[일머리사관학교 디지털헬스케어]

탐색적 데이터 분석

경남대학교 창의융합대학 교수 유현주
comjoo@kyungnam.ac.kr

강의 진행 안내

- **수업 자료 공유 URL**
  - https://github.com/yoohjoo/
  - 수업에 필요한 강의 자료 다운로드 저장

- **소통을 위한 메일 주소:** comjoo@kyungnam.ac.kr
  yoohjoo94@gmail.com

Contents

탐색적 데이터 분석

- Exploratory Data Analysis
- 기초 통계
- 상관 분석
- 회귀 분석

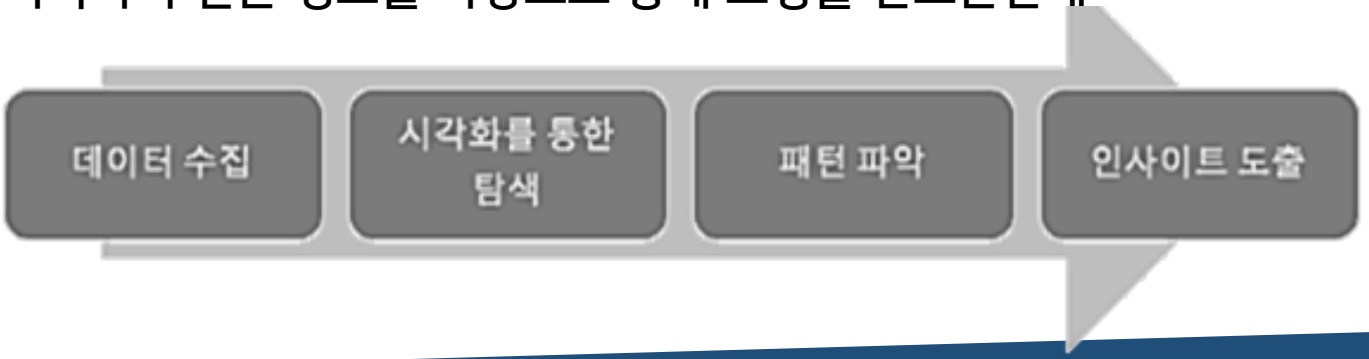
탐색적 데이터 분석(EDA)

Exploratory Data Analysis

데이터 분석

- **CDA(Confirmatory Data Analysis)** : 확증적 데이터 분석
  - 목적을 가지고 데이터를 확보하여 분석하는 방법 (연역적 방법)
  - 관측된 형태나 효과의 재현성평가, 유의성 검정, 신뢰구간 추정 등 통계적 추론하는 단계
  - 보통은 설문조사, 논문에 대한 내용을 입증하는데 많이 사용

- **EDA(Exploratory Data Analysis)** : 탐색적 데이터 분석
  - 쌓여있는 데이터를 기반으로 가설을 세워 데이터를 분석하는 방법 (귀납적 방법)
  - 데이터의 구조와 특징을 파악하여 얻은 정보를 바탕으로 통계 모형을 만드는단계
  - 빅데이터 분석에 사용됨



탐색적 데이터 분석(EDA)

- **데이터 탐색(EDA)**
  - 데이터 분석의 초기 단계에서 데이터의 특성을 살펴보는 과정
  - 데이터 세트를 분석 및 조사하고 주요 특성을 요약하여 데이터의 패턴 발견, 이상 징후 발견, 가설을 테스트, 가정을 확인하는 작업이라 할 수 있음
  - 통계학을 기반으로 데이터의 특성을 발견할 수 있음

- **통계학(統計, Statistics)**
  - 한데 몰아서 어림잡아 계산(표준국어대사전), 현상을 통계에 의하여 관찰 · 연구하는 학문
  - 불확실성에 대한 논리를 부여하는 학문, 경험과학의 한 분야이자 대부분 학문의 기초이며, 다양한 정의가 존재하고 축약하면 자료를 연구하는 학문, 데이터를 분석하는 학문
  - 기술통계학(Descriptive Statistics) : 데이터를 수집, 정리, 요약하여 데이터 의미를 기술(설명)
  - 추론통계학(Inferention Statistics) : 표본 자료에서 얻은 정보를 이용하여 전체 집단(단위)에 대한 정보 및 불확실한 사실에 대해 예측하는 방법과 이론을 제시

빅데이터시대의 기술 통계와 추론 통계

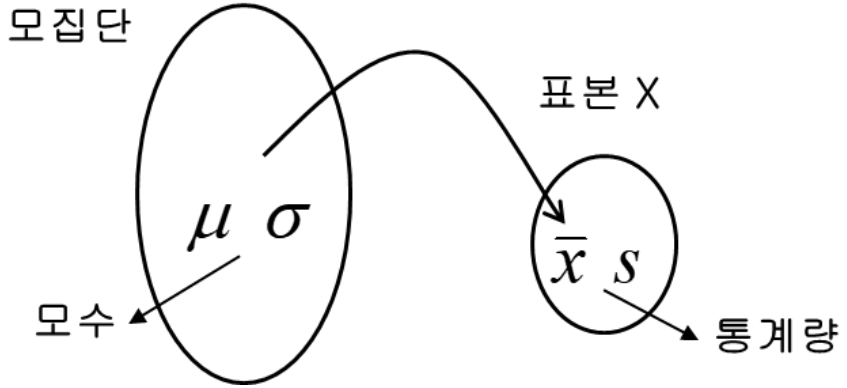
- 기술 통계
  - 기초통계로 데이터 요약 집계하여 데이터를 설명함
- 추론 통계
  - 가설 검정: 모집단에서 샘플링한 표본으로 모집단의 특성을 추론하고 그 결과가 신뢰성이 있는지 검정하는 과정

- **빅데이터 시대**
  - 모집단과 표본집단을 구분하기 보다는 **데이터 전체를 표본으로 하고 수집하지 못한 현실 세계 전체 데이터나 미래에 대한 데이터를 모집단으로 볼 수 있음**

예) 통신 회사의 고객 데이터 전체는 데이터 과학자의 샘플링 단계가 없더라도 표본 집단으로 보고, 수집하지 못한 다른 회사의 고객데이터나 다음 달 고객 데이터는 모집단이라고 볼 수 있음

통계 기초 이론: 모집단과 표본

모집단	표본
관심의 대상이 되는 모든 개체의 관측값이나 측정값의 집합	모집단에서 실제로 추출한 관측값이나 측정값



구분	모수	통계량
대상	모집단의 특성	표본에서 계산한 특성
표시	그리스, 로마자로 표시	알파벳으로 표시
평균	μ	$\bar{x}$
표준편차	σ	s

통계 기초 이론: 기술통계

※ “Descriptive : 묘사하는, 그려서 설명하는”

- 수집한 데이터를 요약 묘사 설명하는 통계 기법
- 기술통계(Descriptive Statistics) 기법
  - 수집한 데이터를 대표하는 값이 무엇인지 찾기 → 대표값
  - 수집한 데이터가 어떻게 퍼져 있는지를 설명하기 → 데이터 분포

- 대푯값
  - 주어진 자료를 대표하는 특정 값
  - 대표값은 자료의 중심적인 경향이나 자료분포의 중심의 위치를 나타냄
  - 평균 / 중앙값 / 최빈값 등

- 데이터 분포
  - 주어진 자료의 퍼짐 정도를 표현하는 값 또는 시각화 차트
  - 분산 / 표준편차 / 사분위값 / 왜도 / 첨도 / 도수분포표 / 히스토그램

통계 기초 이론: 대푯값

- 평균
  - 주로 산술평균을 사용
    - 산술평균: 데이터 모음의 값을 모두 더한 후 데이터 개수로 나눈 값
    - 산술평균외에 용도에 따라 **기하평균, 조화평균, 가중평균** 등을 사용하기도 함
  - 데이터 중 극단적인 값 또는 이상한 값이 섞여 있거나 최대값이나 최소값의 크기가 크게 차이 나면 평균값이 데이터 모음의 대푯값으로 왜곡된 의미를 가질 수 있음

- 중앙값
  - 수집된 데이터 모음의 값을 크기별로 나열하였을 때 가장 중앙에 위치한 값
    - 데이터 개수가 짝수개이면 중앙 위치의 두 데이터의 평균으로 표시하기도 함

- 최빈값
  - 수집된 데이터 모음에서 가장 자주 발생한 값(빈도수가 가장 큰 값)
    - 주로 대소관계가 의미 없는 자료(명목형)에서 많이 사용됨

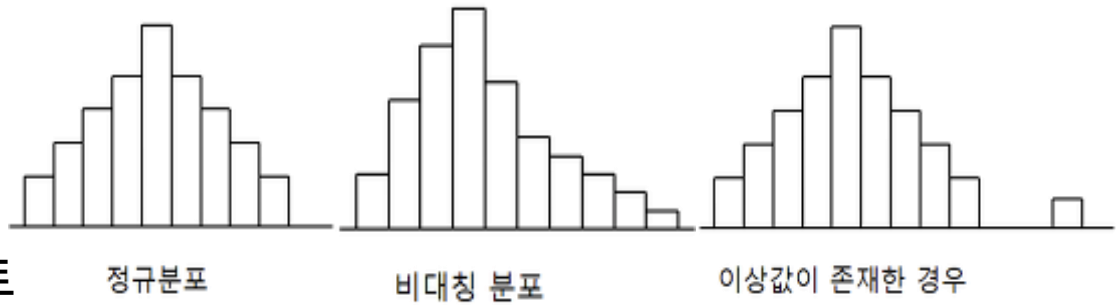
통계 기초 이론: 기술 통계 값의 의미

- 중심성 : 데이터가 어느 부분에 집중되는가?
  - 평균, 중앙값, 최빈값
- 변동성 : 데이터의 중심으로부터 얼마나 떨어져있는가?
  - 분산, 표준편차, 범위(최대값-최소값), 사분위값
- 정규성: 데이터의 분포 모양이 정규 분포인가?
  - 왜도, 첨도

통계 기초 이론: 데이터 분포

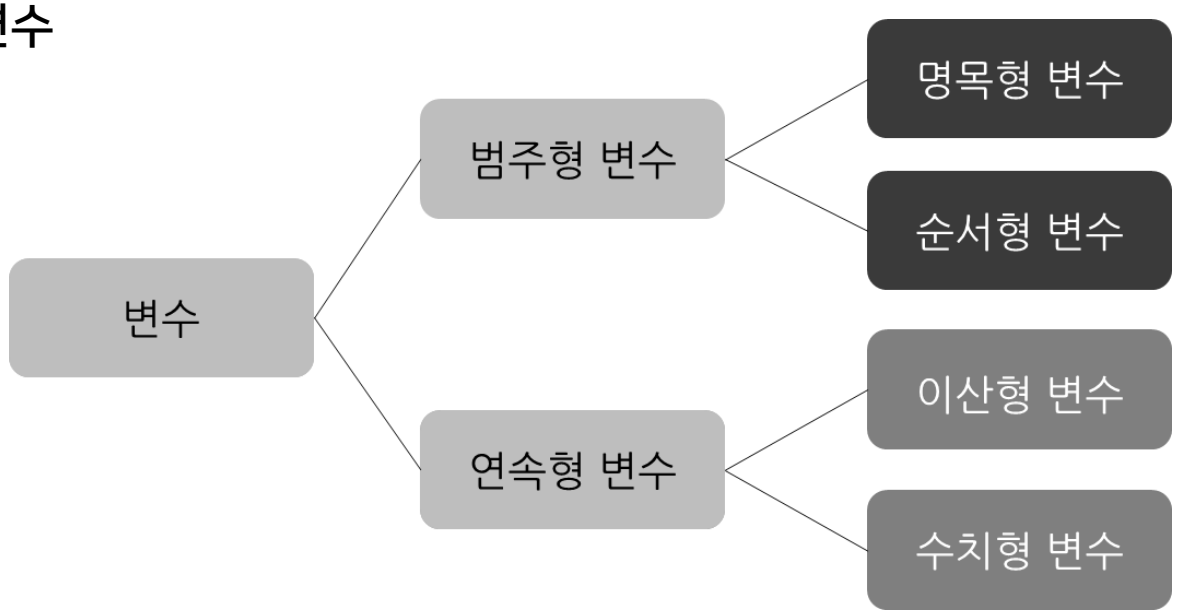
- 도수분포표(Frequency table)
  - 자료의 분포를 구간으로 나누고 각 구간에 속하는 값이 몇 개인지 빈도를 나타낸 표
  - 전체 분포를 요약해서 파악할 수 있음
    - 자료의 개수, 자료 중 최대값과 최소값을 구함
    - 구간(계급수)의 개수를 결정: 자료의 개수나 분포에 따라 다름
      - 각 구간에 5개 이상의 값이 들어가는 것을 추천 (일반적으로 5~15구간 추천)
    - 구간의 폭(계급의 폭)을 구함: 구간폭 = (최대값 – 최소값) / 구간수
      - 정수, 짝수, 5의배수 등의 사용을 추천
    - 구간의 경계값을 구함: 최소값에서 부터 구간폭을 더해서 구간의 경계값을 구함
    - 구간별 값의 개수(도수)를 표시

- 히스토그램(Histogram)
  - 도수분포를 막대그래프로 시각화한 차트
  - 그래프 모양(종모양, 비대칭 종모양 등)으로 데이터 분포를 파악할 수 있음



통계 기초 이론: 변수 유형

- 범주형 변수
  - 연속형 데이터와 범주형 데이터의 가장 큰 차이는 가감승제가 가능하지만 의미가 없다는 것
    - 예를 들어 성별을 1(남자), 2(여자)로 구분하여 이들의 평균을 구하면 1.5가 됨
  - 명목변수, 서열변수
- 연속형 변수
  - 사칙연산이 가능. 예로부터 자연현상에 대해 과학적 수치를 부여하는 과정을 "수치화" 라고 함
  - 등간, 비율변수



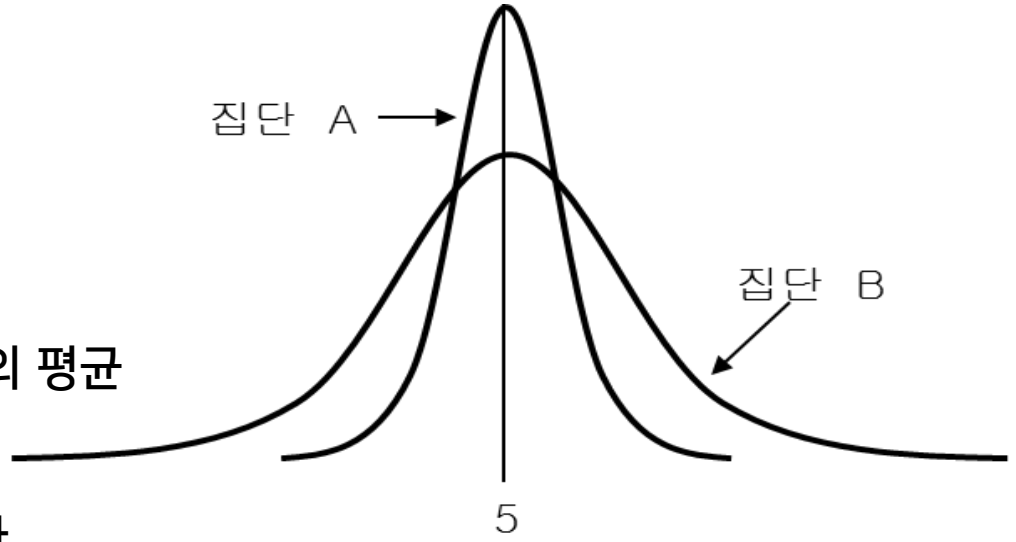
통계 기초 이론: 범주형 변수 분석

- 범주형 변수의 분석의 첫 단계는 해당 변수의 빈도(Frequency)
- 전체데이터 빈도를 계산하는 것부터 시작
- 빈도분석
  - 빈도는 대상 자료에서 반복하는 횟수를 기록하는 것으로 빈도분석을 통해 전반적 분포를 살펴보면서 전체 데이터의 구성을 파악 과정
  - 가장 빈도가 높은 데이터와 가장 빈도가 낮은 데이터를 기록하고 데이터 분석을 시작해야 함
- 교차분석
  - 교차분석은 범주형 변수간 빈도분석을 확장하여 빈도를 교차시킨 분할표(Contingency Table)를 만들어 분석하는 방법
  - 분할표는 교차분석표(Cross table) 라고도 함

통계 기초 이론: 연속형 변수 분석

- 데이터 중심 이해
  - 데이터의 중심은 여러 자료들의 비교하였을 경우 중앙에 위치하는 값으로 자료의 특성을 보여줌
  - 평균, 중위수

- 데이터 퍼짐 정도 이해
  - 분산: 데이터 퍼짐 정도
  - 표준편차: 데이터의 퍼짐 정도를
    동일 한 기준을 적용하기 위하여 편차 제곱합의 평균
  - 범위: 데이터의 흩어진 범위
  - 사분위: 데이터의 분포가 좌우대칭이 아니거나
    이상치가 있는 경우 평균은 극단적으로 치우진 대표성이 없는 값에 의해 영향을 받음
    - 이 경우 자료를 나열하여(시각화) 전체 데이터를 파악할 수 있음
    - 자료를 순서대로 나열했을 때 50%에 위치하는 수가 중위수(Q2) 이고 25%에 위치하는 수가 Q1

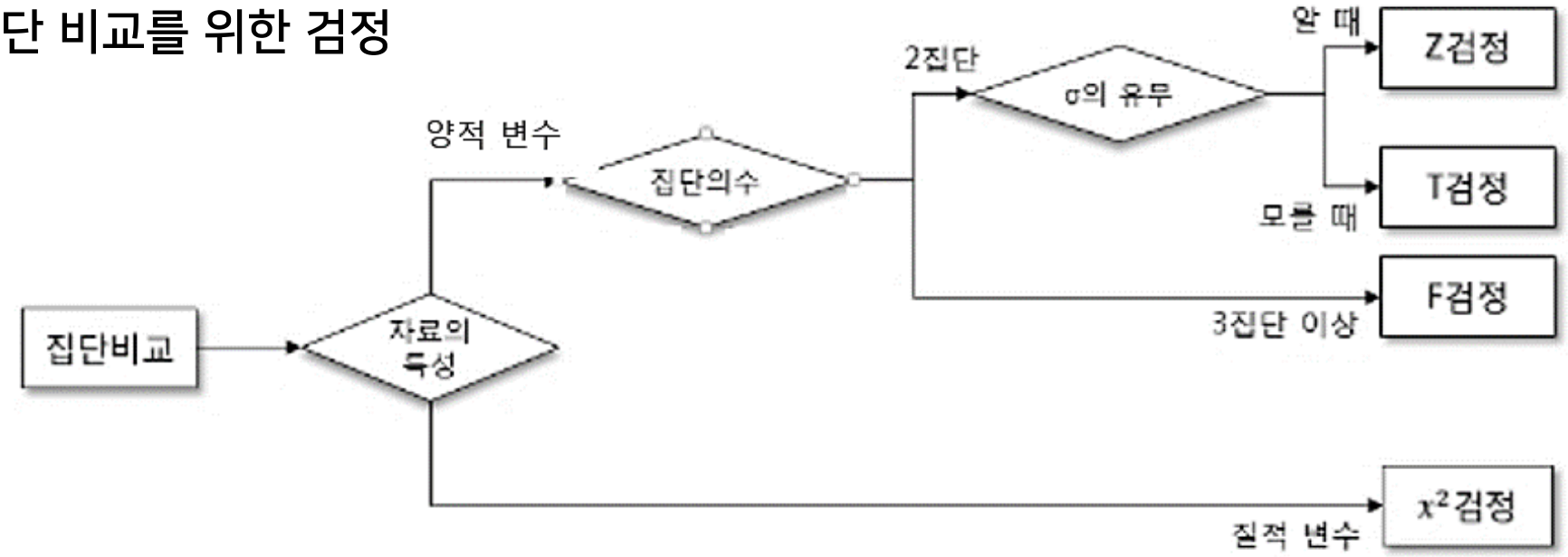


통계 기초 이론: 분석 모형 선택 기준

- 분석하고자 하는 내용(주제) 즉, “가설”을 명확히 규명
- 가설에 맞는 데이터의 변수 척도를 파악 → 변수 척도에 따라 적용할 모형이 정해짐
- 가설 검정의 구분
  - 차이 검정: 집단 간 평균 차이
  - 관계 검정: 함수적 관계 규명
- 종속 변수의 종류가 연속형과 이산형에 따라 분석 모형을 결정
  - 종속변수 **Y** 가 연속형일 때
    - 차이 : T-test, ANOVA(분산분석)
    - 관계 : Regression (회귀분석)
  - 종속변수 **Y** 가 이산형일 때
    - 차이 : Chi-square Independence Test (카이제곱 독립성검정)
    - 관계 : Logistic Regression (로지스틱 회귀분석)

통계 기초 이론: 차이 분석

● 집단 비교를 위한 검정



집단	대상
단일 표본 비교	하나의 집단(단일 표본 - one sample)중 관심 있는 연속형 변수의 모평균이 어떤 특정 값과 같은지 알아보고자 할 때
두 집단 간 평균비교	독립 표본 T : 서로 독립인 두 표본에 의한 모평균 비교 대응 표본 T : 대응하는(Paired) 쌍에 대한 차의 모평균 검정
3 이상 집단간 평균비교	일원배치분산분석 (One-way ANOVA)

통계 기초 이론: T-분석

- T-분석, T-검정은 두 집단의 평균을 비교하는 통계적 검정 방법
- 모집단을 대표하는 표본으로부터 추정된 분산이나 표준편차를 가지고 검정하는 방법으로 "두 모집단의 평균간의 차이는 없다"라는 귀무가설과 "두 모집단의 평균 간에 차이가 있다"라는 대립가설 중에 하나를 선택할 수 있도록 하는 통계적 검정방법
- 단순히 차이의 존재 여부를 떠나 두 집단의 비교가 통계적으로 의미가 있는가를 검정
- 즉, 두 모집단의 차이가 우연에 의해서 인지 아닌지를 검정하는 방법
- T분석의 기본 가설

✓ 가설

- 영 가설 : 집단간의 평균 차이는 없다.
- 대립가설 : 집단간의 평균 차이는 있다.

$$t = \frac{\overline{X} - \mu}{\text{표준오차(SE)}\ \frac{S}{\sqrt{n}}}$$

$\overline{x}$: 두 집단 차이의 평균

μ : 모집단의 평균

S : 두 집단 차이의 표준편차

- 30개 이하의 비교적 적은 수의 표본에 대해 활용 → 30개 이상이면 정규분포의 Z검정
- 모집단의 표준편차를 알 수 없을 때 사용

통계 기초 이론: 상관분석

- 두 연속형 변수 사이 상관관계가 존재하는지를 파악하고, 상관관계의 정도를 확인 하는 것이 상관분석(Correlation analysis)이라 함
- 상관분석에서는 관련성을 파악하는 지표로 상관계수(Correlation coefficient)라는 통계학적 관점에서 선형적 상관도를 확인하여 정도를 파악
- 상관분석은 간단한 분석이지만 머신러닝의 기반이 됨

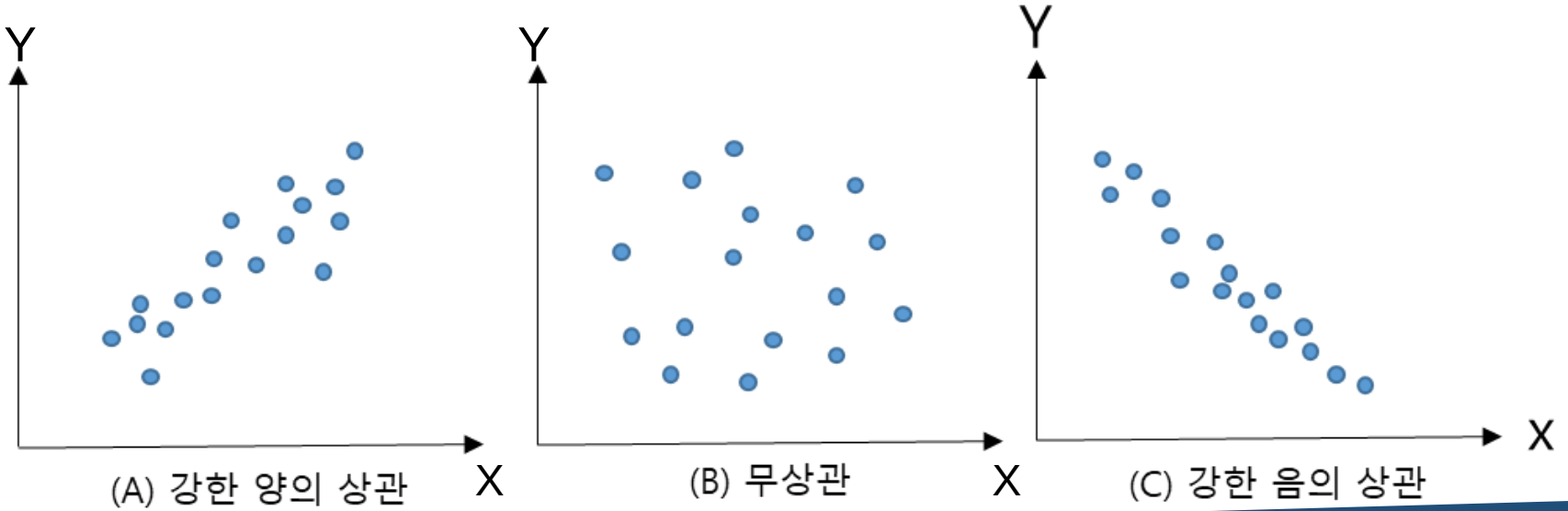
① 산점도(Scatter) 두 변수 상관 파악

② 상관계수 확인

③ 의사결정

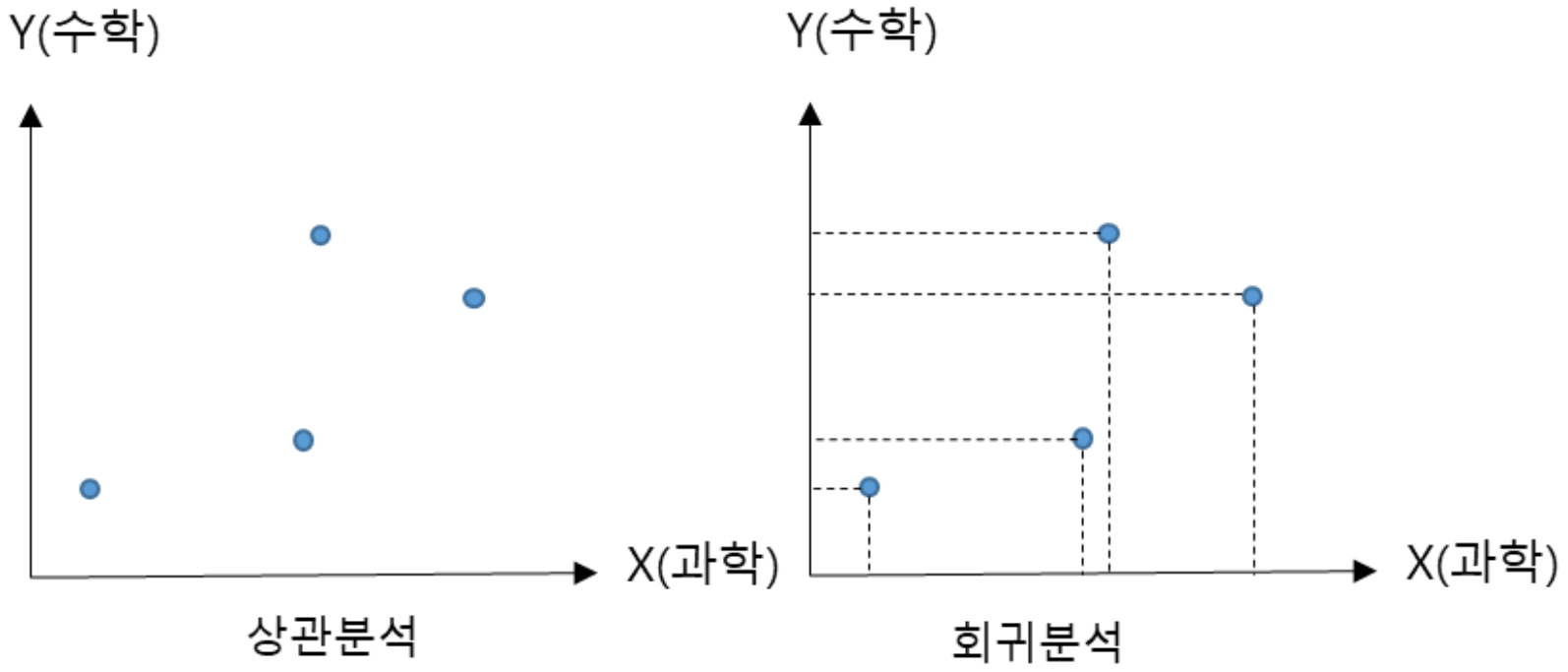
통계 기초 이론: 상관분석

상관	상관계수
양의 상관	+0.1 ~ +0.3 이면, 약한 양의 상관관계 +0.3 ~ +0.7 이면, 뚜렷한 양의 상관관계 +0.7 ~ +1.0 이면, 강한 양의 상관관계 – 그림A
무상관	-0.1 ~ +0.1 이면, 없다고 할 수 있는 상관관계 - 그림 B
음의 상관	-1.0 ~ -0.7 이면, 강한 음의 상관관계 – 그림C -0.7 ~ -0.3 이면, 뚜렷한 음의 상관관계 -0.3 ~ -0.1 이면, 약한 음의 상관관계



통계 기초 이론: 회귀분석

- 상관분석에서는 두 연속형 변수 X(과학)와 Y(수학)의 상관 정도만 알 수 있고 인과관계는 알 수 없었음
- 회귀분석에서는 두 연속형 변수 X와 Y를 독립변수와 종속변수라고 하는 인과관계로 설명
- '과학 점수가 좋으면 수학점수가 좋을까요?' 와 같이 간단 하지만 미래를 예측할 수 있는 머신러닝의 초기 모델이 됨

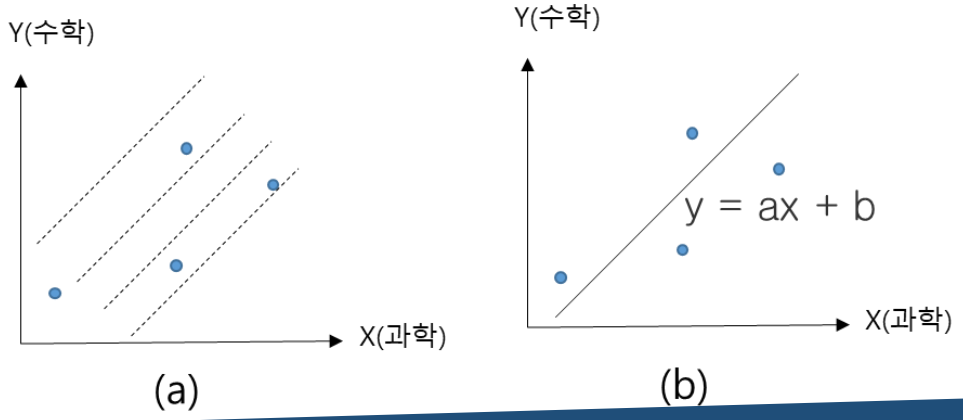


통계 기초 이론: 회귀분석

- 선형회귀분석(Linear Regression Analysis)은 쌍으로 관찰된 연속형 변수들 사이의 관계에 있어서 한 변수를 원인으로 하고 다른 변수들을 결과로 하는 분석
- 독립변수와 종속변수 사이 선형식을 구하고 그 식을 이용하여 변수값 들이 주어 졌을 때 종속변수의 변수 값을 예측하는 분석방법

X	Y
독립변수, 설명변수, 원인변수	종속변수, 반응변수, 결과변수 머신러닝(클래스, 라벨)
다른 변수에 영향을 주는 원인	다른 변수에 영향을 받는 결과

- x변수와 y변수 간의 관계를 y = ax + b 와 같은 하나의 선형 관계식으로 표현
- y = ax + b 인 회귀식에서 독립변수 x가 하나인 것이기에 단순선형회귀분석이라함



통계 기초 이론: 변수 종류에 따른 통계 분석 종류

- 종속변수와 독립변수의 종류에 따라 통계 분석법의 종류

Y 종속변수 (반응변수)	X 독립변수(설명변수)	통계분석법	귀무가설
연속형	범주형(2개 범주)	T-검정, paired T-검정	집단 간 평균이 동일
연속형	범주형(3개 이상)	분산분석(ANOVA)	집단 간 평균이 동일
연속형	연속형	회귀분석	회귀 계수 = 0
연속형	혼합(수치형+범주형)	공분산분석(ANCOVA)	
범주형	범주형	$\chi 2$검정/로짓분석	집단 간 연관성이 없음/ 회귀 계수 = 0
범주형	혼합(수치형+범주형)	로짓분석	
생존시간	혼합(수치형+범주형)	생존분석	

통계 기초 이론: 가설 수립

- 가설(Hypothesis)은 모수에 대한 예상, 주장, 또는 단순한 추측
  - 예) '외계인은 존재한다, 그 사람은 유죄이다' 처럼 아직은 하나의 추측인 사실을 가설이라 함
- 통계적 가설 검정은 가설에 대해 증거를 수집하고 과학적으로 증명하는 과정
- 가설 검정의 첫 단계는 가설 수립부터 시작
  - 검정하고자 하는 모집단의 모수(조사하고자 하는 자료의 평균, 분산, 표준편차, 상관계수)에 대해 항상 귀무가설과 대립가설 두 가지로 수립

영(귀무)가설(null hypothesis)	대립가설(alternative hypothesis)
H0	H1
기각하기를 희망하여 형식화한 가설, 기존에 받아들이던 가설 모수에 관한 귀무가설은 항상 모수의 정확한 값을 지정하도록 진술 될 것인 반면 대립가설에서는 여러 개의 값의 가능성이 허용	표본을 통해 입증하고자 하는 새로운 가설 모수에 대한 관심의 영역 중에서 귀무가설로 지정되지 않은 모든 경우를 포괄적으로 지정
외계인은 존재하지 않는다. 그 사람은 무죄이다	외계인은 존재한다 그 사람은 유죄이다

통계 기초 이론: 가설 검정 단계

구분	사례1
가설 수립	외계인이 존재할까? 외계인이 존재한다는 확실한 증거 수집 전까지는 외계인은 없다라고 한다. H0 : 외계인 = 0. 외계인은 0 명이다. 외계인은 없다 H1 : 외계인 ≠ 0. 외계인은 0 명이 아니다. 외계인은 있다
가설 검정	외계인이 있다라는 증거가 많이 있다 ⇨ 외계인은 존재한다. 외계인이 있다라는 증거가 조금밖에 없다 ⇨ 외계인은 존재한다는 증거가 부족하다
증거 수집	외계인은 존재하는가? 외계인이 존재한다는 객관적인 증거가 97% 있다. 증거가 95%보다 많으므로 외계인은 존재한다

구분	사례2
가설 수립	그 사람은 무죄일까? 유죄일까? 법정에서는 유죄 판결을 받기 전까지 모든 사람들은 무죄 H0 : 기존에 받아들이던 가설 죄=0. 죄가 없다. 무죄 H1 : H0를 기각하기를 바라는 가설 죄≠0. 죄가 있다. 유죄
가설 검정	무죄가 아니라는 증거가 많이 있다(유죄) 무죄가 아니라는 증거가 조금밖에 없다(증거 불충분으로 무죄)
증거 수집	그 사람은 무죄인가? 유죄인가? 유죄라는 객관적인 증거가 80% 있다 증거가 95%보다 적으므로 증거불충분으로 무죄

통계 기초 이론: 오류의 이해

- 유죄라는 객관적인 증거와 외계인이 존재한다는 객관적인 증거는 얼마나 필요할까? 일반적으로 95% 정도가 필요
- 그럼 객관적인 증거가 95%보다 많으면 실제로 외계인은 존재할까(H1)?
- 아니면 객관적인 증거가 95%보다 적으면 실제로 무죄일까(H0)?
- 반드시 그런 것은 아니며 이것을 오류라 하며, 통계적 오류는 다음과 같이 구분

구분	영가설진실(H0)	대립가설진실(H1)
영가설선택**(H0)**	옳은 판단 신뢰수준(1-α)	2종 오류 유의수준(β)
대립가설선택**(H1)**	1종 오류 유의수준(α)	옳은 판단 신뢰수준(1- β)

통계 기초 이론: 유의수준

- 유의 확률 P

비교	결과
P < α = 0.05 **(오류 5% 이하, 95% 이상 진실)**	H1 선택
P >= α = 0.05 **(오류 5% 이상, 95% 이하 진실)**	H0 선택

- 유의수준(α)는 항상 0.05로 고정되어 있을까? 그렇지 않다.
- 유의수준은 일반적으로 0.05를 사용하는데, 연구자의 기준에 따라서 변할 수 있다.
- 그러나, 유의수준은 분석 전에 미리 결정을 하여야 함.
- 유의수준을 0.01로 하면 어떻게 될까? 혹은 유의수준을 0.1로 하면 어떻게 될까?

통계 기초 이론: 통계 결과 해석

- 결과 해석은 유의수준과 유의확률을 비교하여 결정
- 유의수준은 제1종 오류의 최대 허용 한계
  - 유의수준 α 값이 작아지면(오류가 작아진다) 영가설이 틀렸다는 결론을 내리기 어려움
  - 반대로 유의수준 α 값이 커지면 귀무가설이 틀렸다는 결론을 내리기 쉬워짐
- 유의확률(검정통계량)은 P값 또는 P-value 라고 함
- 영가설(H0)이 맞을 경우, 대립가설 쪽의 값이 나올 확률이 얼마나 되는지를 나타내는 값으로 결론적으로 통계는 유의확률 P와 유의수준 α를 비교하여 영가설과 대립가설을 선택하는 과정
- 대부분통계는 새로운 가설을 선택하는게 목적
- 주로 판정 기준을 이렇게 표현

P < 0.05 이하 기준 이면 새로운 대립가설(H1) 을 선택

통계 기초 이론: 통계 절차 요약

① 통계분석방법 선정

② 분석하고자 하는 목적에 따른 귀무가설(영가설)과 대립가설 설정

③ 분석도구(Excel, SPSS, R, Python programming 등) 검정통계량 실행 및 확인

④ 유의수준(α) 결정 : 0.1, 0.05, 0.01

⑤ 유의확률(P) 확인

⑥ 유의확률과 유의수준 비교 (< α)

⑦ 귀무가설 과 대립가설 선택

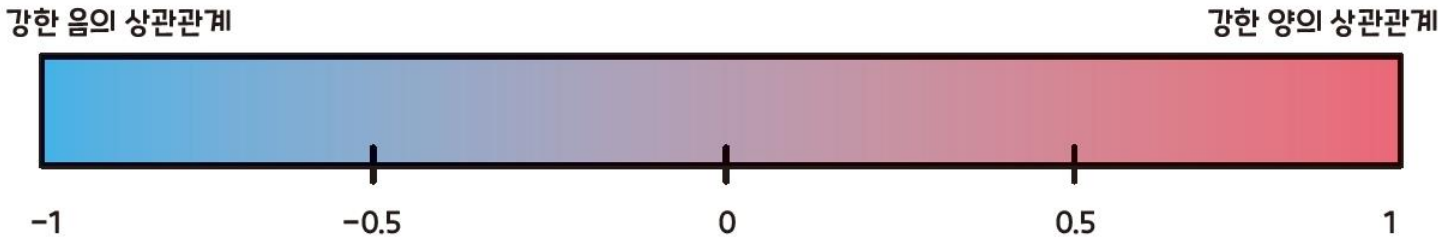
⑧ 분석 결론

상관관계 분석의 개념

상관관계 분석의 개념

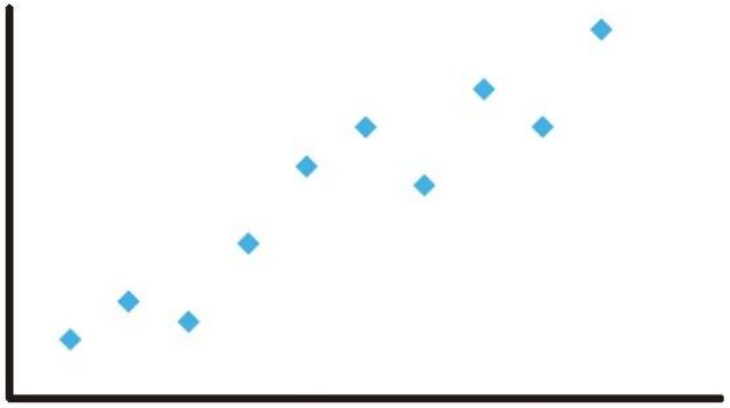
I. 상관관계 분석과 상관계수

- 상관관계 분석(Correlation analysis, 상관분석)
  - 두 변수 사이 관계의 강도와 방향을 파악하는 통계 기법.
  - 상관관계의 강도를 나타낸 수치를 상관계수(Correlation coefficient).
  - 변수 x와 y가 있을 때 두 변수의 상관관계는 다음 세 가지 중 하나.
    - »양의 상관관계: 변수 x가 커질수록 변수 y도 커짐.
    - »음의 상관관계: 변수 x가 커질수록 변수 y는 작아짐.
    - »상관관계 없음: 변수 x가 커질 때 변수 y는 커질 수도, 작아질 수도 있음.



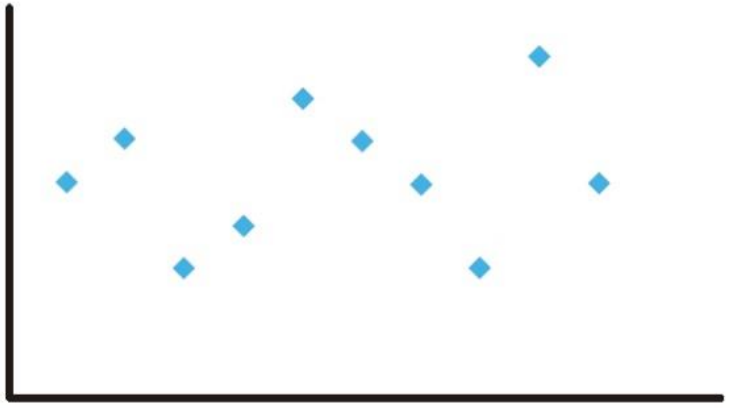
상관관계 분석의 개념

- 상관관계 분석과 상관계수



(a) 강한 양의 선형 상관관계

(b) 약한 양의 선형 상관관계

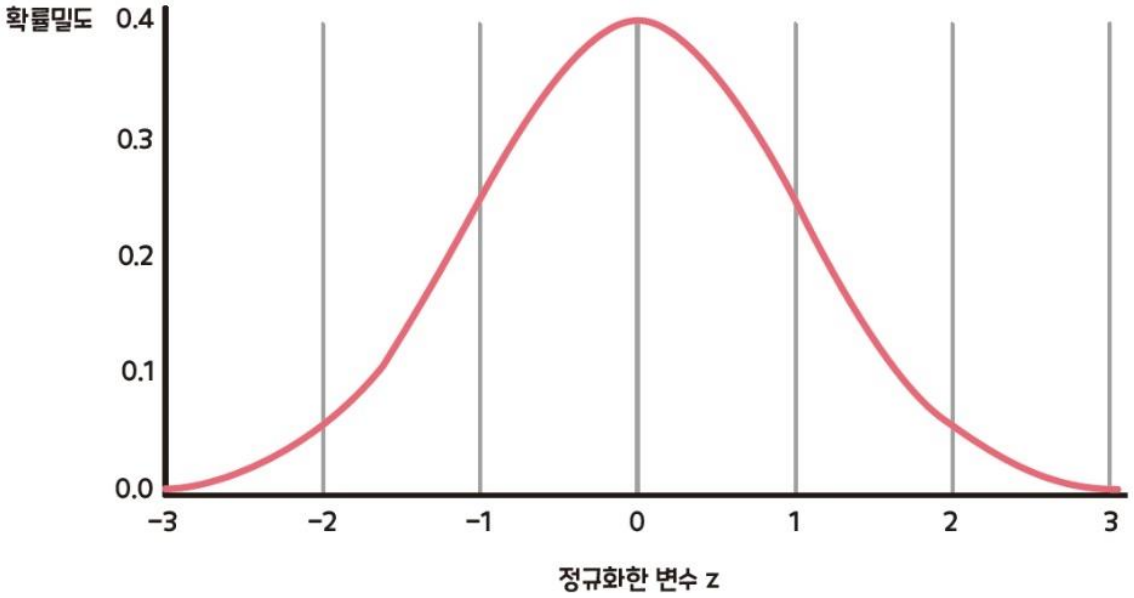


(c) 선형 상관관계 없음

(d) 강한 음의 선형 상관관계

상관관계 분석의 개념

- 상관관계 분석의 세 가지 방법
  - 피어슨 상관분석(Pearson correlation analysis) : 가장 일반적인 상관분석 방법.
  - 스피어만 상관분석(Spearman correlation analysis) : 두 변수가 정규성을 보이지 않을 때 사용하기 적합한 방법.
  - 켄달 상관분석(Kendall correlation analysis) : 스피어만 상관분석과 비슷하나 표본 데이터가 적고 동점이 많을 때 사용하기 적합한 방법.



✓하나 더 알기: 상관관계 분석의 의미

상관계수만 가지고 두 변수 사이의 상관성이 있는지 없는지 판단할 수는 없음.

두 변수에 선형 상관관계가 아닌 다른 상관관계가 있을 수 있음.

상관관계 분석의 개념

- 피어슨 상관분석(1)
  - 어린이가 영어 동요에 노출된 시간과 영어 점수와의 상관관계 분석해보기.

피어슨 상관분석(1)

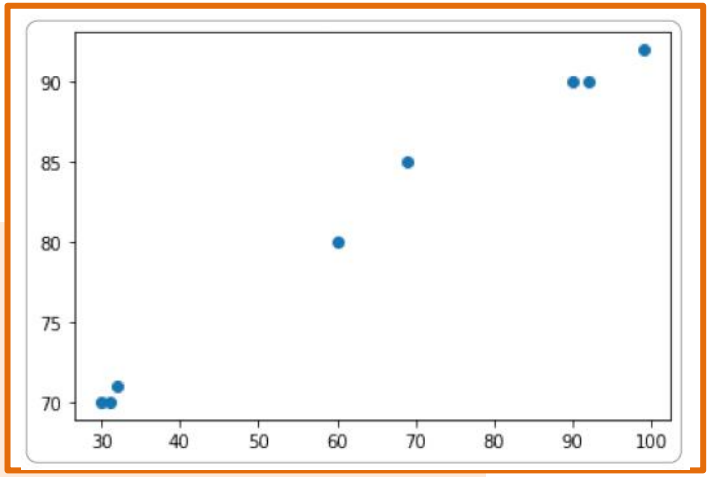
```
import pandas as pd
#리스트에 데이터 삽입하기
engListening = [30, 60, 90]
engScore = [70, 80, 90]

#리스트를 데이터프레임으로 변환하기
data = {'engListening':engListening, 'engScore':engScore}
df = pd.DataFrame(data)

#상관분석 수행하기
coef = df.corr(method='pearson')
print(coef)
```

상관관계 분석의 개념

- 피어슨 상관분석
  - 데이터의 산포도를 산점도 그래프로 확인.



산포도 확인

```
import matplotlib.pyplot as plt

#데이터 추가하기
engListening = [30, 60, 90, 31, 32, 69, 92, 99]
engScore = [70, 80, 90, 70, 71, 85, 90, 92]
data2 = {'engListening':engListening, 'engScore':engScore}
df2 = pd.DataFrame(data2)

#산점도 그래프의 x좌표와 y좌표 설정하기
plt.scatter(df2['engListening'], df2['engScore'])
plt.show()
```

상관관계 분석의 개념

- 피어슨 상관분석
  - 데이터를 추가한 데이터프레임 data2의 선형 상관도는 0.995829.
  - 매우 강한 선형 상관성이 있다고 말할 수 있음.

피어슨 상관분석(2)

```
coef = df2.corr(method='pearson')
print(coef)
```

```
             engListening   engScore
engListening     1.000000   0.995829
engScore         0.995829   1.000000
```

상관관계 분석의 개념

- 스피어만 상관분석과 켄달 상관분석
  - 판다스의 corr() 함수에서 method 인자 'pearson'을 'spearman'과 'kendall'로 변경.
  - 상관계수의 값은 분석 방법 종류에 따라 조금씩 다를 수 있음.

스피어만 상관분석과 켄달 상관분석

```
#스피어만  상관분석
spearmanCoef = df.corr(method='spearman')
print(spearmanCoef)

#켄달  상관분석
kendallCoef = df.corr(method='kendall')
print(kendallCoef)
```

```
               engListening  engScore
engListening           1.0       1.0
engScore               1.0       1.0
               engListening  engScore
engListening           1.0       1.0
engScore               1.0       1.0
```

✓하나 더 알기: 스피어만 상관분석과 피어슨 상관분석의 차이점

피어슨 상관분석은 두 연속 변수 간의 선형 관계를 측정하는 반면,

스피어만 상관분석은 선형인지 여부에 관계없이 변수 간의 단조 연관성(Monotonic relationship)을 측정.

피어슨 상관관계가 스피어만 상관관계보다 데이터의 이상치(Outlier)에 민감하게 반응.

상관관계 분석의 개념

- 피어슨/스피어만/켄달 상관분석을 각각 수행.

상관분석 결과

```
#피어슨 상관분석
pearsonCoef = df3.corr(method='pearson')
print(pearsonCoef)

#스피어만 상관분석
spearmanCoef = df3.corr(method='spearman')
print(spearmanCoef)

#켄달 상관분석
kendallCoef = df3.corr(method='kendall')
print(kendallCoef)
```

상관관계 분석의 개념

- 피어슨/스피어만/켄달 상관분석을 각각 수행.

실행결과

```
             engListening engReading endClass engScore
engListening     1.000000   0.877201 0.703028 0.995829
engReading       0.877201   1.000000 0.808755 0.894111
endClass         0.703028   0.808755 1.000000 0.759453
engScore         0.995829   0.894111 0.759453 1.000000
             engListening engReading endClass engScore
engListening     1.000000   0.826362 0.717256 0.988024
engReading       0.826362   1.000000 0.852757 0.848500
endClass         0.717256   0.852757 1.000000 0.725950
engScore         0.988024   0.848500 0.725950 1.000000
             engListening engReading endClass engScore
engListening     1.000000   0.618284 0.563621 0.963624
engReading       0.618284   1.000000 0.750568 0.679366
endClass         0.563621   0.750568 1.000000 0.584898
engScore         0.963624   0.679366 0.584898 1.000000
```

- 피어슨 상관분석 결과 engListening과 engScore의 상관계수가 0.995829로 가장 큰 선형 상관성을 보였으며, engListening과 engClass의 상관계수가 0.703028로 가장 작은 선형 상관성을 보였음
- 스피어만 상관분석과 켄달 상관분석에서도 engListening과 engScore의 선형 상관성이 가장 크고 engListening과 engClass의 선형 상관성이 가장 작게 나타남

✓하나 더 알기: 켄달 상관분석의 특징

켄달 상관분석은 두 변수 간의 순위를 비교하여 연관성을 계산함.

한 변수가 증가할 때 다른 변수가 함께 증가하는 횟수와 감소하는 횟수를 측정하여 횟수의 차이를 상관계수로 표현하는 방법.

순위로 표현할 수 있는 데이터이거나, 표본 크기가 작거나, 데이터의 순위에 동률이 많을 때 활용.

상관관계 분석의 활용

상관관계 분석의 활용

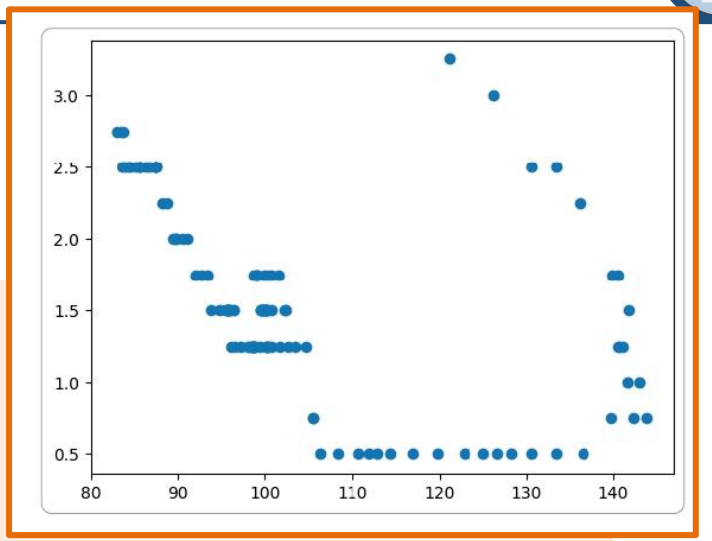
I. 기준금리와 부동산 매매가격

- 기준금리와 아파트 가격의 관계를 분석.

연월(yymm)	부동산 지수	기준금리(%)
1301	83	2.75
1302	83	2.75
1303	83.5	2.75
1304	83.8	2.75
1305	83.9	2.5
...	...	...
2207	136.1	2.25
2208	133.4	2.5
2209	130.6	2.5
2210	126.2	3
2211	121.1	3.25

상관관계 분석의 활용

I. 기준금리와 부동산 매매가격



- 기준금리와 아파트 가격의 관계를 분석.

데이터 준비

```
import pandas as pd
import matplotlib.pyplot as plt

realEstate = [83, 83, 83.5, 83.8, 83.9,
                    (중략), 136.1, 133.4, 130.6, 126.2, 121.1]
interestRate = [2.75, 2.75, 2.75, 2.75, 2.5,
                    (중략), 2.25, 2.5, 2.5, 3, 3.25]
data = {'부동산':realEstate, '금리':interestRate}

plt.scatter(df['부동산'], df['금리'])
plt.show()
```

- 데이터를 딕셔너리 data에 할당하고, 키는 '부동산'과 '금리'로 설정.
- '부동산'이 120 초과일 때는 선형 상관도가 낮을 것으로 예상.

상관관계 분석의 활용

I. 기준금리와 부동산 매매가격

- 기준금리와 아파트 가격의 관계를 분석.
  - 월별 부동산 실거래 매매가격 지수와 기준금리 전체 데이터의 상관관계를 피어슨 방식으로 분석.

피어슨 상관분석

```
df = pd.DataFrame(data)
coef = df.corr(method='pearson')
print(coef)
```

```
              부동산        금리
부동산       1.000000  -0.497677
금리         -0.497677   1.000000
```

  - 두 변수의 상관계수 값은 -0.497677로, 절댓값이 0.5에 가까운 음의 선형 상관관계.
  - 기준금리가 오를수록 부동산 가격이 낮아지고, 기준금리가 내릴수록 부동산 가격이 높아진다는 것으로 해석.

상관관계 분석의 활용

I. 기준금리와 부동산 매매가격

- 기준금리와 아파트 가격의 관계를 분석.

부동산 상승기 상관분석

```
originalData = {'부동산':realEstate, '금리':interestRate}

realEstateIndexList = []
interestList = []
lastIndex = -1

#부동산 지수가 143.8이 될 때까지만 리스트에 데이터 추가하기
for key, value in originalData.items():
   if key == '부동산':
      for i in range(0, len(value)):
         if value[i] == 143.8:
           break
         else:
           realEstateIndexList.append(value[i])
           lastIndex = i
   else:
      for i in range(0, lastIndex + 1):
         interestList.append(value[i])
```

상관관계 분석의 활용

I. 기준금리와 부동산 매매가격

- 기준금리와 아파트 가격의 관계를 분석

부동산 상승기 상관분석

```
data = {'지수':realEstateIndexList, '금리':interestList}
df = pd.DataFrame(data)
coef = df.corr(method='pearson')
print(coef)
```

```
            지수        금리
지수    1.000000  -0.854603
금리   -0.854603   1.000000
```

- 피어슨 상관계수는 -0.854603으로 '지수'와 '금리' 두 변수는 강한 음의 선형 상관관계.

상관관계 분석의 활용

II. 영어 성적과 수학 성적

- 학생 10명의 영어 시험과 수학 시험 등수로 스피어만 상관분석을 수행.

학생	학생1	학생2	학생3	학생4	학생5	학생6	학생7	학생8	학생9	학생10
영어 시험 등수	4	2	1	3	10	8	9	7	6	5
수학 시험 등수	2	1	3	4	8	7	10	5	9	6

영어와 수학 등수 상관분석

```
import pandas as pd

data = {'영어':[4, 2, 1, 3, 10, 8, 9, 7, 6, 5],
'수학':[2, 1, 3, 4, 8, 7, 10, 5, 9, 6]}
df = pd.DataFrame(data)
coef = df.corr(method='spearman')
print(coef)
```

	영어	수학
영어	1.000000	0.818182
수학	0.818182	1.000000

- 스피어만 상관계수 0.818182로 두 변수는 양의 상관관계.

실전분석 GDP 성장률과 인구수의 상관관계 분석

[문제]

인구가 많을수록 GDP 성장 잠재력이 높다고 합니다. 실제로 인구수가 세계 1위인 중국과 2위인 인도의 경제 성장률은 2000년 이후 매우 높은 수준을 유지하고 있습니다. 분석 대상을 G20 회원국으로 한정하여 GDP 성장률과 인구수의 상관분석을 수행하세요.

G20는 G7 국가인 미국, 일본, 영국, 프랑스, 독일, 이탈리아, 캐나다를 비롯하여 신흥경제 12개 국가, 그리고 유럽연합(EU)으로 총 20개 국가입니다. 이 국가들 의 GDP 성장률과 인구수 데이터는 [표 9-5]과 같습니다. 인구가 많은 나라가 GDP 성장률도 더 높은 것처럼 보입니다. 그렇다면 인구수와 GDP 성장률이 정말 비례하는지 분석해 보겠습니다.

실전분석 GDP 성장률과 인구수의 상관관계 분석

국가	GDP 성장률 (2022년, %)	인구수 (2021년, 백만 명)
미국	0.9	334
일본	0.6	125
영국	0.4	67.53
프랑스	0.5	67.65
독일	1.1	83.16
이탈리아	1.7	59.24
캐나다	3.9	38.44
대한민국	1.4	51.74
러시아	−3.7	146
중국	2.9	1,412
인도	6.3	1,380
인도네시아	5.01	273
아르헨티나	5.9	45.81
브라질	3.6	213
멕시코	3.5	126
호주	5.9	25.77
남아프리카공화국	4.1	60.14
사우디아라비아	5.4	34.11
터키	3.9	84.68
유럽연합(EU)	1.9	343

GDP 성장률과 인구수의 상관관계 분석

[해결]

1. 데이터를 딕셔너리에 할당하고, 키는 국가, GDP 성장률, 인구수로 함.
   딕셔너리를 데이터프레임으로 변경하여 출력.

```
import pandas as pd

data = {'국가': ['미국', '일본', '영국', '프랑스', '독일', '이탈리아', '캐나다',
              '대한민국', '러시아', '중국', '인도', '인도네시아', '아르헨티나',
              '브라질', '멕시코', '호주', '남아프리카공화국', '사우디아라비아',
              '튀르키예', '유럽연합(EU)'],
        'GDP 성장률': [0.9, 0.6, 0.4, 0.5, 1.1, 1.7, 3.9, 1.4, -3.7, 2.9, 6.3,
                    5.01, 5.9, 3.6, 3.5, 5.9, 4.1, 5.4, 3.9, 1.9],
        '인구수': [334, 125, 67.53, 67.65, 83.16, 59.24, 38.44, 51.74, 146, 1412,
                1380, 273, 45.81, 213, 126, 25.77, 60.14, 34.11, 84.68, 343]}

df = pd.DataFrame(data)
print(df)
```

GDP 성장률과 인구수의 상관관계 분석

[해결]

1. 데이터를 딕셔너리에 할당하고, 키는 국가, GDP 성장률, 인구수로 함.
   딕셔너리를 데이터프레임으로 변경하여 출력.

	국가	GDP 성장률	인구수
0	미국	0.90	334.00
1	일본	0.60	125.00
2	영국	0.40	67.53
3	프랑스	0.50	67.65
4	독일	1.10	83.16
5	이탈리아	1.70	59.24
6	캐나다	3.90	38.44
7	대한민국	1.40	51.74
8	러시아	-3.70	146.00
9	중국	2.90	1412.00
10	인도	6.30	1380.00
11	인도네시아	5.01	273.00
12	아르헨티나	5.90	45.81
13	브라질	3.60	213.00
14	멕시코	3.50	126.00
15	호주	5.90	25.77
16	남아프리카공화국	4.10	60.14
17	사우디아라비아	5.40	34.11
18	튀르키예	3.90	84.68
19	유럽연합(EU)	1.90	343.00

GDP 성장률과 인구수의 상관관계 분석

[해결]

2. 데이터프레임 df로 피어슨, 스피어만, 켄달 상관분석을 모두 수행.

```
pearsonCoef = df.corr(method='pearson')
print("Pearson Correlation Analysis")
print(pearsonCoef)

spearmanCoef = df.corr(method='spearman')
print("\nSpearman Correlation Analysis")
print(spearmanCoef)

kendallCoef = df.corr(method='kendall')
print("\nKendall Correlation Analysis")
print(kendallCoef)
```

GDP 성장률과 인구수의 상관관계 분석

[해결]

2. 데이터프레임 df로 피어슨, 스피어만, 켄달 상관분석을 모두 수행.

```
Pearson Correlation Analysis
            GDP 성장률      인구수
GDP 성장률    1.000000   0.198924
인구수        0.198924   1.000000

Spearman Correlation Analysis
            GDP 성장률      인구수
GDP 성장률    1.000000  -0.196388
인구수       -0.196388   1.000000

Kendall Correlation Analysis
            GDP 성장률      인구수
GDP 성장률    1.000000  -0.137568
인구수       -0.137568   1.000000
```

GDP 성장률과 인구수의 상관관계 분석

[해결]

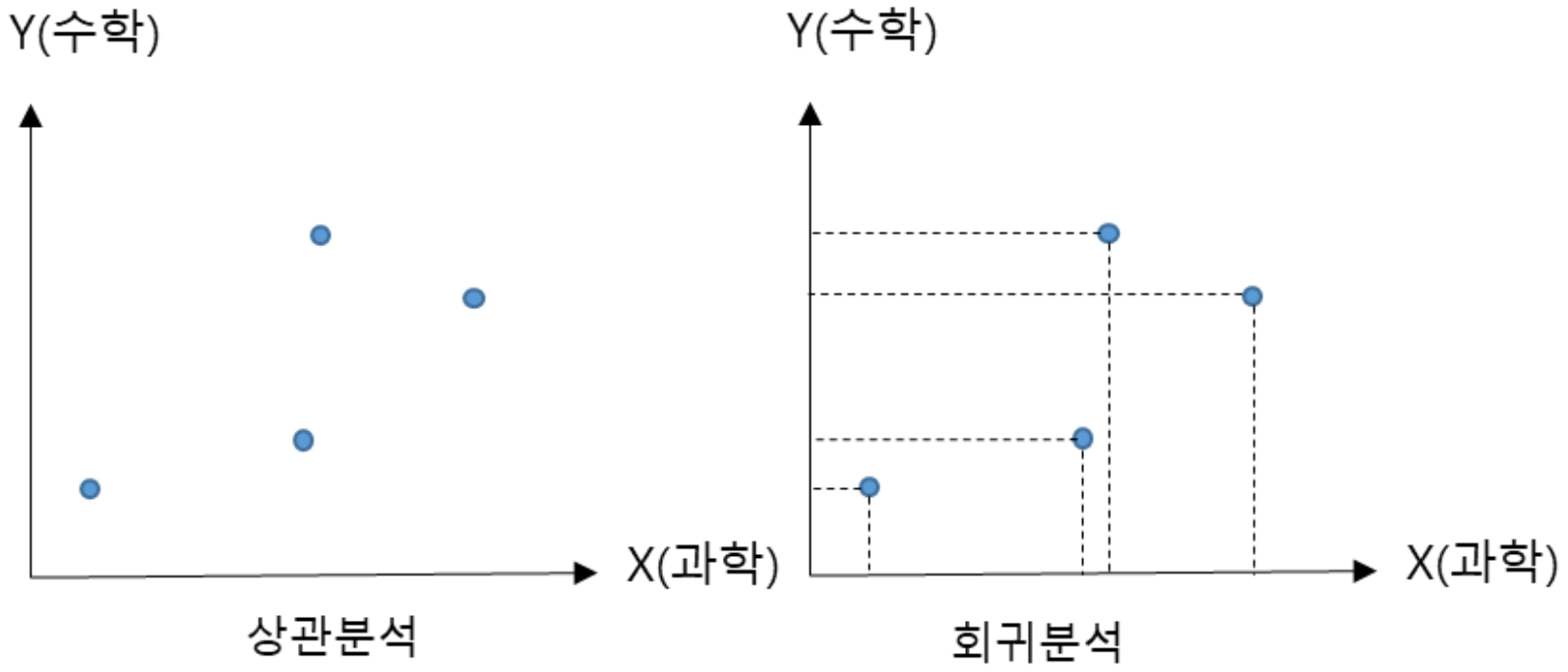
3. 세 가지 상관분석 방법으로 분석한 결과 상관계수의 절댓값이 모두 0.5에 훨씬 못 미침.
   따라서 GDP 상승률과 인구수는 선형 상관관계가 없음. 그러나이 분석에는 두 가지 한계가 있음.

   첫째는 전 세계에 코로나19라는 특수한 요인이 작용한 시기의 전년대비 GDP 성장률이라는 점.
   둘째는 표본 수가 적다는 점.

회귀분석

회귀분석

- 상관분석에서는 두 연속형 변수 X(과학)와 Y(수학)의 상관 정도만 알 수 있고 인과관계는 알 수 없었음
- 회귀분석에서는 두 연속형 변수 X와 Y를 독립변수와 종속변수라고 하는 인과관계로 설명
- '과학 점수가 좋으면 수학점수가 좋을까요?' 와 같이 간단 하지만 미래를 예측할 수 있는 머신러닝의 초기 모델이 됨

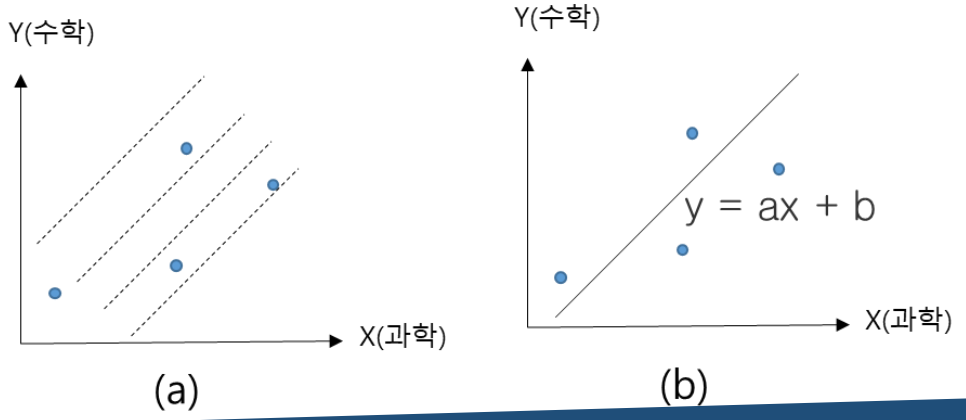


회귀분석

- 선형회귀분석(Linear Regression Analysis)은 쌍으로 관찰된 연속형 변수들 사이의 관계에 있어서 한 변수를 원인으로 하고 다른 변수들을 결과로 하는 분석
- 독립변수와 종속변수 사이 선형식을 구하고 그 식을 이용하여 변수값 들이 주어 졌을 때 종속변수의 변수 값을 예측하는 분석방법

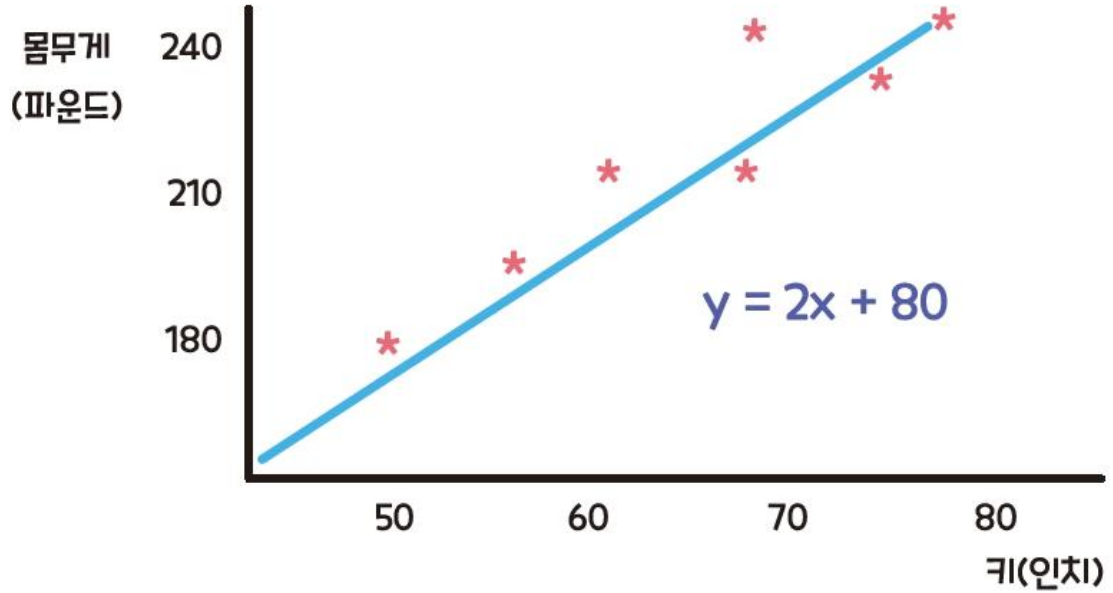
X	Y
독립변수, 설명변수, 원인변수	종속변수, 반응변수, 결과변수 머신러닝(클래스, 라벨)
다른 변수에 영향을 주는 원인	다른 변수에 영향을 받는 결과

- x변수와 y변수 간의 관계를 y = ax + b 와 같은 하나의 선형 관계식으로 표현
- y = ax + b 인 회귀식에서 독립변수 x가 하나인 것이기에 단순선형회귀분석이라함



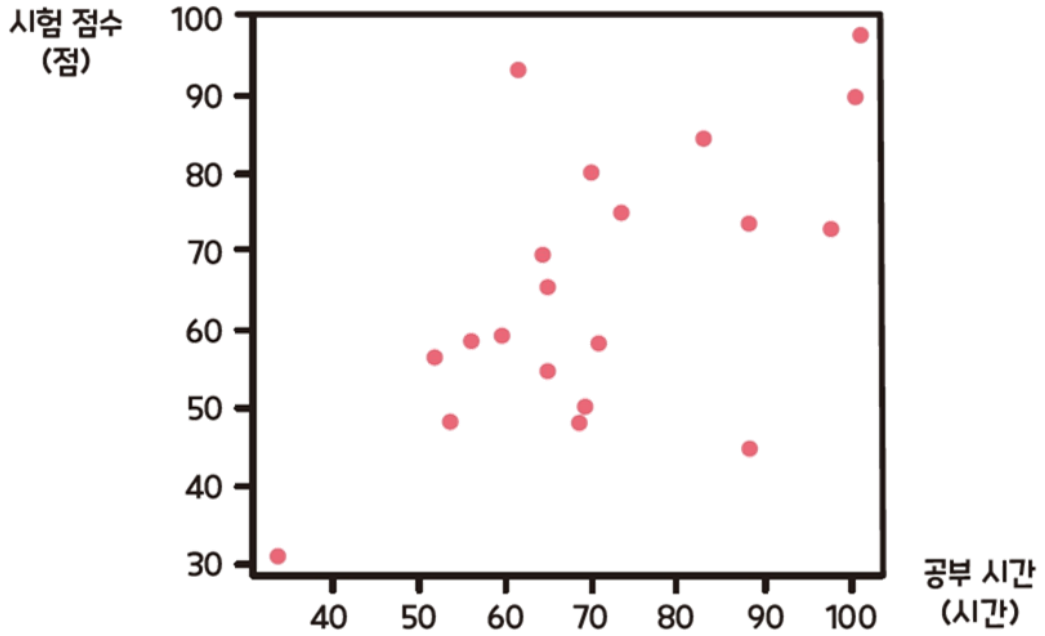
선형 회귀분석의 개념

- 선형 회귀분석의 모형
- 선형 회귀분석(Linear regression analysis) : 두 개 또는 그 이상의 변수 간 인과관계의 패턴을 원래 모습과 가장 가깝게 추정하는 분석 방법.
  - 함수 y = 2x + 80의 그래프, 이를 선형 회귀분석의 모형(Model)이라고 부름.
  - 선형 회귀분석은 x 변수가 원인, y 변수가 결과로 인과관계여야 한다는 조건 있음.



선형 회귀분석의 개념

- 선형 회귀분석의 모형
  - 회귀분석에서 원인인 x 변수는 독립변수(Independent variable), 결과인 y 변수는 종속변수(Dependent variable)라고 부름.
  - 어떠한 결과의 원인이 되는 독립변수가 한 개일 때 단순 선형 회귀분석(Simple linear regression analysis), 두 개 이상이면 다중 선형 회귀분석(Multiple linear regression analysis).
  - 공부 시간과 시험 점수의 인과관계를 분석해보면 종속변수는 시험 점수이며 독립변수는 공부 시간만 고려.



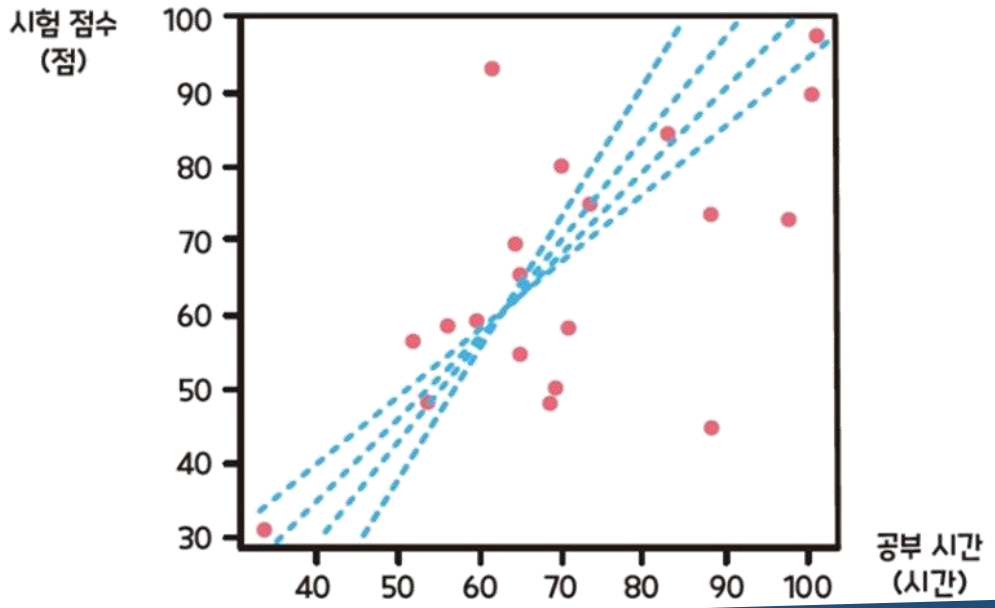
선형 회귀분석의 개념

- 선형 회귀분석의 모형
  - 점들을 직선 하나로 표현하고자 함.
  - 직선을 일차함수 y = mx + b로 표현한다면 계수 m은 이 직선의 기울기이므로 양수일 것
  - 일반적인 선형 회귀모형은 다음 식과 같이 독립변수 x_i앞에 계수 β_i가 붙음.
    마지막 ε은 오차 항

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots\cdots + \varepsilon$$

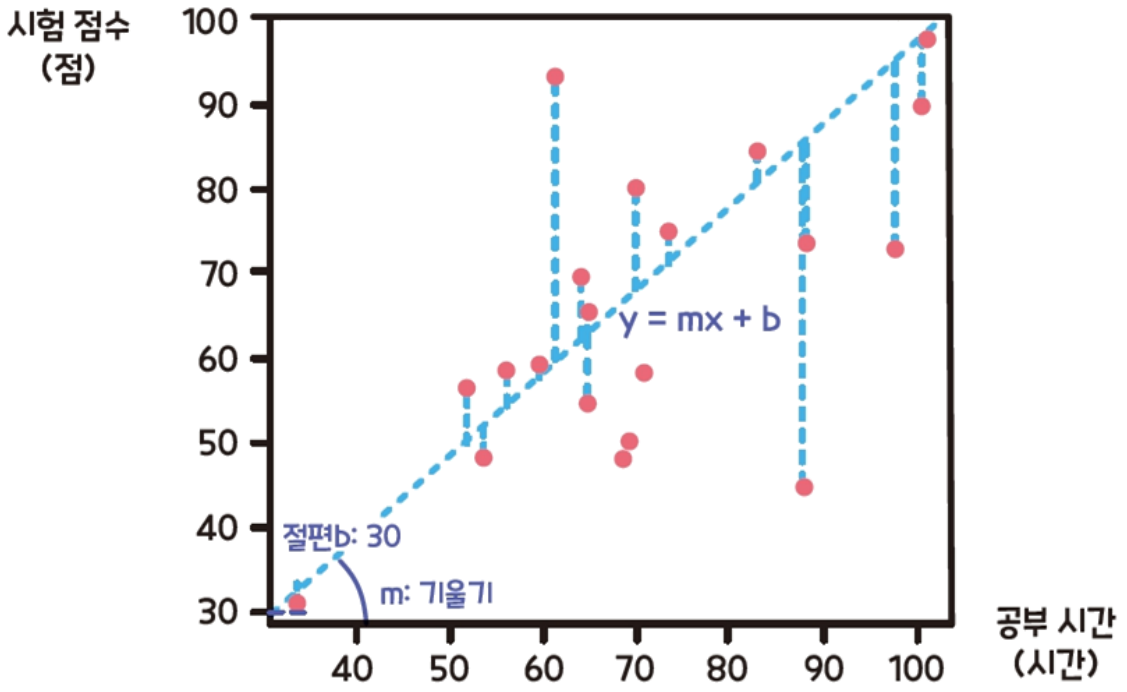
β_i : x_i의 계수 　　오차

  - 데이터를 표현하는 직선은
    여러 개 존재할 수 있음.



선형 회귀분석의 개념

- 선형 회귀분석의 모형
  - 최대한 많은 점과 거리가 가까운 직선이 좋은 직선.
  - 점에서 직선까지 y축과 평행한 선분을 그렸을 때 모든 선분 길이의 합을 최소로 하는 직선 찾는 것.
  - 점이 가리키는 값과 직선이 예측하는 값의 차를 잔차(Residual)라고 부름.



선형 회귀분석의 개념

- 선형 회귀분석의 모형

✓하나 더 알기: 더미 변수

더미 변수(Dummy variable)는 독립변수를 0과 1로 변환하여 '예'와 '아니오'로 나타낼 수 있는 변수.

더미 변수를 여러 개 두면 '예'와 '아니오'만으로 결과를 세 가지 이상으로 구분할 수 있음.

어린이, 청소년, 성인으로 구분.

① 데이터가 어린이일 때 '어린이' 더미 변수는 1이고 '청소년' 더미 변수는 0.

② 청소년일 때 '어린이' 더미 변수는 0이고 '청소년' 더미 변수는 1.

③ 어린이도 청소년도 아닌 성인은 두 더미 변수에 모두 0을 대입하여 표현할 수 있음.

따라서 구분하고자 하는 데이터의 종류가 N개일 때 더미 변수 N-1개를 선언하면 됨.

	어린이 더미 변수	청소년 더미 변수
어린이	1	0
청소년	0	1
성인	0	0

선형 회귀분석의 개념

- 선형 회귀분석의 모형
- 결정계수
  - 선형 회귀분석에서 모형이 데이터의 패턴을 얼마나 효과적으로 보여주는지 수치화한 값
  - 결정계수 R2(R square)를 다음 수식과 같이 정의.

$$R^2 = \frac{(Q - Q_e)}{Q}$$

Q = 전체 데이터의 편차 제곱의 합

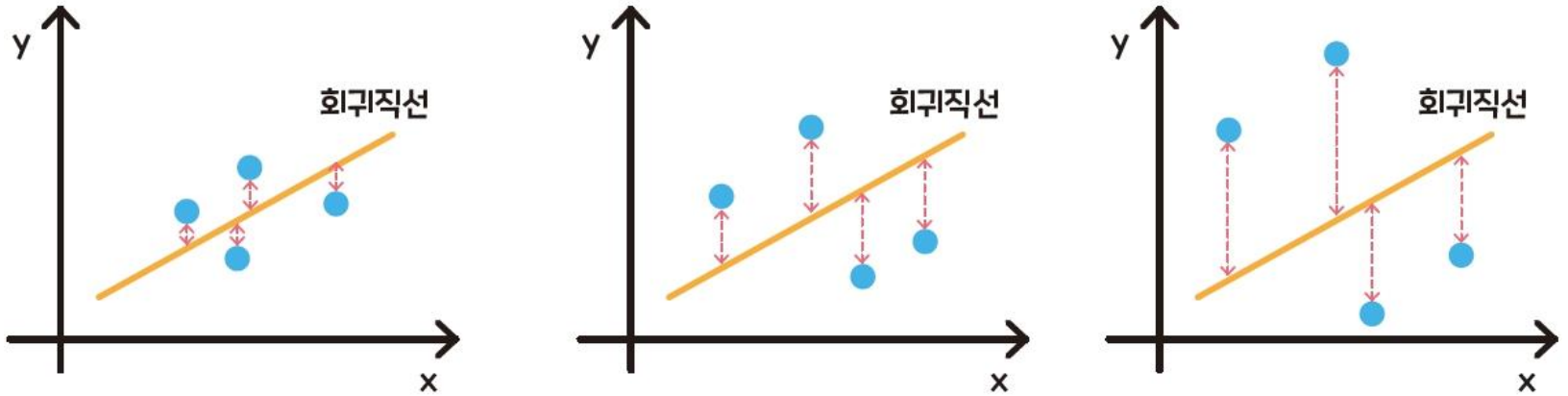
Q_e = 전체 데이터의 잔차 제곱의 합

선형 회귀분석의 개념

● 선형 회귀분석의 모형

● 결정계수

- 편차(Deviation)는 평균과 실제 값의 차이.
- 결정계수 R^2 는 0 이상 1 이하의 값으로 계산됨.
- R^2값이 1에 가까울 때 잔차가 작고 예측의 정밀도가 높음.
  반면 R^2 값이 0에 가까울 때 잔차가 커 예측의 정밀도가 낮음.



(a) R^2가 1에 가까운 경우 (b) R^2가 0.5에 가까운 경우 (c) R^2가 0에 가까운 경우

선형 회귀분석의 개념

- 선형 회귀분석의 모형
  - 수정된 결정계수
    - 결정계수 R^2 는 독립변수의 개수가 많을수록 커지는 경향을 보임.
    - 이러한 문제를 해결하기 위해 다중 선형 회귀분석에서
      수정된 결정계수 adj. R^2 (adjusted R square)로 설명력을 나타냄.

$$adj.R^2 = \frac{(n-1)}{(n-p-1)(1-R^2)}$$

$R^2 =$ 결정계수 $p =$ 독립변수의 개수 $n =$ 표본 수

선형 회귀분석의 개념

- 선형 회귀분석의 해석
- 통계적 가설검정
  - 모집단에 대한 추측을 하고 표본의 정보를 기준으로 가설이 타당한지 판정하는 방법.
  - 통계적 가설에는 두 종류가 있음.

    통계학에서 처음부터 거짓일 것으로 기대하는 가설인 귀무가설(Null hypothesis),

    입증하고자 하는 가설인 대립가설(Alternative hypothesis).
  - 실험 결과를 보고 귀무가설을 채택하거나 대립가설을 채택하는 기준을 세워 두어야 하고, 그 기준을 유의수준(Significance level)이라고 함.
  - 귀무가설이 참일 때 실제 결과가 실험 결과와 같을 확률은 p-값(p-value, 유의확률)이라고 부름.

01. 선형 회귀분석의 개념

II. 선형 회귀분석의 해석

- 선형 회귀분석 과정
  - 선형 회귀분석을 수행하고 해석하는 과정은 크게 세 단계

① 결과인 종속변수를 y로 두고, 원인이 되는 독립변수를 xi로 둠.

② 설명력 R^2또는 adj.R^2값을 확인. 결정계수가 0.6 또는 0.4 이상이면 해당 회귀모형이 설명력을 갖추었다고 인정.

③ 각 독립변수의 p-값이 유의수준보다 작은지 확인. p-값이 유의수준 이상인 변수를 제외하고 남은 독립변수가 결과에 영향을 주는 원인.

선형 회귀분석의 활용

● 연봉과 직장 만족도

● 연봉이 직장 만족도와 얼마나 관계 있는지 파악할 수 있음.

직장 만족도(점)	60	75	70	85	90	70	65	95	70	80
연봉(만 원)	3,000	4,200	4,000	5,000	6,000	3,800	3,500	6,200	3,900	4,500

▪ 원인인 독립변수는 연봉이며 결과인 종속변수는 직장 만족도.

산점도 확인

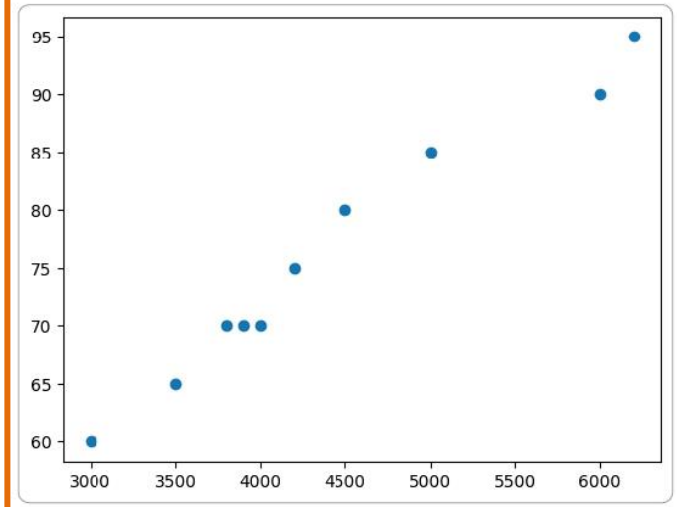
```
import pandas as pd
import matplotlib.pyplot as plt

x = [3000, 4200, 4000, 5000, 6000, 3800, 3500, 6200, 3900, 4500]
y = [60, 75, 70, 85, 90, 70, 65, 95, 70, 80]
data = {'x': x, 'y': y}
df = pd.DataFrame(data)
plt.scatter(df['x'], df['y'])
plt.show( )
```

선형 회귀분석의 활용

● 연봉과 직장 만족도

실행결과



- 리스트 x에 독립변수인 연봉을 할당, 리스트 y에 종속변수인 직장 만족도를 할당.
- 맷플롯립의 scatter( ) 함수로 산점도를 그리면 산점도는 우상향 형태로 나타남.

선형 회귀분석의 활용

- 연봉과 직장 만족도
  - 산점도를 보면 두 변수에 양의 상관관계가 있음을 알 수 있는데,
    인과관계가 얼마나 강한지 선형 회귀분석으로 확인 하려함.
  - 선형 회귀모형은 다음과 같은 형태로 설정.

```
종속변수 ~ 독립변수1 + 독립변수2 + 독립변수3 + ···
```

단순 선형 회귀분석

```
from statsmodels.formula.api import ols
from sklearn.linear_model import LinearRegression

fit = ols('y ~ x', data=df).fit( )
print(fit.summary( ))
```

  - ols( ) 함수는 선형 회귀분석을 수행하는 함수.
    첫 번째 인자로 회귀모형 'y ~ x', 두 번째 인자로 데이터를 입력.

선형 회귀분석의 활용

- 연봉과 직장 만족도

실행결과

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y R-squared:                       0.971
Model:                            OLS Adj. R-squared:                  0.968
Method:                 Least Squares F-statistic:                     271.0
Date:                Mon, 27 Mar 2023 Prob (F-statistic):           1.87e-07
Time:                        07:13:44 Log-Likelihood:                -20.111
No. Observations:                  10 AIC:                             44.22
Df Residuals:                       8 BIC:                             44.83
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     29.0004      2.926      9.913      0.000      22.254      35.747
x              0.0107      0.001     16.463      0.000       0.009       0.012
==============================================================================
Omnibus:                        0.346 Durbin-Watson:                   2.871
Prob(Omnibus):                  0.841 Jarque-Bera (JB):                0.447
Skew:                           0.286 Prob(JB):                        0.800
Kurtosis:                       2.136 Cond. No.                     2.07e+04
==============================================================================
```

선형 회귀분석의 활용

- 연봉과 직장 만족도
  - 첫째, 결정계수 R^2값이 0.971이므로 표본 데이터들에 대한 설명력이 97.1%이고 결정계수가 0.6을 크게 초과하므로 모형의 정밀도가 높음.
  - 둘째, 회귀모형의 독립변수 x의 유의수준은 0.000으로 0.05 미만이므로 x는 유의한 독립변수.
  - 셋째, 독립변수 x의 계수는 0.0107.

$$y = 29.0004 + (0.0107) \times x + \varepsilon$$

선형 회귀분석의 활용

- 직장 만족도의 요인 분석
  - 연봉 외에 일평균 휴식시간(분)과 일평균 근무시간(시간) 추가하여 각 변수의 영향을 분석.

직장 만족도(점)	60	75	70	85	90	70	65	95	70	80
연봉(만 원)	3,000	4,200	4,000	5,000	6,000	3,800	3,500	6,200	3,900	4,500
일평균 휴식시간(분)	120	60	100	100	50	120	90	40	120	120
일평균 근무시간(시간)	8	6	10	8	10	10	9	7	8	9

  - 결과인 직장 만족도가 종속변수이며, 나머지는 독립변수.
  - 회귀모형 companySatisfaction ~ salary + breakTime + workingTime을 입력하여 다중 선형 회귀분석을 수행.

선형 회귀분석의 활용

- 직장 만족도의 요인 분석

다중 선형 회귀분석

```
from statsmodels.formula.api import ols
from sklearn.linear_model import LinearRegression

salary = [3000, 4200, 4000, 5000, 6000, 3800, 3500, 6200, 3900, 4500]
breakTime = [120, 60, 100, 100, 50, 120, 90, 40, 120, 120]
workingTime = [8, 6, 10, 8, 10, 10, 9, 7, 8, 9]
companySatisfaction = [60, 75, 70, 85, 90, 70, 65, 95, 70, 80]
data = {'salary': salary, 'breakTime': breakTime, 'workingTime': workingTime,
        'companySatisfaction': companySatisfaction}
df = pd.DataFrame(data)

fit = ols('companySatisfaction ~ salary + breakTime + workingTime',
data=df).fit( )
print(fit.summary( ))
```

선형 회귀분석의 활용

- 직장 만족도의 요인 분석

실행결과

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     companySatisfaction  R-squared: 0.988
Model:                             OLS  Adj. R-squared: 0.982
Method:                  Least Squares    F-statistic: 164.0
Date:                 Mon, 27 Mar 2023    Prob (F-statistic): 3.81e-06
Time:                         07:30:46             Log-Likelihood: -15.777
No. Observations:                   10             AIC: 39.55
Df Residuals:                        6             BIC: 40.77
Df Model:                            3
Covariance Type:             nonrobust
==============================================================================
                 coef    std err          t      P>|t|           [0.025 0.975]
------------------------------------------------------------------------------
Intercept     24.9819      5.353       4.667      0.003           11.884 38.080
salary        0.0120       0.001      15.895      0.000            0.010 0.014
breakTime     0.0668       0.027       2.491      0.047            0.001 0.132
workingTime    -0.9718      0.412     -2.356      0.057           -1.981 0.037
==============================================================================
Omnibus:                        0.929 Durbin-Watson:      2.500
Prob(Omnibus):                  0.628 Jarque-Bera (JB):   0.752
Skew:                          -0.441 Prob(JB):           0.686
Kurtosis:                       1.986 Cond. No.           5.06e+04
==============================================================================
```

선형 회귀분석의 활용

- 직장 만족도의 요인 분석
  - 첫째, 수정된 결정계수 $adj.R^2$는 0.988이므로 이 선형 회귀모형의 설명력이 98.8%
  - 둘째, 변수 salary는 p-값이 0.000이고 변수 breakTime은 0.047로 0.05 보다 작고 workingTime의 p-값은 0.057로 0.05 보다 큼.
    따라서 연봉과 휴식시간은 유의한 독립변수이고, 근무시간은 유의하지 않은 독립변수.
  - 셋째, 유의수준 결과에 따라서 salary와 breakTime을 독립변수로 하는 모형이 구성됨.

$$companySatisfaction$$

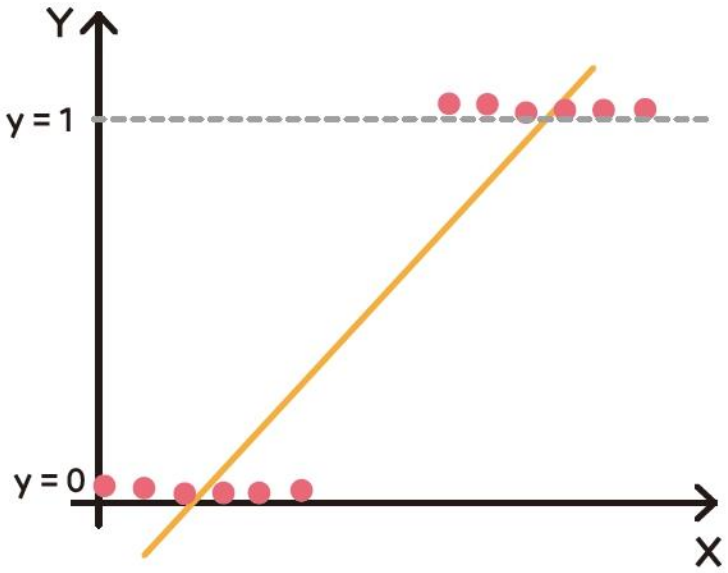
$$= 24.9819 + (0.0120) \times salary + (0.0668) \times breakTime + \varepsilon$$

로지스틱 회귀분석의 개념

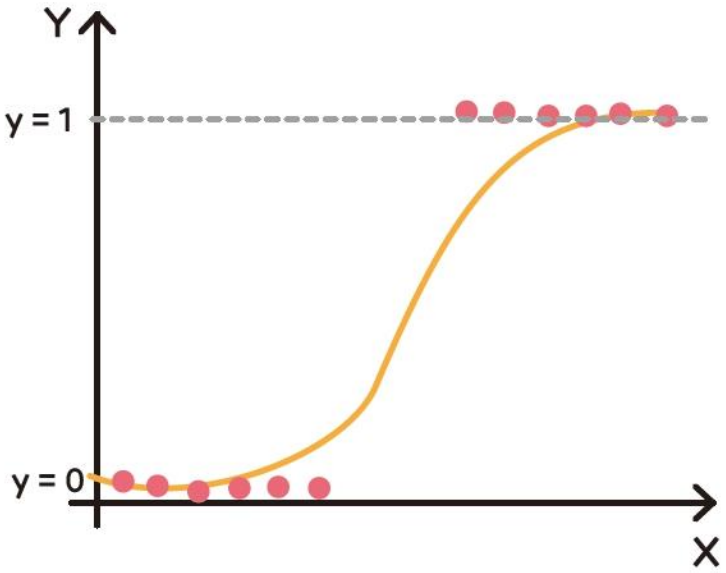
- 로지스틱 회귀모형
  - 로지스틱 회귀분석(Logistic regression analysis)은 결과인 종속변수에 미치는 요인들을 독립변수로 두고 각 독립변수의 영향을 설명.
  - 로지스틱 회귀분석의 종속변수는 범위에 제한이 있음.
    로지스틱 회귀분석의 종속변수는 0에서 1사이의 값.

로지스틱 회귀분석의 개념

- 로지스틱 회귀모형
  - 선형 회귀분석과 로지스틱 회귀분석의 모형.
  - 왼쪽의 선형 회귀분석은 표본 데이터를 직선으로 그룹화.
  - 반면 오른쪽의 로지스틱 회귀분석은 표본 데이터가 0에서 1사이의 값으로 그룹화되어 있음.



(a) 선형 회귀분석



(b) 로지스틱 회귀분석

로지스틱 회귀분석의 개념

- 로지스틱 회귀모형
  - 로지스틱 회귀분석 결과 오즈비를 얻음. 오즈비(Odds Ratio, OR)는 우리말로 승산비.
  - 사건이 발생할 확률을 p라고 할 때, 오즈비를 수식으로 나타내면 다음과 같음.

$$\text{OR} = \frac{\text{사건이 발생할 확률}}{\text{사건이 발생하지 않을 확률}} = \frac{p}{1-p}$$

  - 대학 합격이라는 사건에서 합격을 1, 불합격을 0으로 정했을 때 합격할 확률이 0.8이라면 오즈비는 4. 이는 대학에 합격할 확률이 불합격할 확률보다 4배 높다는 뜻.

$$\text{OR} = \frac{p}{1-p} = \frac{0.8}{1-0.8} = 4$$

로지스틱 회귀분석의 개념

● 로지스틱 회귀분석의 해석

● 로지스틱 회귀분석의 해석 과정은 다음과 같음.

- 첫째, 선형 회귀분석과 마찬가지로 각 독립변수의 p-값을 확인함. p-값이 유의수준보다 작은 독립변수를 통계적으로 유의한 변수라고 판단.
- 둘째, 오즈비를 구해 각 독립변수가 종속변수를 1로 만들 확률을 비교.

● 사과 가격이 사과 판매 여부에 미치는 영향을 분석해보기. 종속변수는 사과 판매 여부로, 사과가 판매되면 1이고 판매되지 않으면 0. 독립변수는 사과 가격.

사과 판매 여부	1	1	1	1	1	1	1	1	1
가격(원)	1,500	2,000	5,000	3,000	3,500	2,500	4,000	4,500	3,000
사과 판매 여부	0	0	0	0	0	0	0	0	
가격(원)	4,500	4,000	4,500	5,500	6,500	5,000	3,500	7,000	

로지스틱 회귀분석의 개념

- 로지스틱 회귀분석의 해석

로지스틱 회귀분석

```
import statsmodels.api as sm
import pandas as pd
import numpy as np

sales = [1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0]
price = [1500, 2000, 5000, 3000, 3500, 2500, 4000, 4500, 3000,\
         4500, 4000, 4500, 5500, 6500, 5000, 3500, 7000]
data = {'sales': sales, 'price': price}
df = pd.DataFrame(data)

logis = sm.Logit.from_formula('sales ~ price', data=df).fit( )
print(logis.summary( ))
print('OR')
print(np.exp(logis.params)))
```

로지스틱 회귀분석의 개념

- 로지스틱 회귀분석의 해석

실행결과

```
Optimization terminated successfully.
        Current function value: 0.430873
        Iterations 7
                          Logit Regression Results
==============================================================================
Dep. Variable:          sales    No. Observations:                       17
Model:                  Logit    Df Residuals:                           15
Method:                   MLE    Df Model:                                1
Date:        Mon, 03 Apr 2023    Pseudo R-squ.:                      0.3768
Time:                00:48:37    Log-Likelihood:                    -7.3248
converged:               True    LL-Null:                          -11.754
Covariance Type:       nonrobust    LLR p-value:                    0.002917
==============================================================================
              coef     std err       z     P>|z|         [0.025 0.975]
------------------------------------------------------------------------------
Intercept    6.5752      3.300    1.993    0.046          0.108 13.042
price       -0.0016      0.001   -2.008    0.045         -0.003 3.75e-05
==============================================================================
OR
Intercept 717.058841
price 0.998433
dtype: float64
```

로지스틱 회귀분석의 개념

- 로지스틱 회귀분석의 해석

  - 우선 각 독립변수가 유의한지 확인.
  - 독립변수 price의 p-값은 0.04로 0.05 미만이므로 유의한 변수.
    y절편인 Intercept의 p-값 0.046도 확인할 수 있음.
  - 변수 price의 오즈비가 0.998433이므로 가격을 올렸을 때 판매될 가능성이 판매되지 않을 가능성의 0.998433배.

로지스틱 회귀분석의 활용

- 타이타닉 탑승자 생존여부 예측
  - 타이타닉호 데이터에서 생존과 사망의 요인을 분석해보기.

타이타닉 탑승자 데이터

```
import seaborn as sns

titanic = sns.load_dataset('titanic')
print(titanic)
```

```
     survived pclass     sex    age   sibsp parch     fare  embarked class \
0           0      3    male   22.0       1      0   7.2500         S Third
1           1      1  female   38.0       1      0  71.2833         C First
2           1      3  female   26.0       0      0   7.9250         S Third
3           1      1  female   35.0       1      0  53.1000         S First
4           0      3    male   35.0       0      0   8.0500         S Third
..        ...    ...     ...    ...     ...    ...      ...           ... ...
886         0      2    male   27.0       0      0  13.0000         S Second
887         1      1  female   19.0       0      0  30.0000         S First
888         0      3  female    NaN       1      2  23.4500         S Third
889         1      1    male   26.0       0      0  30.0000         C First
890         0      3    male   32.0       0      0   7.7500         Q Third
```

로지스틱 회귀분석의 활용

● 타이타닉 탑승자 생존여부 예측

실행결과(계속)

```
            who    adult_male   deck   embark_town  alive alone
0           man          True   NaN    Southampton     no False
1         woman         False     C    Cherbourg      yes False
2         woman         False   NaN    Southampton    yes True
3         woman         False     C    Southampton    yes False
4           man          True   NaN    Southampton     no True
..          ...           ...    ...           ...     ... ...
886         man          True   NaN   Southampton     no True
887       woman         False     B   Southampton    yes True
888       woman         False   NaN   Southampton     no False
889         man          True     C   Cherbourg      yes True
890         man          True   NaN   Queenstown      no True
[891 rows x 15 columns]
```

로지스틱 회귀분석의 활용

- 타이타닉 탑승자 생존여부 예측

타이타닉 탑승자 데이터의 로지스틱 회귀분석

```
import statsmodels.api as sm
import numpy as np
from sklearn.preprocessing import LabelEncoder

encoder = LabelEncoder( )
encoder.fit(titanic['sex'])
sex = encoder.transform(titanic['sex'])
titanic['sex'] = sex

model = sm.Logit.from_formula('survived ~ pclass + sex + age + fare + parch + sibsp', data=titanic)
logit = model.fit( )
print(logit.summary( ))

print("OR")
print(np.exp(logit.params))
```

로지스틱 회귀분석의 활용

- 타이타닉 탑승자 생존여부 예측

실행 결과

```
Optimization terminated successfully.
             Current function value: 0.445244
             Iterations 6

                           Logit Regression Results
==============================================================================
Dep. Variable:            survived     No. Observations:                    714
Model:                       Logit     Df Residuals:                        707
Method:                       MLE       Df Model:                            6
Date:          Mon, 03 Apr 2023       Pseudo R-squ.:                  0.3408
Time:                  00:59:54       Log-Likelihood:                -317.90
converged:                    True     LL-Null:                       -482.26
Covariance Type:       nonrobust      LLR p-value:                 5.727e-68
==============================================================================
              coef        std err        z        P>|z|         [0.025 0.975]
--------------------- ------------------------------------------------------------
Intercept     5.3890        0.604      8.926      0.000          4.206 6.572
pclass       -1.2422        0.163     -7.612      0.000                  -1.562 -0.922
sex          -2.6348        0.220    -11.998      0.000                  -3.065 -2.204
age          -0.0440        0.008     -5.374      0.000                  -0.060 -0.028
fare          0.0022        0.002      0.866      0.386                  -0.003 0.007
parch        -0.0619        0.123     -0.504      0.614                  -0.303 0.179
sibsp        -0.3758      0.127      -2.950      0.003                  -0.625 -0.126
==============================================================================
```

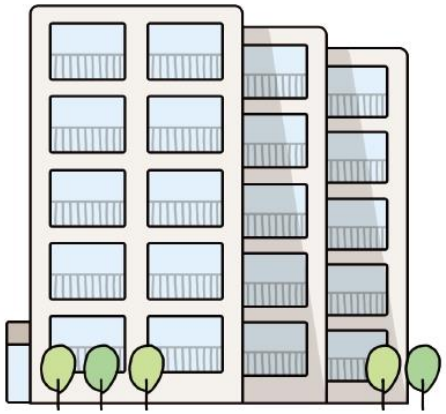
로지스틱 회귀분석의 활용

- 타이타닉 탑승자 생존여부 예측
  - 첫째, 독립변수 중 p-값이 0.05 미만인 변수는 pclass, sex, age, sibsp.
  - 이들 변수의 계수가 모두 음수이므로
    독립변수의 값이 증가할 때 생존 가능성이 낮아지는 것으로 판단.
  - 둘째, 유의한 독립변수 중 age의 오즈비가 가장 크고, sex의 오즈비가 가장 작음.

아파트 매매가격의 요인 분석

[문제]

아파트 매매가격은 변동하는 값입니다. 어떤 요인이 매매가격에 얼마나 영향을 미치는지 알고 싶습니다. 이럴 때 선형 회귀분석을 수행하여 인과관계를 알아낼 수 있습니다.



[해결]

1. 가장 먼저 종속변수와 독립변수를 설정. 아파트 매매가격을 종속변수로 하고, 이에 영향을 미치는 예상 요인들을 독립변수로 함. 아파트 매매가격에 영향을 주는 것은 면적**(size)**, 아파트가 얼마나 오래되었는지**(age)**, 주변 편의시설일 것으로 예상됨.

아파트 매매가격의 요인 분석

[해결]

1. 표는 아파트 매매가격과 앞에서 설정한 독립변수.

price	size	age	kindergarten	elementarySchool	busStop	hospital	mart
174,000	152	19	22	10	13	19	19
156,500	118	19	22	10	13	19	19
168,000	118	19	22	10	13	19	19
145,000	85	19	22	10	13	19	19
...	...	...	...	...	...	...	...
100,000	59	11	4	12	29	14	14
139,500	128	11	4	12	29	14	14
160,500	128	11	4	12	29	14	14
150,000	115	11	4	12	29	14	14

아파트 매매가격의 요인 분석

[해결]

2. 선형 회귀모형을 다음과 같이 설정하여 분석을 수행.

```
price ~ size + age + kindergarten + elementarySchool + busStop + hospital + mart
```

```
import pandas as pd
import matplotlib.pylab as plt
from statsmodels.formula.api import ols
from sklearn.linear_model import LinearRegression

price = [174000, 156500, 168000, 145000, (중략), 100000, 139500, 160500, 150000]
size = [152, 118, 118, 85, (중략), 59, 128, 128, 115]
age = [19, 19, 19, 19, (중략), 11, 11, 11, 11]
kindergarten = [22, 22, 22, 22, (중략) , 4, 4, 4, 4]
elementarySchool = [10, 10, 10, 10, (중략) , 12, 12, 12, 12]
busStop = [13, 13, 13, 13, (중략) , 29, 29, 29, 29]
hospital = [19, 19, 19, 19, (중략) , 14, 14, 14, 14]
mart = [19, 19, 19, 19, (중략) , 14, 14, 14, 14]
```

아파트 매매가격의 요인 분석

[해결]

3. 각 변수 리스트를 데이터프레임으로 변환하고 선형 회귀분석을 수행.

```
data = {'price': price, 'size': size, 'age': age, 'kindergarten':
kindergarten, 'elementarySchool': elementarySchool, 'busStop': busStop,
'hospital': hospital, 'mart': mart}
df = pd.DataFrame(data)

fit = ols('price ~ size + age + kindergarten + elementarySchool + busStop
+ hospital + mart', data=df).fit( )
print(fit.summary( )))
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.876
Model:                            OLS   Adj. R-squared:                  0.862
Method:                 Least Squares   F-statistic:                     62.45
Date:                Mon, 03 Apr 2023   Prob (F-statistic):           1.11e-25
Time:                        02:26:04   Log-Likelihood:                -734.71
No. Observations:                  70   AIC:                             1485.
Df Residuals:                      62   BIC:                             1503.
Df Model:                           7
Covariance Type:            nonrobust
```

아파트 매매가격의 요인 분석

[해결]

3. 각 변수 리스트를 데이터프레임으로 변환하고 선형 회귀분석을 수행.

```
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      1.169e+05   1.23e+05      0.948      0.347    -1.3e+05    3.64e+05
size             534.9026     43.081     12.416      0.000     448.785     621.021
age            -1460.9677   1754.535     -0.833      0.408   -4968.233    2046.298
kindergarten    1927.0880    591.638      3.257      0.002     744.421    3109.755
elementarySchool -1599.1185   3858.456     -0.414      0.680   -9312.062    6113.825
busStop          -13.2131    730.790     -0.018      0.986   -1474.042    1447.616
hospital         737.2488    891.948      0.827      0.412   -1045.730    2520.227
mart           -1372.4907   3583.901     -0.383      0.703   -8536.606    5791.625
==============================================================================
Omnibus:                        4.208   Durbin-Watson:     2.150
Prob(Omnibus):                  0.122   Jarque-Bera (JB):  3.332
Skew:                          -0.446   Prob(JB):          0.189
Kurtosis:                       3.589   Cond. No.          1.17e+04
==============================================================================
```

아파트 매매가격의 요인 분석

[해결]

3. 각 변수 리스트를 데이터프레임으로 변환하고 선형 회귀분석을 수행.

수정된 결정계수 adj.R^2가 0.862이므로 이 모형은 86.2%의 설명력을 갖췄음.

유의한 변수는 p-값이 0.000인 size와 0.002인 kindergarte뿐.

변수 size의 계수는 534.9026이며, kindergarten의 계수는 1927.0880.

$$price = 0.0000169 + (534.9026) \times size + (1927.0880) \times kindergarten + \varepsilon$$

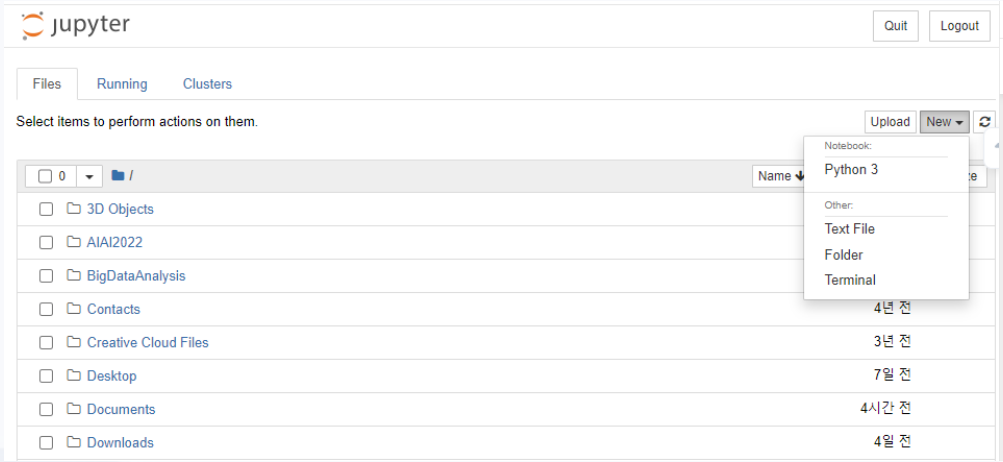
4. 모형 해석:

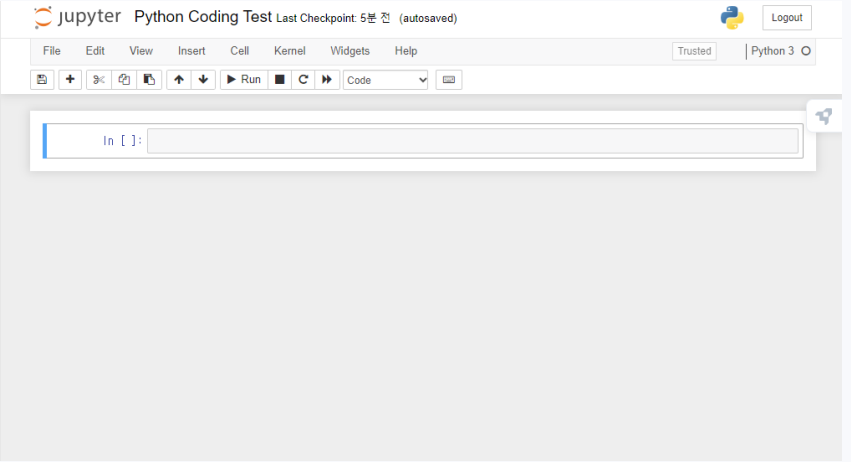
아파트 매매가격에 영향을 미치는 요인은 면적과 유치원까지의 거리. 그러나 특정 지점까지 도보 소요시간이 길어질수록 매매가격이 비싸진다는 해석은 일반적이지 않음.

따라서 데이터의 시간적, 공간적 범위를 넓히고 독립변수도 추가하여 다시 분석할 필요가 있음.

학습활동: Python 코딩 실습

Python 코드 소스 파일 작성 실습





Thank you!

See you next time.

담당교수 : 유 현 주

comjoo@uok.ac.kr