2024.07.20(토)

정규세션 1주차

기초통계(1)

21기 이소영, 22기 정혜정

CONTENTS

목차

CHAPTER 01

확률의 관점

01 확률의 관점

빈도주의와 베이지안

빈도주의(Frequentist)

확률을 **사건의 빈도**로 보는 관점

베이지안(Bayesian)

확률을 **주장에 대한 신뢰도**로 해석하는 관점

확률을 해석하는 관점의 차이

동전의 앞면이 나올 확률이 50%

빈도주의 : 100번 동전을 던졌을 때 50번은 앞면이 나온다
베이지안 : 동전의 앞면이 나왔다는 주장의 신뢰도가 50%이다

TOBIG'S

01 확률의 관점

빈도주의와 베이지안

빈도주의(Frequentist)	베이지안(Bayesian)
확률을 사건의 빈도로 보는 관점	확률을 주장에 대한 신뢰도로 해석하는 관점
모수(parameter)는 고정된 상수	모수(parameter)는 변하는 수
데이터가 분포를 가짐 → 데이터 분포를 가정하여 모수에 대한 추론	데이터는 관찰된 sample (고정) → 데이터를 기반 정보를 업데이트하여 모수를 추정
데이터가 부족할 경우 → 신뢰도가 낮고 잘못된 추론을 할 수 있음	사전지식의 확실성/불확실성 관련 → 사전확률에 따라 결과가 크게 바뀔 수 있음

베이즈 정리

$$\underbrace{P(H|E)}_{\text{사후확률(posterior)}} = \frac{\overbrace{P(E|H)}^{\text{가능도(likelihood)}} \overbrace{P(H)}^{\text{사전확률(prior)}}}{P(E)}$$

[사전확률과 사후확률의 관계를 나타내는 정리]

불확실성 하에서 의사결정을 하는데 있어 중요한 도구

특히 새로운 정보가 주어졌을 때 이전에 믿었던 것을 어떻게 업데이트할지 알려줌

예측, 분류, 의사결정 문제 등 다양한 분야에서 응용

베이즈 정리

$$\underbrace{P(H|E)}_{\text{사후확률(posterior)}} = \frac{\overbrace{P(E|H)}^{\text{가능도(likelihood)}} \overbrace{P(H)}^{\text{사전확률(prior)}}}{P(E)}$$

- H (Hypothesis) : 가설 or ‘어떤 사건이 발생했다는 주장’을 의미
- E (Evidence) : 새로운 정보
- P(H) : 어떤 사건이 발생했다는 주장에 관한 신뢰도 = 사전확률
- P(H|E) : 새로운 정보를 받은 후 갱신된 신뢰도 = 사후확률
- P(E|H) : 추가정보 = 가능도

TOBIG'S

베이즈 정리

사전확률(prior)

$$P(H|E) = \frac{P(E|H)\ P(H)}{P(E)}$$

삼색이 고양이 성별?

성염색체와 관련된 이유로 삼색이 고양이는 거의 대부분이 암컷이라고 알려져 있다.

고양이의 크기, 행동 등의 정보(feature)가 없이도
사전 지식만 가지고 판별O

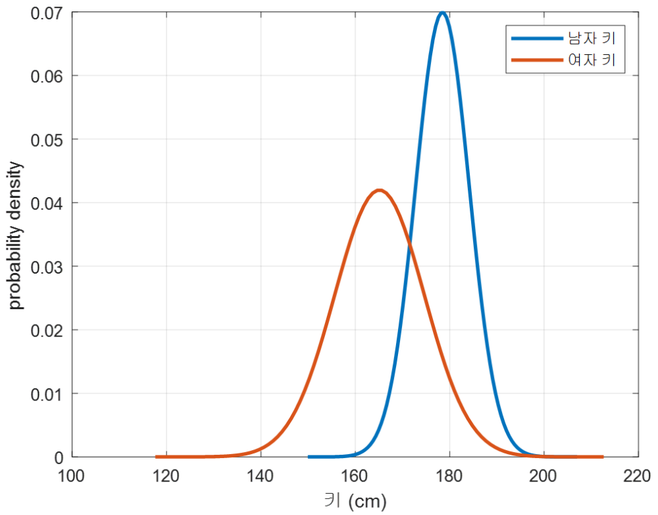
이러한 판별에 도움을 주는 확률값

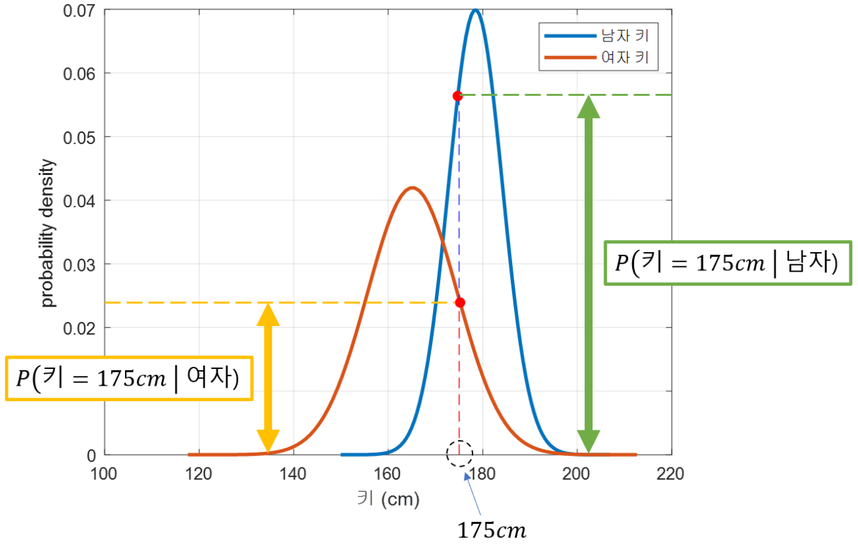
베이즈 정리

$$P(H|E) = \frac{\overbrace{P(E|H)}^{\text{가능도(likelihood)}}\ P(H)}{P(E)}$$

[키(특정 정보)에 따라 이 사람이 여자인지 남자인지 판별하기]

키 175cm





여자라고 생각하는 것 보다 남자라고 생각하는 것이
더 “가능성”이 커 보인다는 것을 알 수 있음

베이즈 정리

예시1)
질병 A의 발병률은 0.1%로 알려져있음.
이 질병이 실제로 있을 때 질병이 있다고 검진할 확률(민감도)은 99%,
질병이 없을 때 없다고 실제로 질병이 없다고 검진할 확률(특이도)는 98%라고 함.
만약 어떤 사람이 질병에 걸렸다고 검진 받았을 때, 이 사람이 정말로 질병에 걸렸을 확률은?

베이즈 정리

예시2)
예제1에서 한 번 양성 판정을 받았던 사람이 두 번째 검진을 받고 또 양성 판정을 받았을 때,
이 사람이 실제로 질병에 걸린 확률은?

CHAPTER 02

확률적 모델링을 위한 확률론

이산확률분포 vs 연속확률분포

이산확률분포

확률변수가 가질 수 있는 값이
셀 수 있는 경우의 분포

베르누이 분포
이항 분포
포아송 분포

연속확률분포

확률변수가 가질 수 있는 값이
연속적인 실수인 경우의 분포

정규 분포
지수 분포
감마 분포

02 확률적 모델링을 위한 확률론

이산확률분포

베르누이 분포 (Bernoulli Distribution)

두 가지 가능한 결과 중 하나가 나오는 실험
실험 결과 → “성공”, “실패”로 구분

예)
동전 던지기 1번 → 앞면 / 뒷면 ?

이항 분포 (Binomial Distribution)

베르누이 시행을 n번 반복했을 때
k번 성공할 확률

예)
동전을 10번 던져서
앞면(성공)이 나오는 횟수?

포아송 분포 (Poisson Distribution)

단위시간 또는 단위공간에서
어떤 사건이 몇 번 발생할 것인가 표현

예)
특정 시간동안 걸려오는 전화의 수 ?

02 확률적 모델링을 위한 확률론

연속확률분포

정규 분포 (Normal distribution)

데이터가 평균을 중심으로
대칭적으로 분포하는 경우

지수 분포 (Exponential Distribution)

포아송 분포
: 일정 시간동안 발생하는 사건의 횟수
→ 지수 분포
: 특정 사건이 발생할 때까지의
대기 시간이나 간격

예)
하루 평균 3명의 환자가 내원한다고 했을 때,
첫번째 환자가 5시간 안에 내원할 확률

감마 분포 (Poisson Distribution)

지수분포의 일반화 된 형태
= 감마분포는 지수분포의 합
예)
낚시를 하는데 어부가 물고기를
30분에 한 마리씩 잡음. 어부가 4마리
물고기를 잡을 때까지 걸리는 시간이
2~4시간일 확률은?
- 물고기 한 마리 잡는데 시간 : 지수함수
- 물고기 4마리 잡는데 시간 : 감마함수

TOBIG'S

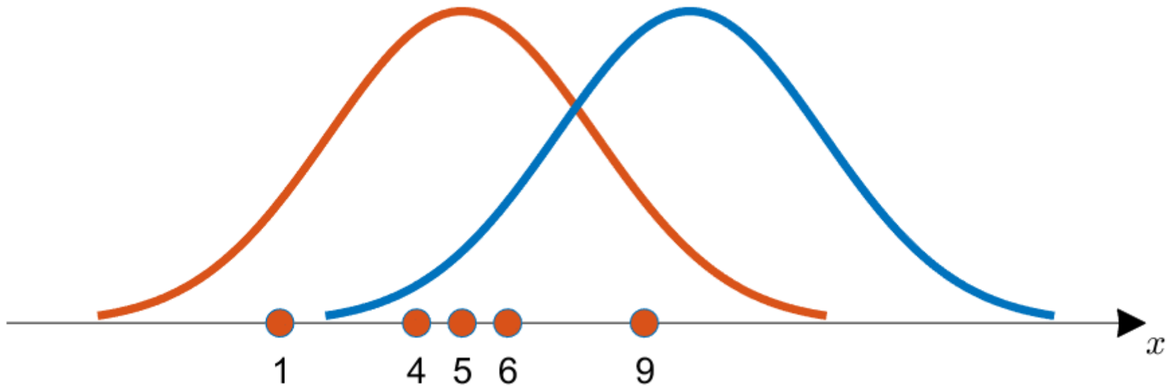
02 확률적 모델링을 위한 확률론

모수 추정 과정

MLE (최대우도법, Maximum Likelihood Estimation)

어떤 확률변수에서 표집한 값들을 토대로
각 가설마다 계산된 우도값 중 가장 큰 값을 고르는 방법

x={1,4,5,6,9} 데이터를 얻었다고 가정



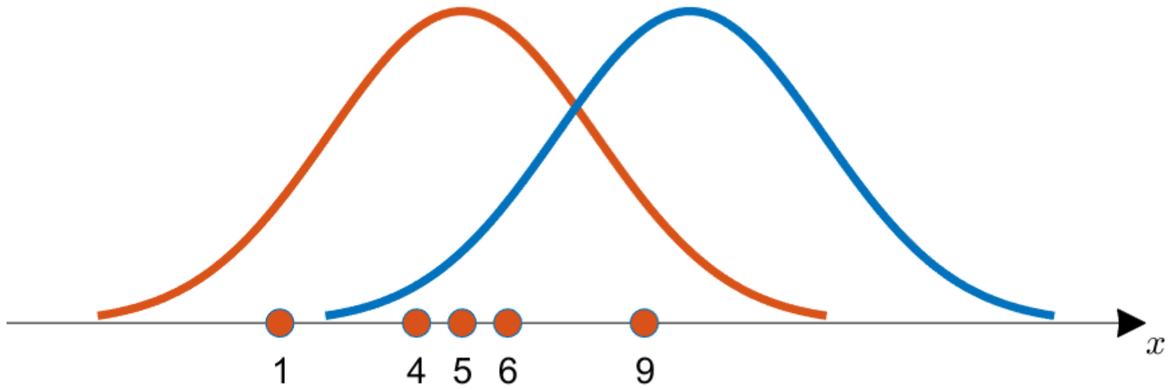
데이터 xx는 주황색 곡선과 파란색 곡선 중
어떤 곡선으로부터 추출되었을 확률이 더 높을까?

모수 추정 과정

MLE (최대우도법, Maximum Likelihood Estimation)

어떤 확률변수에서 표집한 값들을 토대로
각 가설마다 계산된 우도값 중 가장 큰 값을 고르는 방법

x={1,4,5,6,9} 데이터를 얻었다고 가정



데이터 xx는 **주황색 곡선**과 파란색 곡선 중
어떤 곡선으로부터 추출되었을 확률이 더 높을까?

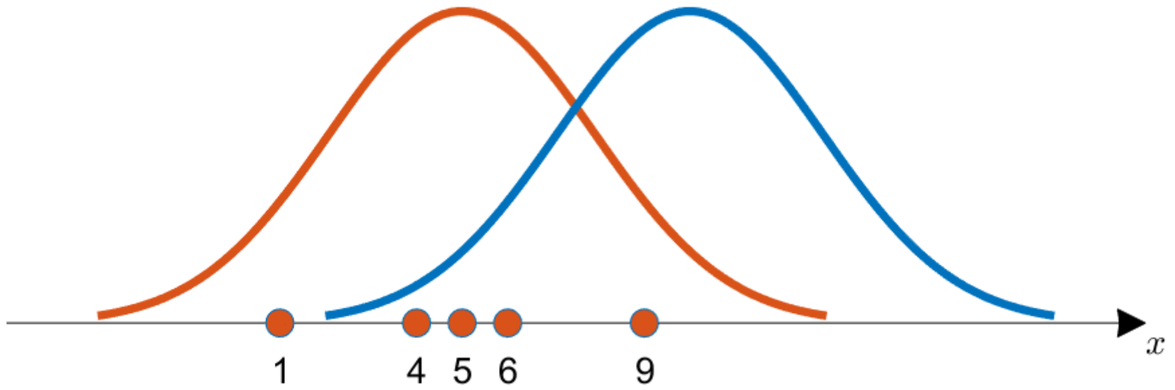
02 확률적 모델링을 위한 확률론

모수 추정 과정

MLE (최대우도법, Maximum Likelihood Estimation)

어떤 확률변수에서 표집한 값들을 토대로
각 가설마다 계산된 우도값 중 가장 큰 값을 고르는 방법

x={1,4,5,6,9} 데이터를 얻었다고 가정

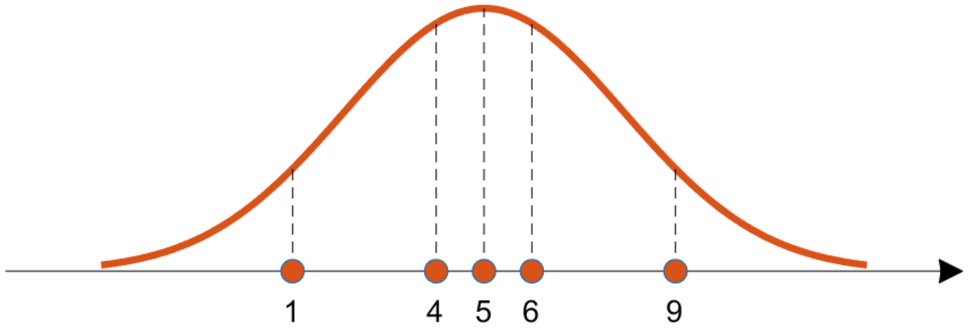


우리가 데이터를 관찰함으로써
이 데이터가 추출되었을 것으로 생각되는
분포의 특성을 추정할 수 있음

모수 추정 과정

MLE (최대우도법, Maximum Likelihood Estimation)

어떤 확률변수에서 표집한 값들을 토대로
각 가설마다 계산된 우도값 중 가장 큰 값을 고르는 방법



likelihood : 지금 얻은 데이터가 이 분포로부터 나왔을 가능도

$$P(x|\theta) = \prod_{k=1}^{n} P(x_k|\theta)$$

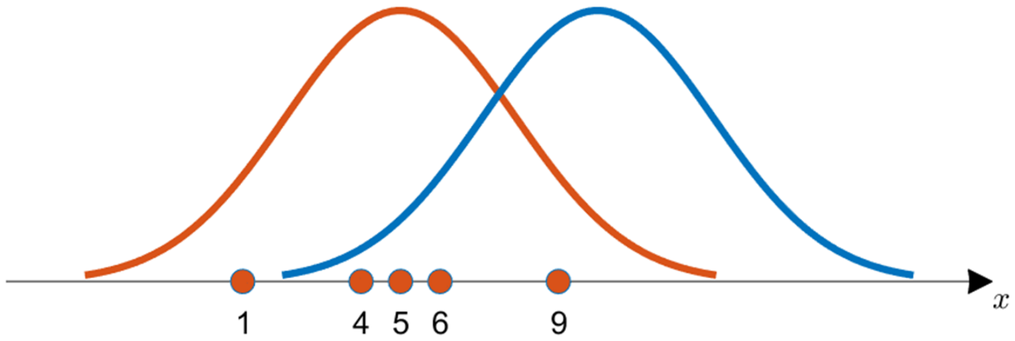
각 포인트로부터 해당 분포에 대한 높이들 계산
→ 데이터들이 갖는 확률밀도함수(pdf)에 대한 높이를 알 수 있음
→높이들을 모두 곱함
→ 높이 하나하나를 likelihood에 기여하는 기여도라고 함.

02 확률적 모델링을 위한 확률론

모수 추정 과정

MLE (최대우도법, Maximum Likelihood Estimation)

어떤 확률변수에서 표집한 값들을 토대로
각 가설마다 계산된 우도값 중 가장 큰 값을 고르는 방법



likelihood : 지금 얻은 데이터가 이 분포로부터 나왔을 가능도

$$P(x|\theta) = \prod_{k=1}^{n} P(x_k|\theta)$$

주황색 likelihood 기여도 계산한 값과
파란색 likelihood 기여도 계산한 값 다름

→ 최대화 할 수 있는 θ 값 찾는 것이 목표

CHAPTER 03

몬테카를로

몬테카를로 기본 개념

몬테카를로 (Monte Carlo)

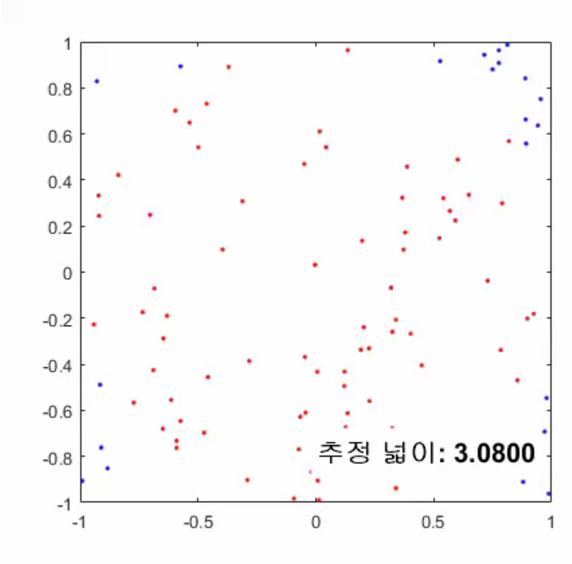
통계적인 수치를 얻기 위해 수행하는 '시뮬레이션' 같은 것
반복된 무작위 추출을 이용하여 함수의 값을 근사하는 알고리즘

왜 굳이 이런 시뮬레이션을 하는가?
→ 통계학의 특성상 무한히 많은 시도를 거쳐야만 진짜 정답이 무엇인지 알 수 있으나
그렇게 하기가 현실적으로 어렵기 때문에 유한한 시도만으로 정답을 추정하는데 의미가 있음

몬테카를로 기본 개념

몬테카를로 (Monte Carlo)

통계적인 수치를 얻기 위해 수행하는 '시뮬레이션' 같은 것
반복된 무작위 추출을 이용하여 함수의 값을 근사하는 알고리즘



가장 유명한 예제 : 원의 넓이 계산

가로, 세로의 길이가 2인 정사각형 안에 점을 무수히 많이 찍으면서
중심으로부터의 거리가 1이하이면 빨간색으로 칠하고, 그렇지 않으면 파란색으로 칠해줌

→ 전체적으로 찍은 점의 개수와 빨간색으로 찍힌 점의 개수의 비율을 계산하여
원래의 사각형 면적인 4를 곱해주면 반지름이 1인 원의 넓이 대략적으로 추정 가능
→ 점이 많이 찍힐 수록 3.14에 수렴

CHAPTER 04

회귀 분석

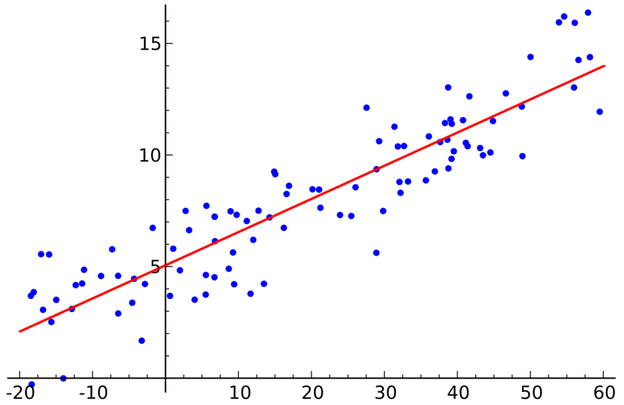
회귀 분석

회귀분석

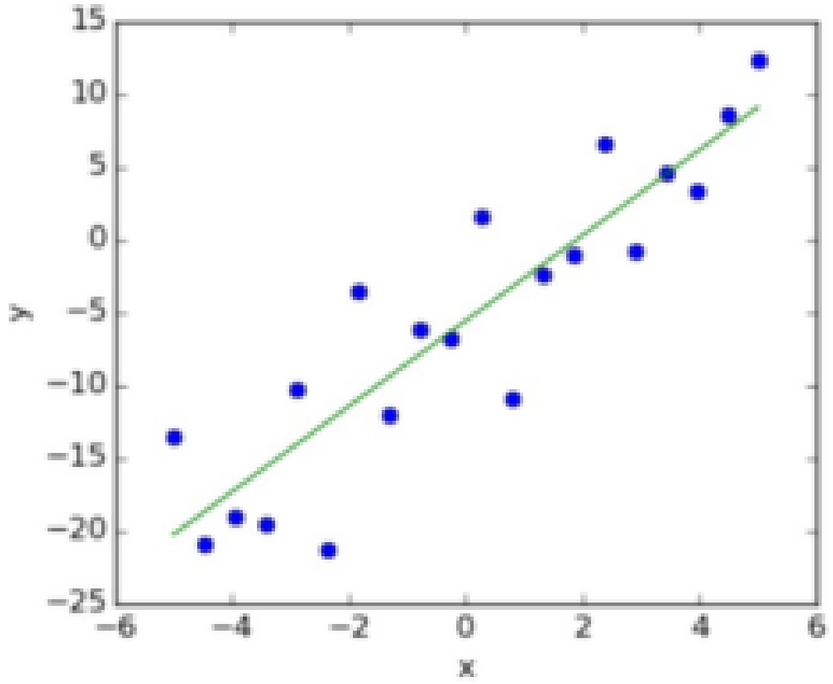
- 설명변수(X)에 대응하는 반응변수(Y)와 가장 비슷한 값($\hat{Y}$)을 출력하는 함수를 찾는 과정
- 변수들의 관계를 기술하고 형태를 파악하는 통계적인 기법

선형회귀분석

- 반응변수와 한 개 이상의 설명변수와의 선형 상관관계를 모델링하는 회귀분석 기법
- 예) 해당 연도 수확량(X)에 따른 열매 개수(Y)



단순 선형 회귀 분석



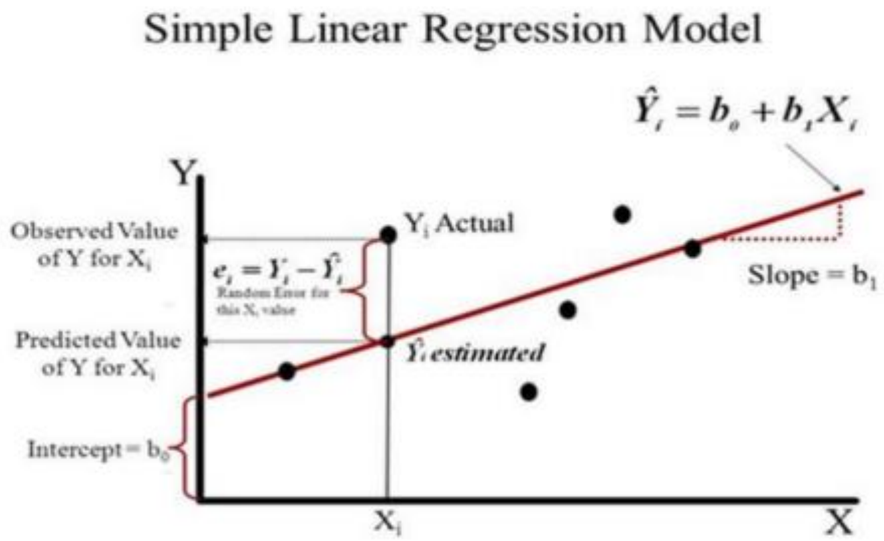
$$Y = \beta_0 + \beta_1 X + \varepsilon$$

β0, β1 : 회귀계수, ε : 오차항

→ 3가지 변수를 추정하는 것이 단순 선형 회귀 분석의 목표

단순 선형 회귀 분석

최소제곱추정(LSE, Least Squares Estimation)



잔차(e = y - $\hat{y}$)의 제곱합을 최소화

→ 최적의 회귀계수(모수) 추정

→ 값이 작을수록 좋은 모델

회귀 진단

1. 선형성

2. 독립성

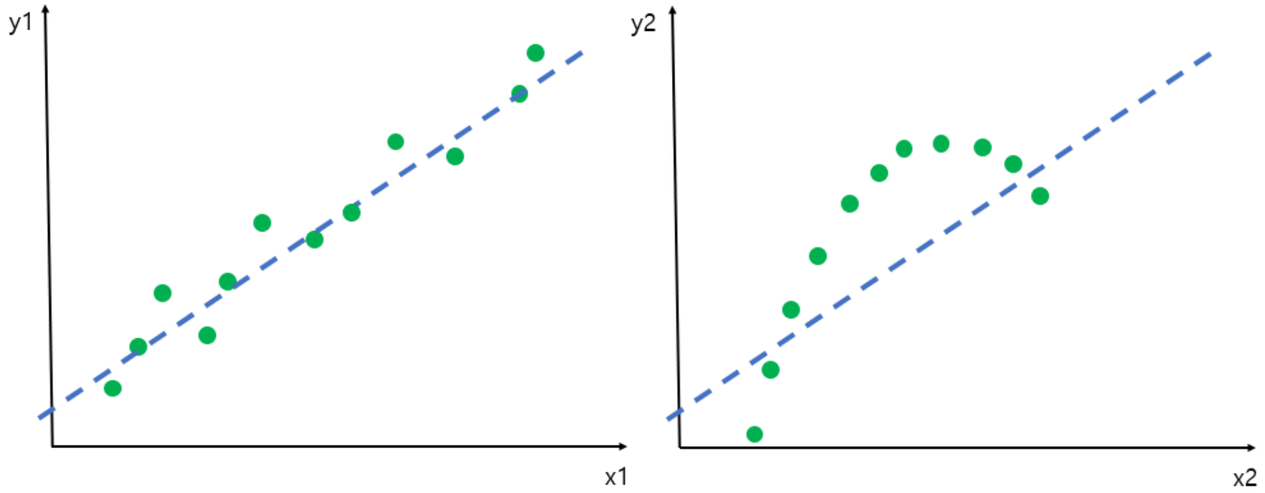
3. 등분산성

4. 정규성

회귀 진단

1. 선형성

- 종속 변수 Y와 독립 변수 X 간에 선형 관계가 있다는 것
  = Y와 X 사이의 관계를 직선으로 설명할 수 있어야 함

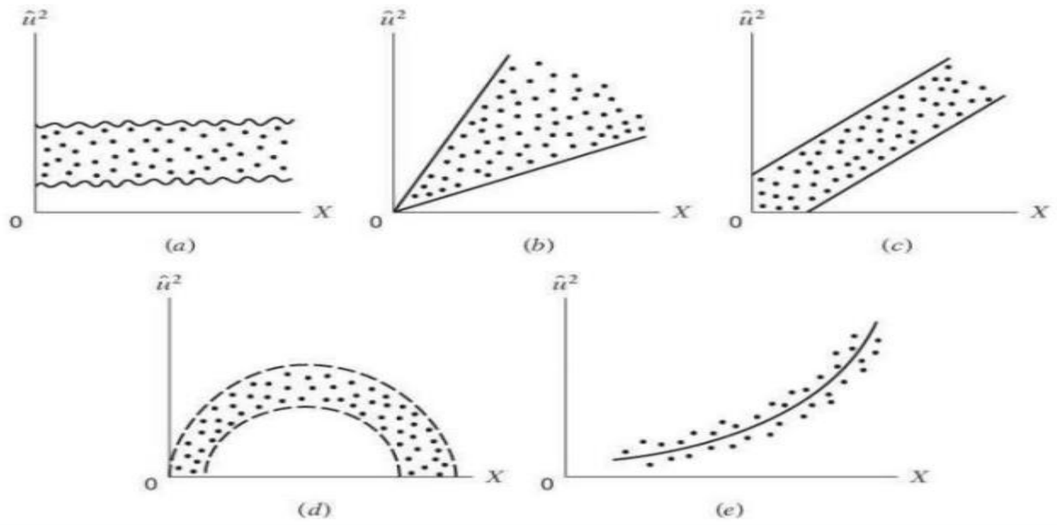


- 검증 방법
  - 산점도 → X, Y 사이의 관계를 시각적으로 확인

회귀 진단

2. 독립성

- 독립 변수 X의 각 관측값이 서로 독립적
  = 한 관측값의 X가 다른 관측값의 X에 영향을 주지 않아야 함



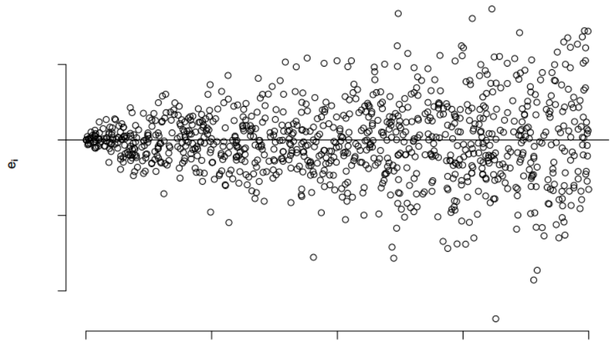
- 패턴을 보이면 X
  → (a)를 제외한 나머지는 독립성 가정 위배

- 검증 방법
  - Durbin-Watson 검정(듀런 왓슨), ACF

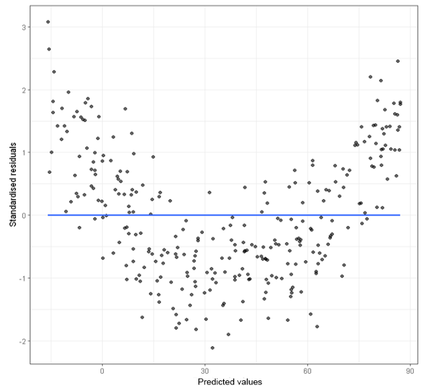
회귀 진단

3. 등분산성

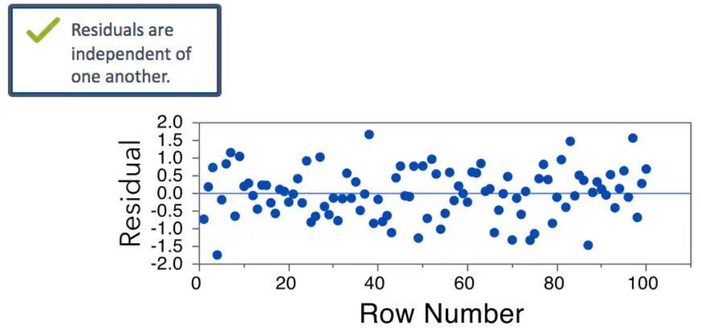
- 독립 변수 X의 값에 대해 종속변수 Y의 분산이 일정하다는 것을 의미
  = 회귀 모델이 다양한 X값에 대해 일관된 예측 정확도 유지해야 함



점점 넓어짐
= 잔차에 정보가 남아있음



잔차에 2차 함수 형태가 남아있음
= 잔차에 정보가 남아있음



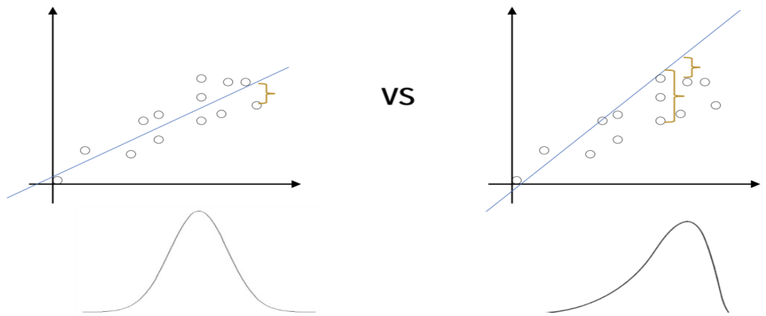
잔차가 0에 대해서 대칭, 무작위 형태
잔차에 아무런 정보도 남아있지 않음

회귀 진단

4. 정규성

- 회귀모형의 오차항이 정규 분포를 따름
- 정규성 가정이 무너지면 t-검정 등을 할 수 없게 됨

▪ **회귀분석의 진단**



- 회귀모델을 잘 만들었을 경우, 잔차는 정규분포를 따른다.

- 검증 방법
  - QQ-plot, Shapiro-Wilk 검정, Kolmogorov-Smirnov 검정

THANK YOU

감사합니다