VIIII. 선형회귀분석과 확장

1. <u>선형회귀분석의 배경 및 개요</u>
2. 주요 개념: 오차항 가정, 최소제곱법, 결정계수, 회귀식, 결과해석 등
3. 선형회귀분석의 예측에의 활용과 오차의 측정
4. 선형 모형의 확장: 포아송/로지스틱 회귀 모형 개념 소개
5. 우도와 최대 우도 추정 방법의 이해

1. 선형회귀분석의 배경 및 개요

- **데이터분석의 목적 → 예측 or 추론**

지도학습 (Supervised Learning)

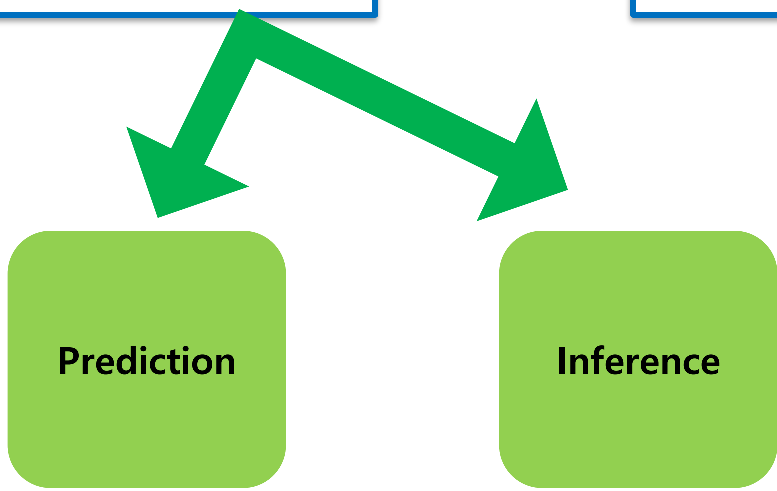
독립(설명)변수를 바탕으로 종속(반응)변수
예측 모형 제시

예:회귀/분류 모형

비지도학습 (Unsupervised Learning)

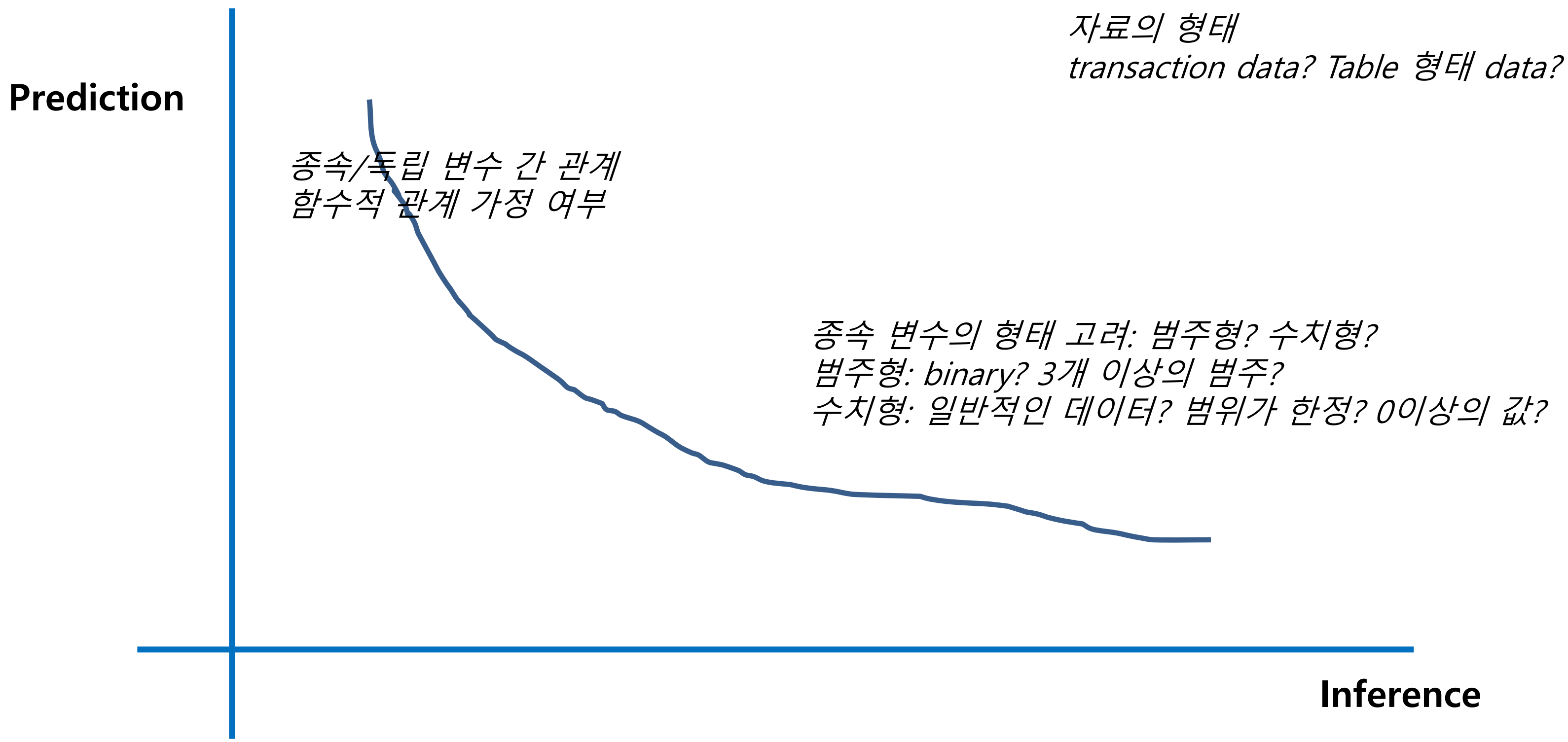
Target(종속변수/반응변수)가 없으며, 독립(설명)변수 간의 관계나 이를 바탕으로 개체들을 구분

예:군집 분석, 연관성 분석, 주성분 / 요인분석



1. 선형회귀분석의 배경 및 개요

➢ **데이터 분석의 목적**



1. 선형회귀분석의 배경 및 개요

- **Regression?**
  - 영국의 우생학자 Francis Galton (1822-1911)

- 아버지와 아들의 키의 관계를 연구하며 Regression이라는 용어를 처음 사용
- 아버지의 키가 큰 경우, 아들의 키는 작거나 아버지의 키가 작은 경우 아들의 키는 크며, 이들의 신장은 평균으로 가려는 경향
- 부모의 키가 아들의 키에 영향을 주지만, 아들의 키는 그 세대 전체의 평균 신장으로 회귀

- **선형 회귀분석 (Linear Regression)**
- **목적 :**
  - 반응 인자(Response variable)와 하나 이상의 예측 인자(Predictor Variables) 사이의 관계를 표본으로부터 추정하여 수학적 모형을 만들고, 이를 통해 반응 인자에 대한 예측을 하는 방법

- **선형회귀는 데이터에 Straight line(기울기와 Y 절편)을 적합(fit)시키는 과정**
  - X변수들이 독립변수/Predictor 변수, Y변수가 종속변수/ Response 변수
  - 선형회귀를 통해 얻어진 Line은 기울기와 Y절편으로 나타내며, 알려진 X값에 대한 Y 값 예측

1. 선형회귀분석의 배경 및 개요
2. <u>주요 개념: 오차항 가정, 최소제곱법, 결정계수, 회귀식, 결과해석 등</u>
3. 선형회귀분석의 예측에의 활용과 오차의 측정
4. 선형 모형의 확장: 포아송/로지스틱 회귀 모형 개념 소개
5. 우도와 최대 우도 추정 방법의 이해

2. 주요 개념: 오차항 가정, 최소제곱법, 결정계수, 회귀식 등

- 종류(인자 수에 의한 분류)
  - *단순회귀분석(Simple Linear Regression Analysis*)

    : 반응인자 1개와 예측 인자 1개로 구성 (예) $Y = b_0 + b_1X_1$
  - *다중 회귀분석(Multiple Regression Analysis)*

    : 반응인자 1개와 두 개 이상의 예측 인자로 구성 (예) $Y= b_0 + b_1X_1 + b_2X_2 + \ldots +b_kX_k$

- Residual(잔차): 이렇게 예측된 Y와 실제 Y의 차이
  - 관측치가 Straight line에서 얼마나 떨어져있는지를 나타냄
  - Least squares : 기울기와 Y절편인 b0 와 b1 를 구하는 방법, Residual errors의 Square의 합을 최소화하는 기울기와 절편을 찾음

- 유의사항
  - 통계적 추론을 하기 위해서는, 잔차에 대한 가정이 필요
  - 등분산성, 선형성, 정규성 (선형성은 변수와 잔차의 Scatter plot 이용하여 확인)

2. 주요 개념: 오차항 가정, 최소제곱법, 결정계수, 회귀식 등

➢ **다중 선형회귀 분석**

- 선형회귀분석은 설명변수가 반응변수에 어떤 효과를 주는지를 모형화하는데 사용
- 특히, 어떤 변수가 일정량 변화 시, 다른 변수들도 그 변화양에 각 기울기가 곱해진 만큼 변화하는 것을 가정
- 여러 개의 설명변수에 대해서는 다중선형회귀분석을 함

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

2. 주요 개념: 오차항 가정, 최소제곱법, 결정계수, 회귀식 등

- **선형회귀 검정**
  - 회귀 계수에 대한 검정
  - 회귀계수=0이면 그 회귀계수에 해당하는 독립변수는 종속변수와 관계가 없다고 해석
  - 회귀계수=0을 H0으로 보고, 다음의 통계량을 통해 검정

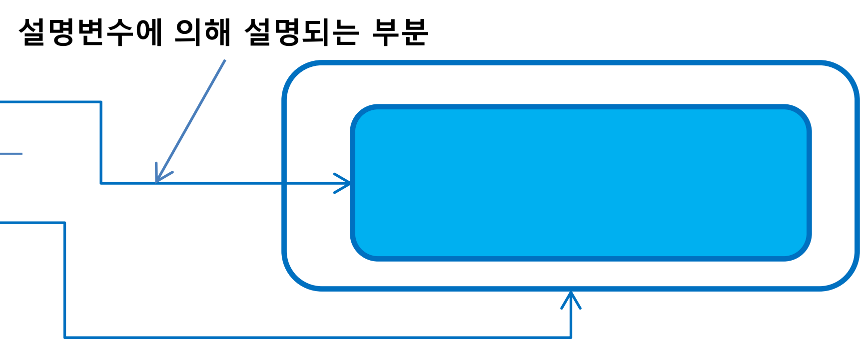
$$t = \frac{\hat{b}}{s.e.(\hat{b})}$$

  - 검정통계량에 대한 확률을 구해서 그 확률이 유의수준보다 작은 경우 H0을 기각
  - 회귀계수는 0이 아니며, 구해진 회귀계수는 유의함

2. 주요 개념: 오차항 가정, 최소제곱법, 결정계수, 회귀식 등

- **결정계수 : R-Squared**
  - 전체제곱합(SST)
    - 실제 반응변수의 값과 예측된 반응변수의 값의 차이의 제곱의 합
    - 회귀제곱합과 잔차제곱합으로 나눠질 수 있음
  - 회귀제곱합(SSR)
    - 예측된 각 반응변수의 값에서 예측된 반응변수의 평균을 뺀 값의 제곱
  - 잔차제곱합(SSE)
    - 관측된 실제 각 반응변수의 값에서 예측된 반응변수의 평균을 뺀 값의 제곱
  - 결정계수
    - 회귀제곱합/전체제곱합, 이 값이 1에 가까울 수록 회귀모형이 데이터를 잘 설명

$$R^2 = \frac{SSR}{SST}$$



VIIII. 선형회귀분석과 확장

1. 선형회귀분석의 배경 및 개요
2. 주요 개념: 오차항 가정, 최소제곱법, 결정계수, 회귀식, 결과해석 등
3. <u>선형회귀분석의 예측에의 활용과 오차의 측정</u>
4. 선형 모형의 확장: 포아송/로지스틱 회귀 모형 개념 소개
5. 우도와 최대 우도 추정 방법의 이해

3. 선형회귀분석의 예측에의 활용과 오차의 측정

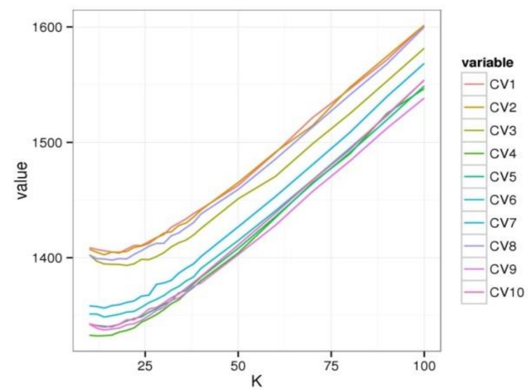
➢ **Sampling (전체 데이터에서 표본을 추출하는 과정)**

1) Simple random sampling : 무작위 추출(복원, 비복원)

2) Stratified random sampling : 층을 결정하고(층: 데이터의 어떤 범주) 각 층에서 무작위 추출

3) Stratified sampling with equal size : 각 층에서 비율을 같게 추출

➢ **Data Partitioning ※ 모형 적합 및 평가를 위해서 필요**



1) Cross validation
   - Training data set: 모형 수립용
   - Validation data set: 수립된 모형의 검증용
   - Test data set: 수립된 모형의 적용

2) 10-fold cross validation
   - 보유 데이터를 10등분하여 9등분은 training, 1등분은 validating으로 쓰는데, 10개에 대해 돌아가며 10번 validation 실시

VIIII. 선형회귀분석과 확장

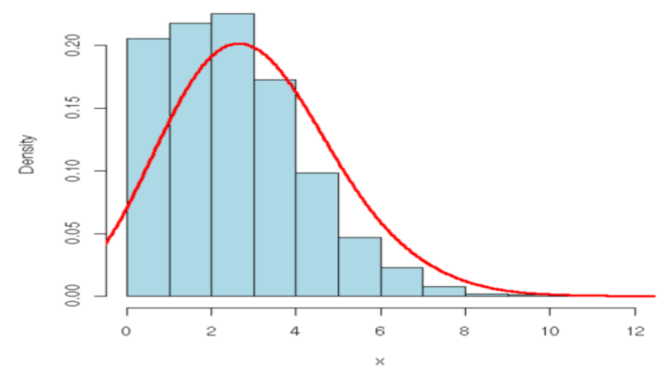
1. 선형회귀분석의 배경 및 개요
2. 주요 개념: 오차항 가정, 최소제곱법, 결정계수, 회귀식, 결과해석 등
3. 선형회귀분석의 예측에의 활용과 오차의 측정
4. <u>선형 모형의 확장: 포아송/로지스틱 회귀 모형 개념 소개</u>
5. 우도와 최대 우도 추정 방법의 이해

4. 선형 모형의 확장: 포아송/로지스틱 회귀 모형 개념 소개

- **Count data**
  - Count data 속성
    - Discrete, Skewed distribution
    - Zero outcome의 비율이 높음
    - 항상 0보다 큼
  - Count 데이터에는 OLS가 어울리지 않음
    - X와 Y의 관계가 비선형
    - Count는 hetero-skedastic하기 때문임(OLS는 등분산성을 가정)
    - OLS로는 종속변수에서 양수 값만 나오도록 예측할 수 없음

    -> OLS(Ordinary Least Square) 외의 다른 추정방법을 사용하는 선형모형이 필요

- **Poisson Regression**
  - Loglinear model이라고도 함
  - Poisson 분포를 가정
  - 특정 지역/개인에게서 특정 사건의 Count에 대한 데이터
  - 빈도이므로 음수는 나오지 않음
  - 데이터가 다음과 같은 분포라면 고려해 볼 수 있음



4. 선형 모형의 확장: 포아송/로지스틱 회귀 모형 개념 소개

- **Poisson Regression**
  - X, Y 관계가 비선형이기 때문에 선형 관계를 만들어서 분석할 수 있도록 처리해야 함
  - Link function: 종속변수를 Transform하기 위해 사용

Poisson: $G(y) = \log(y)$

  - Transform된 Y에 대해 설명변수에 대한 선형 equation을 대입, 다음과 같은 Poisson Regression을 얻음

$$\log(y) = \alpha + \beta x$$

  - Log(y)는 설명변수들에 대해 선형적으로 움직임

- **Poisson Regression의 한계**
  - Over dispersion과 잔차 분포의 heterogeneity
  - 종속 변수의 분산이 평균보다 큰 경우, underestimated standard errors와 overestimated significance of regression parameters의 가능성

4. 선형 모형의 확장: 포아송/로지스틱 회귀 모형 개념 소개

- **Logistic Regression**

  - ✓ Discrete Response Variable의 Modeling
    - Response가 Categorical이거나 Binary이고, 설명변수는 Categorical 또는 Numerical인 경우

  - ✓ Discrete Response Variable의 예

    - **<u>Binary : Logistic Regression</u>**
      - YES / NO
      - 1 또는 0
      - Acceptable 또는 Not acceptable
      - 발생 또는 미발생

    - Discrete Variable with ordering : Ordinal Logistic Regression
      - YES / MAYBE / NO
      - 좋아함 / 보통 / 싫어함

    - Discrete variable without ordering : Multinomial logistic Regression
      - 치킨버거 / 치즈버거 / 불고기버거 / 새우버거

4. 선형 모형의 확장: 포아송/로지스틱 회귀 모형 개념 소개

➢ Logistic Regression

- logit(p)는 다음과 같으며, 이것은 odds의 log와 같음

$$logit(p) = \beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2 + \ldots + \beta_n \times x_n$$

- Odds: P / (1 - P), 어떤 일이 발생할 비율을 발생하지 않을 비율로 나눈 값

$$logit(p) = \log \frac{p}{1-p}$$

- P를 다시 표시하면 아래와 같으면, 모형도 아래와 같이 다시 표현할 수 있음

$$p = \frac{e^{logit(p)}}{1 + e^{logit(p)}} \Rightarrow p = \frac{e^{\beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2 + \ldots + \beta_n \times x_n}}{1 + e^{\beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2 + \ldots + \beta_n \times x_n}}$$

4. 선형 모형의 확장: 포아송/로지스틱 회귀 모형 개념 소개

- **Logistic Regression과 같은 분류모형의 평가**
  - Confusion Matrix

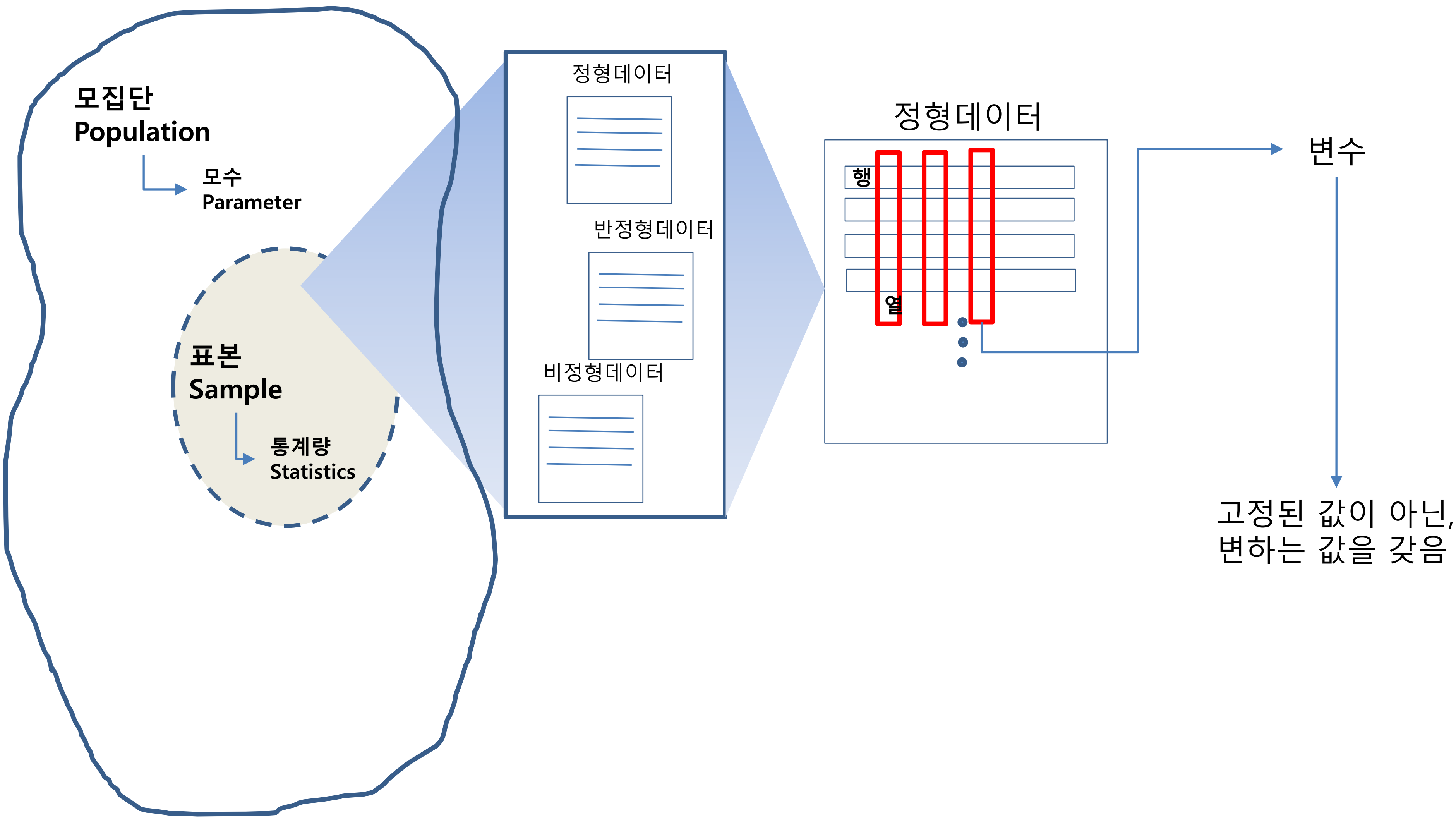
	실제 Y	실제 N
예측 Y	True Positive(TP)	False Positive(FP)
예측 N	False Negative(FN)	True Negative(TN)

  - N=TP+FP+FN+TN
  - 예측 결과에 따라 True, False 구분
  - 예측 값에 따라 Positive, Negative 구분

Metric	Formula	설명
정분류율 or Accuracy	(TP+TN)/N	전체 결과 중 맞게 분류한 비율
오분류율	(FP+FN)/N	전체 결과 중 잘못 분류한 비율
Precision	TP/(TP+FP)	Y로 예측된 것 중 실제로도 Y인 비율
민감도(Recall, Sensitivity, TP Rate, Hit Rate)	TP/(TP+FN)	실제 Y를 Y로 예측한 비율
특이도 (Specificity)	TN/(FP+TN)	실제 N을 N으로 예측한 비율
FP Rate(False Alarm Rate)	FP/(FP+TN)	Y가 아닌데 Y로 예측된 비율이며 (1-특이도)와 동일

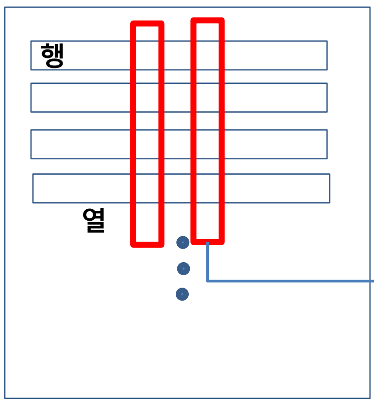
1. <u>데이터부터 회귀분석까지</u>
2. 머신러닝
3. 데이터분석 패러다임의 변화

1. 데이터부터 회귀분석까지



1. 데이터부터 회귀분석까지

정형데이터



변수

어떤 사건에 대한 기록
고정된 값이 아닌,
다양한 값을 갖음

불확실한 값,
불확실한 정도: 확률

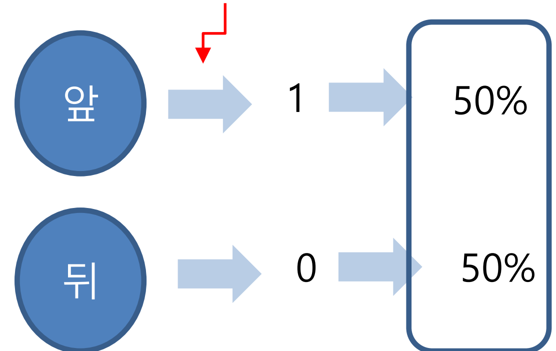
0~1사이 값으로 표현

어떤 사건과
확률을 연결해주기
위해
확률변수와 확률 분
포 필요

예를 들어, 동전을 던지면,

앞과 뒤를 숫자로 표현하면
다루기 편해지는…

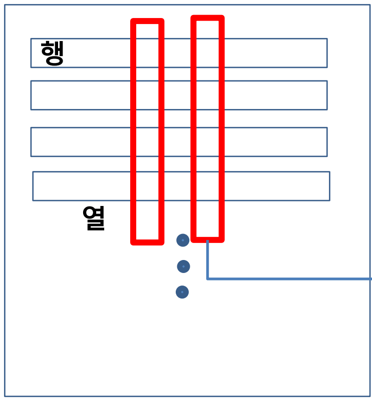
앞면이 나온 횟수를 세어
숫자로 표현하는 대응 규칙을
고려할 수 있음: 확률변수



이때 각 경우에 대한 확률을
표시해줌: 확률 분포

1. 데이터부터 회귀분석까지

정형데이터



변수

어떤 사건에 대한 기록
고정된 값이 아닌,
다양한 값을 갖음

다양한 값이 발생하는
확률은 확률 분포로 나타내며,
그 관계는 확률밀도함수로 표현,
확률분포의 종류도 다양

이산형 확률분포: 어떤 사건이 갖는 값이 셀 수 있는 경우, 이항분포, 포아송 분포

연속형 확률분포: 어떤 사건이 갖는 값이 셀 수 없는 경우, 정규분포, 표준 정규분포

1. 데이터부터 회귀분석까지

다양한 값이 발생하는
확률은 확률 분포로 나타내며,
그 관계는 확률밀도함수로 표현,
확률분포의 종류도 다양

이 자료를 효과적으로
이해하려면: 요약이 필요

표본분포
-모집단에서 일정한 크기로 뽑을 수 있는 표본을 모두 뽑았을 때 그 표본의 특성치(통계량)의 확률분포

중심극한정리
- 표본을 뽑았을 때, n이 충분히 크다면 모집단의 분포모양에 관계없이 표본평균 X 는 근사적으로 정규분포

t – 분포
-서로 다른 두 집단의 평균의 통계 검정

$\chi 2$ - 분포
-서로 다른 2개 이상 집단의 비율의 통계 검정

F - 분포
-서로 다른 2개 이상 집단의 분산의 균질성 검증

통계량
통계량

집중화경향
어디에 값이
주로 몰려있는지…
:mean, median

산포도
평균을 중심으로
값이 얼마나 퍼져
있는지…
:분산, 표준편차

1. 데이터부터 회귀분석까지

통계량
- 한 변수
  - 집중화경향
    :mean, median
  - 산포도
    :분산, 표준편차
- 두 변수
  - 어떤 관계인지
    :공분산
  - 표준화된 관계
    :상관관계
    -1~1

과연 모집단의
특성인 모수를 잘 나타낼까?
확인하는 방법은?...

통계량으로 모수를 추정하는 것을 통계적 추론이라 하며,
가설검정을 통해 할 수 있음

정규성 검정

귀무가설(H_0) : 정규분포를 따른다.

대립가설(H_1) : 정규분포를 따르지 않는다.

t 검정

귀무가설(H_0) : $\mu_1 = \mu_2$ (두 모집단의 평균은 같다.)

대립가설(H_1) : $\mu_1 \neq \mu_2$ (두 모집단의 평균은 다르다.)

Paired t 검정

귀무가설(H_0) : $\delta = 0$ (두 모집단의 평균은 같다.)

대립가설(H_1) : $\delta \neq 0$ (두 모집단의 평균은 다르다.)

F 검정

귀무가설(H_0) : (두 모집단의 산포는 같다.)

대립가설(H_1) : (두 모집단의 산포는 다르다.)

카이제곱 검정

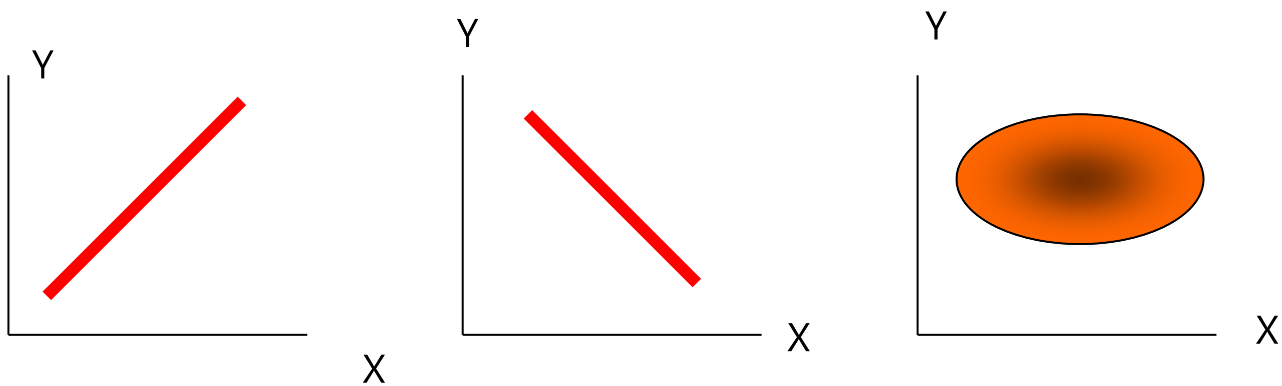
귀무가설(H_0) : 두 모집단은 독립적이다.

대립가설(H_1) : 두 모집단은 종속적이다.

•Step #1	가설 설정
	귀무가설(H_0)과 대립가설(H_1)을 세운다.
•Step #2	유의 수준(α) 결정
•Step #3	P-Value 산출
•Step #4	귀무가설(H_0)의 기각 여부 결정
	If P-Value < 유의수준(α)이면, 귀무가설 (H_0) 기각

1. 데이터부터 회귀분석까지

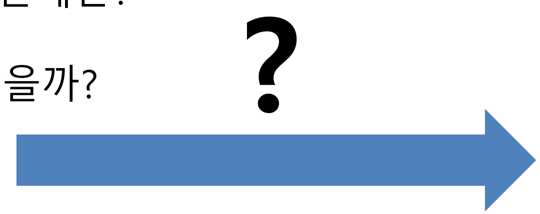
Review:상관분석(Correlation Analysis)



r = 1　　r = -1　　r = 0

Y와 X의 관계가 있음을 알았는데, 과연 둘의 인과관계는?

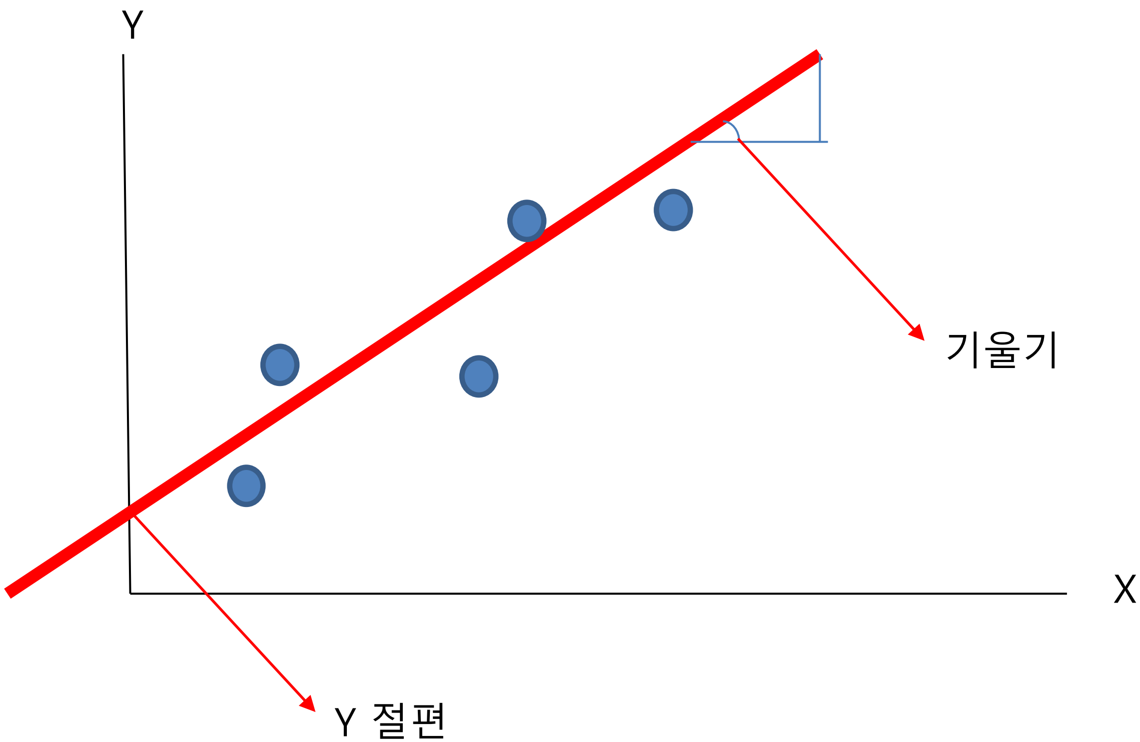
=> X로 인해 Y는 어떤 영향을 받을까?



Y: 종속변수
X: 독립변수

1. 데이터부터 회귀분석까지

주어진 X와 Y의 값을 그려본다면, 아래처럼 표시됨



X와 Y 각각의 값은 좌표가 되어 점으로 표시

여러 개의 점을 한 번에 설명할 수 있는
방법이 필요
-효율적이지만, 아주 정확하지는 않음

점들을 가장 잘 나타내는
하나의 직선으로 표시해 보기!

직선은 기울기와 Y절편만 있으면 그릴 수 있음

더 나아가 X와 Y의 관계도 표시

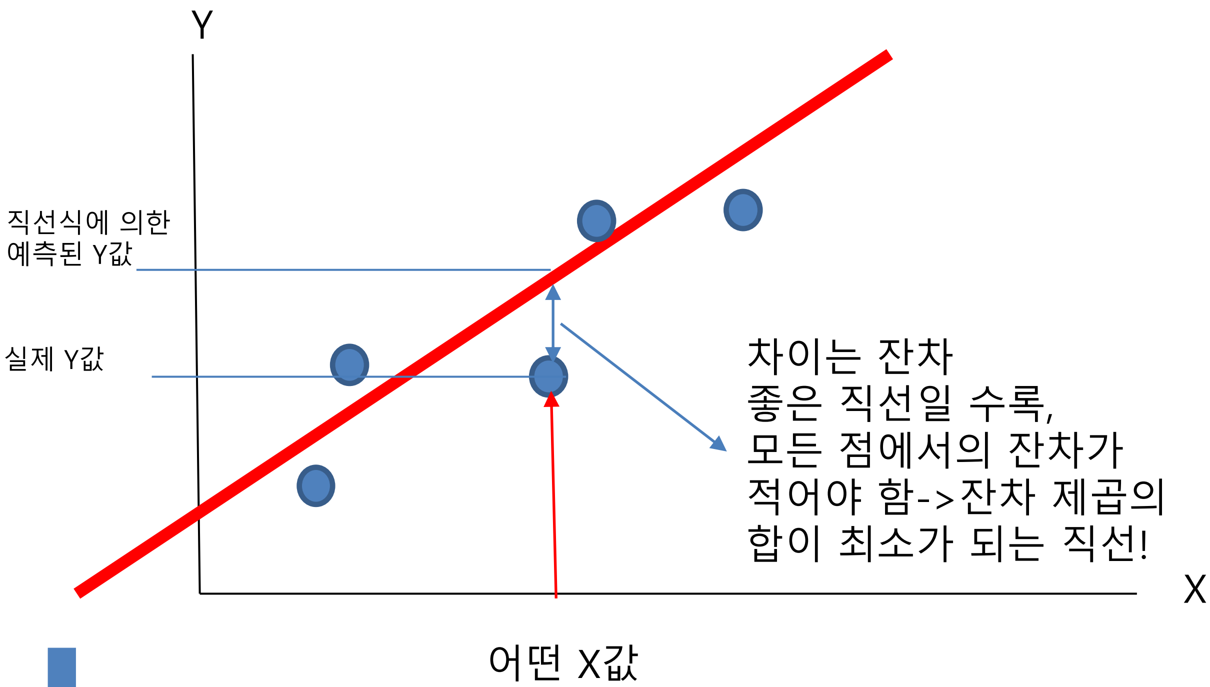
Y = 기울기*X+Y 절편

선형회귀분석

참고로 비선형회귀분석도 있고, 다른 방식의
회귀분석도 많이 있음.

1. 데이터부터 회귀분석까지

선형회귀분석



이 모형이 자료 전체를 잘 설명하는지는
결정계수로 파악
결정계수: 0~1사이 값, 선택한 X변수가 Y 변수의 잔차제곱의 합을 잘 설명해주는 정도를 알려줌

선형회귀분석

기울기의 해석: X가 1단위 증가 시 Y의 변화

모집단에도 우리가 사용한 X, Y에 상응하는 값들이 있고 모집단에서의 기울기가 있음

그렇지만, 우리는 주어진 자료(표본)로만 기울기를 알아내야 함....

기울기를 추정해야 하는 문제가 되며 통계량이 모수를 잘 나타내는지를 알기 위해 가설검정을 함

예:

추정된 기울기가 0.5
H0: 기울기=0
H1: 기울기!=0

이때 이 기울기의 p-value는 0.01이라면, 귀무가설이 기각되어, 추정된 기울기는 통계적으로 유의!