χ^2-값 $= 5.78$)

	금메달	은메달	동메달
28회 아테네	9 - 12 = -3	12 - 9 = 3	9 - 9 = 0
29회 베이징	13 - 13 = 0	10 - 10 = 0	9 - 10 = -1
30회 런던	13 - 11 = 2	8 - 8 = 0	7 - 9 = -2
31회 리우	9 - 8 = 1	3 - 6 = -3	9 - 6 = 3

→

	금메달	은메달	동메달
28회 아테네	$(-3)^2 = 9$	$3^2 = 9$	$0^2 = 0$
29회 베이징	$0^2 = 0$	$0^2 = 0$	$(-1)^2 = 1$
30회 런던	$2^2 = 4$	$0^2 = 0$	$(-2)^2 = 4$
31회 리우	$1^2 = 1$	$(-3)^2 = 9$	$3^2 = 9$

→

	금메달	은메달	동메달
28회 아테네	9/12 = **0.75**	9/9 = **1**	0/9 = **0**
29회 베이징	0/13 = **0**	0/10 = **0**	1/10 = **0.1**
30회 런던	4/11 = **0.36**	0/8 = **0**	4/9 = **0.44**
31회 리우	1/8 = **0.13**	9/6 = **1.5**	9/6 = **1.5**

교차표의 카이제곱값 계산 예제

원본 교차표

	X	Y	행합계
A	35	15	50
B	15	35	50
열합계	50	50	100

행/열 비중

	X	Y	행비중
A	?	?	50%
B	?	?	50%
열비중	50%	50%	100%

빈칸 채우기

	X	Y	행비중
A	25%	25%	50%
B	25%	25%	50%
열비중	50%	50%	100%

기댓값 계산

	X	Y	행합계
A	25	25	50
B	25	25	50
열합계	50	50	100

"−"

"="

	X	Y	행합계
A	10	−10	0
B	−10	10	0
열합계	0	0	0

"□²"

	X	Y	행합계
A	100	100	0
B	100	100	0
열합계	0	0	0

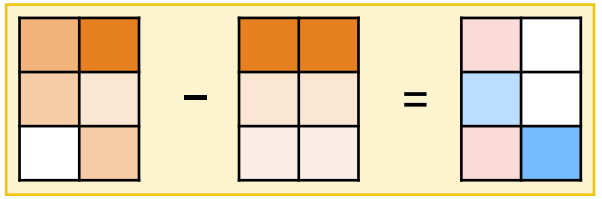
"÷"

	X	Y	행합계
A	4	4	0
B	4	4	0
열합계	0	0	0

"+"

χ^2 −값 : 16

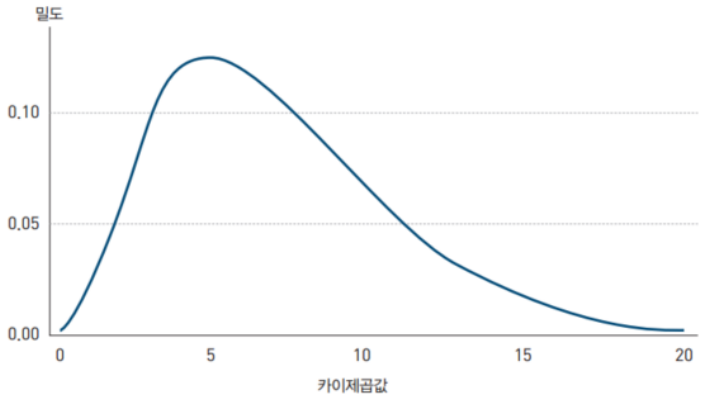
실제값 − 기준값 → 차이



카이제곱 분포의 자유도

카이제곱 분포의 자유도(Degree of freedom)

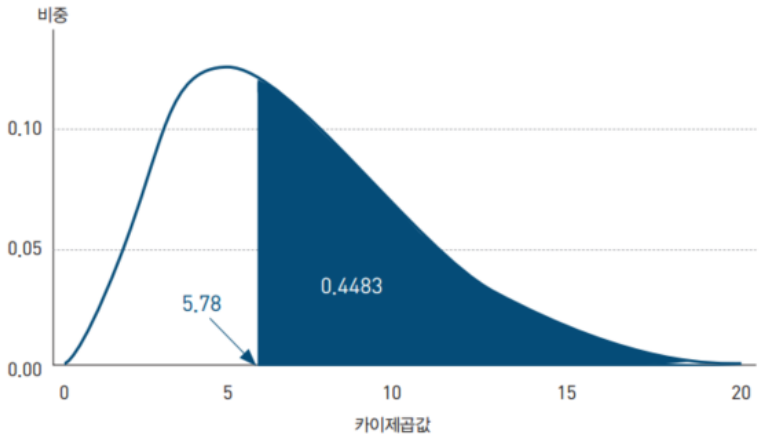
- 일반적으로 카이제곱 분포는 제곱합으로 계산됨
- 몇 개의 값을 제곱하고 더했느냐에 따라서 분포의 형태가 결정
- 자유도 계산
  - (1번째 범주형 변수의 수준의 수 -1) ×(2번째 범주형 변수의 수준의 수 -1)
  - 예제) 올림픽 메달 데이터의 자유도
    - **(대회 수 -1)×(메달 종류 -1)** $= (4-1) \times (3-1) = 6$
  - 자유도가 6인 카이제곱 분포는 왼쪽과 같음

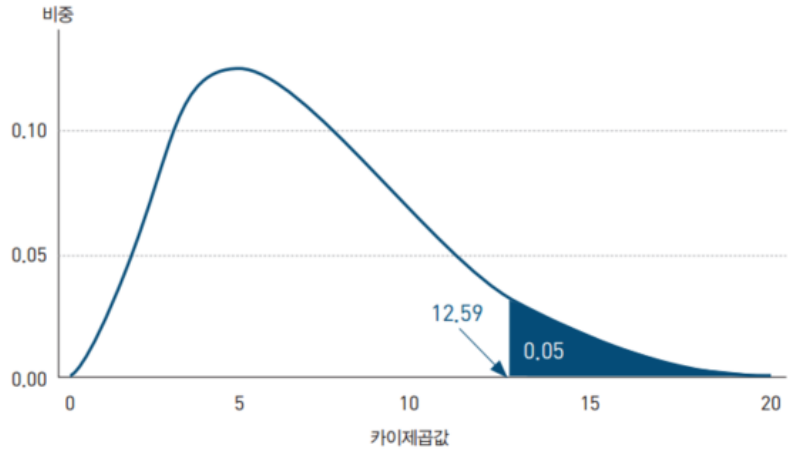


카이제곱 분포에서 p-값의 계산

카이제곱 분포에서 계산된 카이제곱 값의 상대적 위치를 확인

- 계산된 카이제곱 값 5.78을 자유도가 6인 카이제곱 분포에 넣어 p-값을 계산
  - p-값 $= 0.4483 = 44.83\%$
- 계산된 p-값이 유의수준 5%보다 큼
  - 따라서 교차표의 차이는 유의미하다고 볼 수 없음
  - 카이제곱 값이 적어도 12.59는 넘어야 유의하다고 볼 수 있음





F-분포를 활용한 분산분석

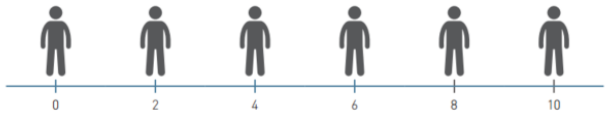
분산분석(ANOVA ; ANalysis Of VAriance)

- 그룹별 평균의 차이에 대한 통계적 검정 과정
- 전체 평균에 비해 그룹별 평균의 차이가 충분히 차이가 나는지를 확인
- 데이터에서 F-값을 계산하고, F-분포를 활용해서 p-값을 계산

분산분석 예제

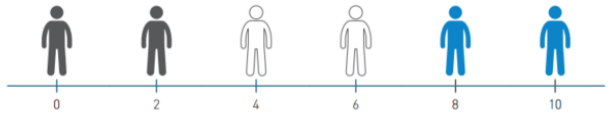
첫번째 단계 - 평균 계산

- 전체 평균의 계산



$$\bar{y} = \frac{0+2+4+6+8+10}{6} = \frac{30}{6} = 5$$

- 그룹별 평균의 계산



$$\overline{yg_1} = \frac{0+2}{2} = 1, \quad \overline{yg_2} = \frac{4+6}{2} = 5, \quad \overline{yg_3} = \frac{8+10}{2} = 9$$

분산분석 예제

두번째 단계 - 세가지 제곱합 계산

- 전체 평균을 활용한 제곱합

- 분산을 계산했던 식과 거의 동일함
- 각 관측치에서 전체 평균을 빼고 제곱해서 합계

$$(0-5)^2+(2-5)^2+(4-5)^2+(6-5)^2+(8-5)^2+(10-5)^2 = 70$$

- 그룹별 평균을 활용한 제곱합

- 전체 평균이 아닌 그룹별 평균을 빼고 제곱해서 합계

$$\{(0-1)^2+(2-1)^2\}+\{(4-5)^2+(6-5)^2\}+\{(8-9)^2+(10-9)^2\}=6$$

- 관측치가 아닌 그룹 평균을 활용한 제곱합

- 관측치 대신 그룹별 평균을 활용하여 그룹별 평균과 전체 평균의 차이로 계산한 제곱합

$$\{(1-5)^2+(1-5)^2\}+\{(5-5)^2+(5-5)^2\}+\{(9-5)^2+(9-5)^2\}=64$$

분산분석 예제

세번째 단계 - F-값의 계산

- 세가지 제곱합의 관계

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(yg_i-\bar{y})^2 + \sum_{i=1}^{n}(y_i-yg_i)^2$$

$$70 \quad = \quad 64 \quad + \quad 6$$

- 분산분석 표 작성

구분	제곱합	자유도	평균제곱합
공부 방법(그룹 간)	64	2	32
개인차(그룹 내)	6	3	2
점수	70	5	14

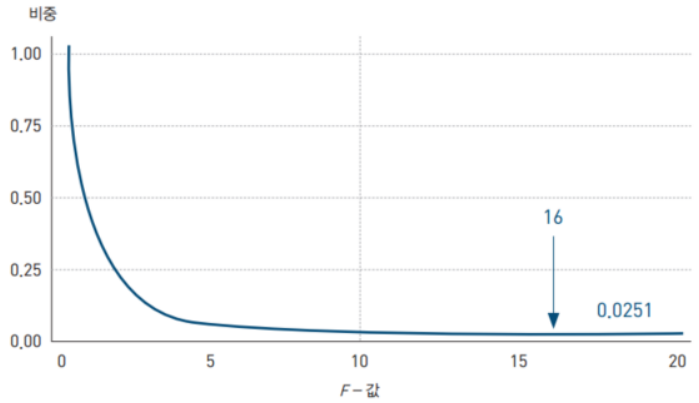
- F-값의 계산

$$F-\text{값} = \frac{\text{설명할 수 있는 부분의 평균제곱합}}{\text{설명할 수 없는 부분의 평균제곱합}} = \frac{32}{2} = 16$$

분산분석 예제

네번째 단계 - F-분포를 활용한 p-값의 계산

- F-분포는 두 개의 자유도에 따라 모양이 결정
  - 위의 분산분석표에 있는 (그룹의 수 -1) 과 (관측치 수 - 그룹의 수)
  - 이 예제에서는 자유도가 2, 3
- 계산된 F-값 16을 자유도가 2, 3인 F-분포에 넣어 p-값을 계산



- p-값은 2.51%
  - 유의수준이 5%라고 하면, p-값이 유의수준보다 작음
  - 따라서 그룹에 따른 평균차이가 유의미하다고 할 수 있음

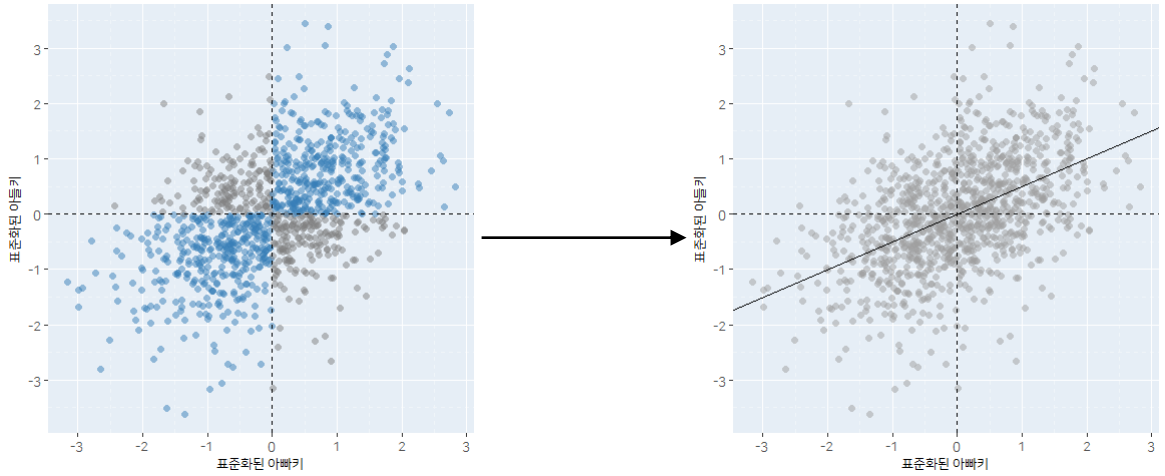
[선형 회귀 모형]

상관계수와 두 변수의 직선관계

상관계수

- 표준화된 두 연속형 변수의 공분산
- 표준화된 두 변수의 산점도에서 두 변수의 관계를 설명하는 기울기에 해당
- 예제) 키 데이터에서 두 변수의 관계
  - 아빠 키와 아들 키의 상관계수 : 0.5

$$\text{표준화된 아들 키} = 0.5 \times \text{표준화된 아빠 키}$$



상관계수를 활용한 단순 회귀 직선

단순 회귀(Simple regression) 직선

- 두 연속형 변수의 관계를 직선의 방정식으로 설명

$$y = a + bx$$

- 상관계수를 활용하여 직선의 방정식을 계산 가능
  - 상관계수로 표준화된 두 변수의 관계를 표현

$$\frac{(\text{아들 키} - \text{아들 키의 평균})}{\text{아들 키의 표준편차}} = 0.5 \times \frac{(\text{아빠 키} - \text{아빠 키의 평균})}{\text{아빠 키의 표준편차}}$$

  - 양변에 아들 키의 표준편차를 곱하고, 아들 키의 평균을 더함

$$\text{아들 키} = 0.5 \times \frac{\text{아들 키의 표준편차}}{\text{아빠 키의 표준편차}} \times (\text{아빠 키} - \text{아빠 키의 평균}) + \text{아들 키의 평균}$$

  - 실제 값으로 계산

$$\text{아들 키} = 0.5 \times \frac{7.15\text{cm}}{6.97\text{cm}} \times (\text{아빠 키} - 171.93) + 174.46$$

$$\Leftrightarrow \text{아들 키} = 0.514 \times \text{아빠 키} + 86.07\text{cm}$$

회귀모형을 활용한 예측

회귀 직선을 활용한 예측

- 회귀 모형을 활용해서 아들 키에 대한 예측 값 $\hat{y}$ 계산 가능

$$\hat{y} = 86.07 + 0.514x$$

- 예제) 아빠 키가 $170cm$인 가족의 아들 키 예측 값은 $173.45cm$

$$86.07cm + 0.514 \times 170cm = 173.45cm$$

회귀모형의 성능평가 예제

제곱합을 활용한 회귀모형의 성능 평가

- 관측치와 전체 평균의 차이에 대한 제곱합

$$\sum_{i=1}^{1,078}(y_i - \bar{y})^2 = 55{,}049$$

- 예측 값과 전체 평균의 차이에 대한 제곱합
  - 예측값이 전체 평균보다 얼마나 더 관측치를 잘 설명하는 지를 의미함

$$\sum_{i=1}^{1,078}(\hat{y_i} - \bar{y})^2 = 13{,}836$$

- 관측치와 예측 값의 차이에 대한 제곱합
  - 예측을 했음에도 설명할 수 없는 개인차를 의미함

$$\sum_{i=1}^{1,078}(y_i - \hat{y_i})^2 = 41{,}213$$

분산의 분해와 결정계수

세 제곱합의 관계

- 어떤 회귀 모형이든 세 제곱합은 아래와 같은 관계가 성립

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y_i}-\bar{y})^2+\sum_{i=1}^{n}(\hat{y_i}-\bar{y})^2$$

- 해석
  - 관측치 간의 차이(첫번째 제곱합)는
  - 모형으로 설명되는 부분(두번째 제곱합)과 설명할 수 없는 개인차(세번째 제곱합)로 분해
- 예제) 키 데이터 $55{,}049 = 13{,}836 + 41{,}213$

결정계수

- 전체 제곱합 중 모형으로 설명되는 부분의 비중

$$R^2=\sum_{i=1}^{n}(\hat{y_i}-\bar{y})^2 \,/\, \sum_{i=1}^{n}(y_i-\bar{y})^2$$

- 예제) 키 데이터의 결정계수 : $R^2=\frac{13{,}836}{55{,}049}=0.25$
  - 아빠 키는 아들 키 차이의 25% 설명할 수 있다.