2023 D&A

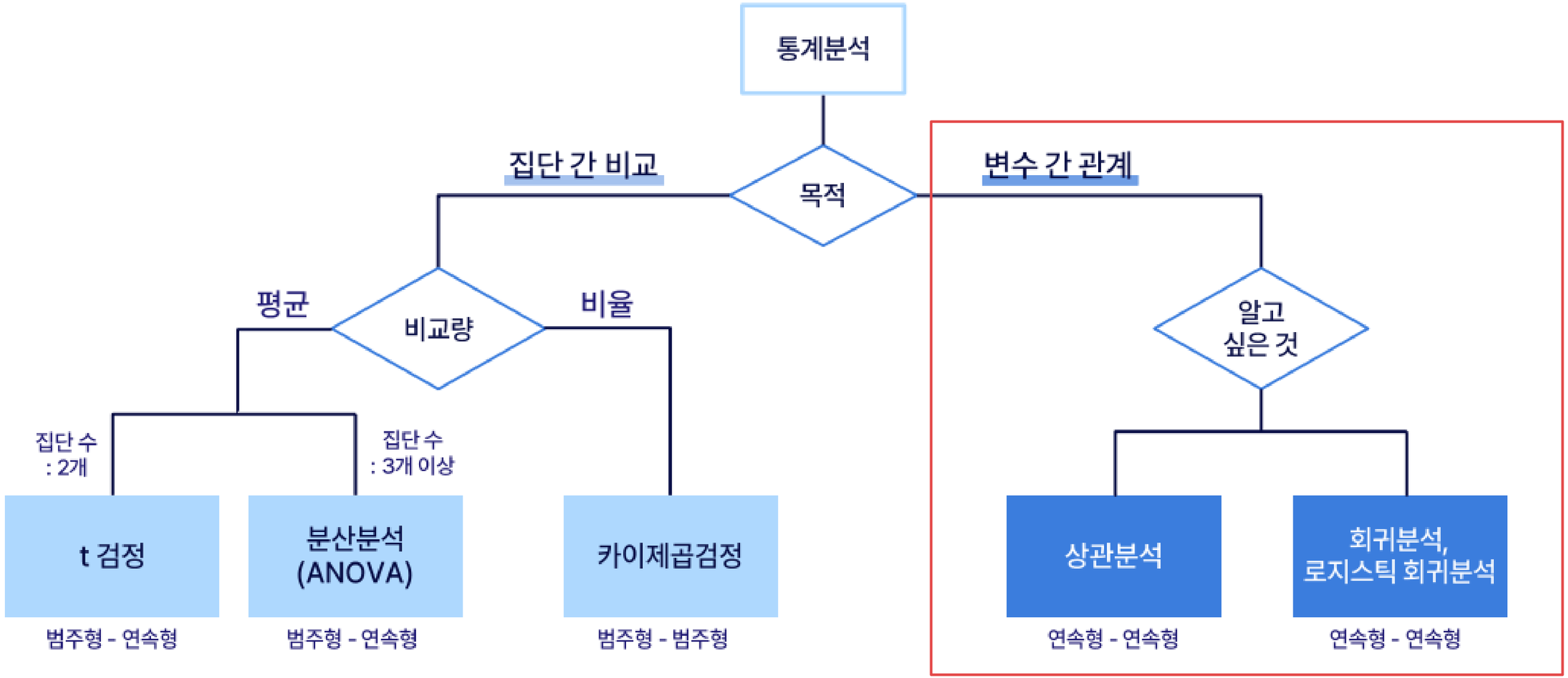
Basic Session 8차시

통계분석 II

CONTENTS

Today

통계분석 II – 상관분석, 회귀분석

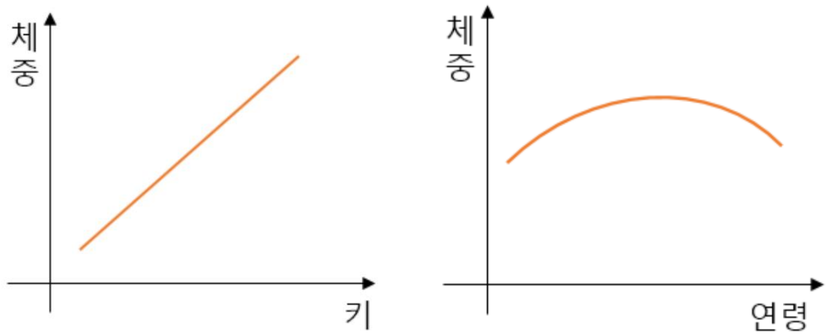


상관분석

상관분석의 기본 개념

상관분석이란?

연속형 변수로 측정된 두 변수 간의 선형적 관계를 분석하는 기법
Ex) 키와 몸무게의 상관관계 분석 => 높은 상관관계
Ex) 체중과 연령의 상관관계 분석 => 낮은 상관관계



* 연속형 변수 vs 범주형 변수
- 연속형 변수 ex) 온도, 키, 체중, 나이 등
- 범주형 변수 ex) 성별(남/여), 직종(서비스직,전문직,기술직) 등

상관분석의 사용 이유?

회귀분석 등에 사용되는 변수들 간의 상관관계를 분석하여
분석대상 변수들이 적절하게 선정되었는지를 제시해주기 위해 사용

+ 두 변수에 대해 상관분석을 하였더니 두 변수가 동일하거나 매우 높은 수준의 상관관계를 지닌 경우,
⇒변수 선택이 잘못됨! 이라고 파악 가능
⇒두 변수가 거의 동일한 변수라는 것은 그 변수만의 효과를 알아내기가 어려워 특정변수의 유의성이 상실됨

상관분석
공분산과 상관계수

공분산

두 변수 간의 관계를 나타내는 통계적 지표

$$cov(x, y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

- 측정 단위에 영향을 받음(kg vs g, km vs mile)

Ex) 100점짜리 두 과목의 공분산은 낮지만(상관성 낮음) 100점만점이기 때문에 큰 값 도출
Ex) 10점짜리 두 과목의 공분산은 높지만(상관성 높음) 10점 만점이기 때문에 작은 값 도출

상관계수

표준편차로 나눠서 -1부터 1사이의 값을 갖도록 함(단위화)

$$cor(x, y) = \frac{cov(x, y)}{sd(x)sd(y)}$$

상관분석

상관계수의 기본 개념

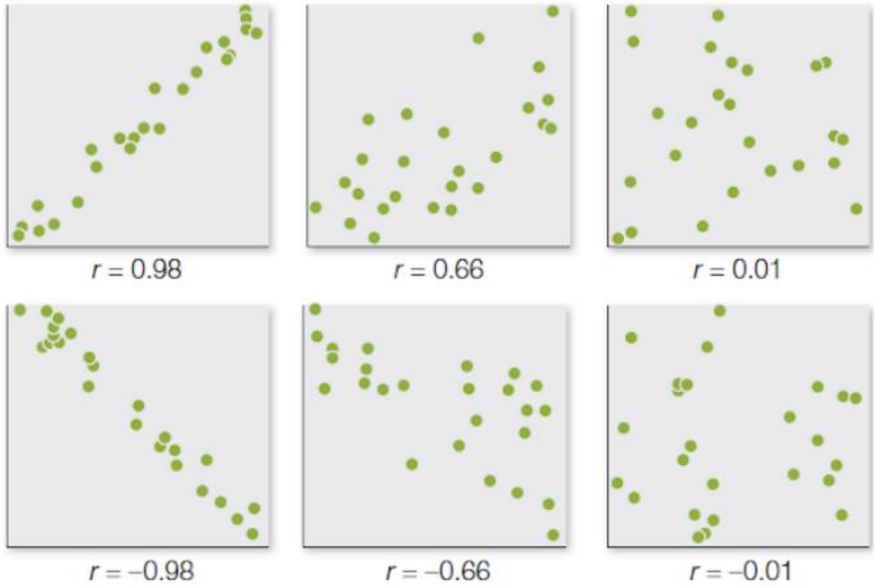
상관계수란?

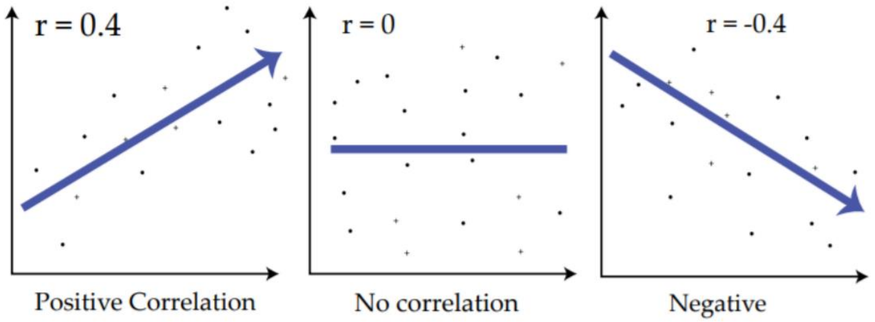
두 변수 사이의 선형적인 관계 정도를 수치적으로 나타낸 계수

- 상관계수가 음수면 음의 상관, 양수이면 양의 상관을 가짐
- 상관계수가 0일 경우, 선형의 상관관계가 아님(무상관)
- -1 ~1 사이

Ex) 키와 몸무게의 상관관계
⇒키가 커지면 몸무게 또한 대체적으로 증가
⇒양의 상관관계를 지닌다고 표현

Ex) 담배와 기대 수명의 상관관계
⇒담배를 많이 피울 경우, 기대 수명이 줄어든다
⇒음의 상관관계를 지닌다고 표현





상관분석

상관계수의 종류

피어슨(Pearson) 상관계수

X,Y 간의 선형 상관 관계를 계량화한 수치

범위: -1 ~ 1 사이의 값

참고) 피어슨 상관계수가 0일 경우, 선형 상관관계가 없다는 것을 의미함.
=> 비선형 관계는 존재할 수도 있음
=> 0이라고 상관관계가 없는 것이 아님!

$$r_{XY} = \frac{\sum_i^n \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_i^n \left(X_i - \overline{X}\right)^2}\sqrt{\sum_i^n \left(Y_i - \overline{Y}\right)^2}}$$

스피어만(Spearman) 상관계수

값에 순위를 매겨 순위에 대해 상관계수를 구하는 방법(서열척도) <-> 피어슨(연속변수)

Ex)
- 수학 과목의 석차와 영어 과목의 석차의 상관관계 -> 스피어만 상관계수
- 수학 점수와 영어 점수의 상관관계 -> 피어슨 상관계수

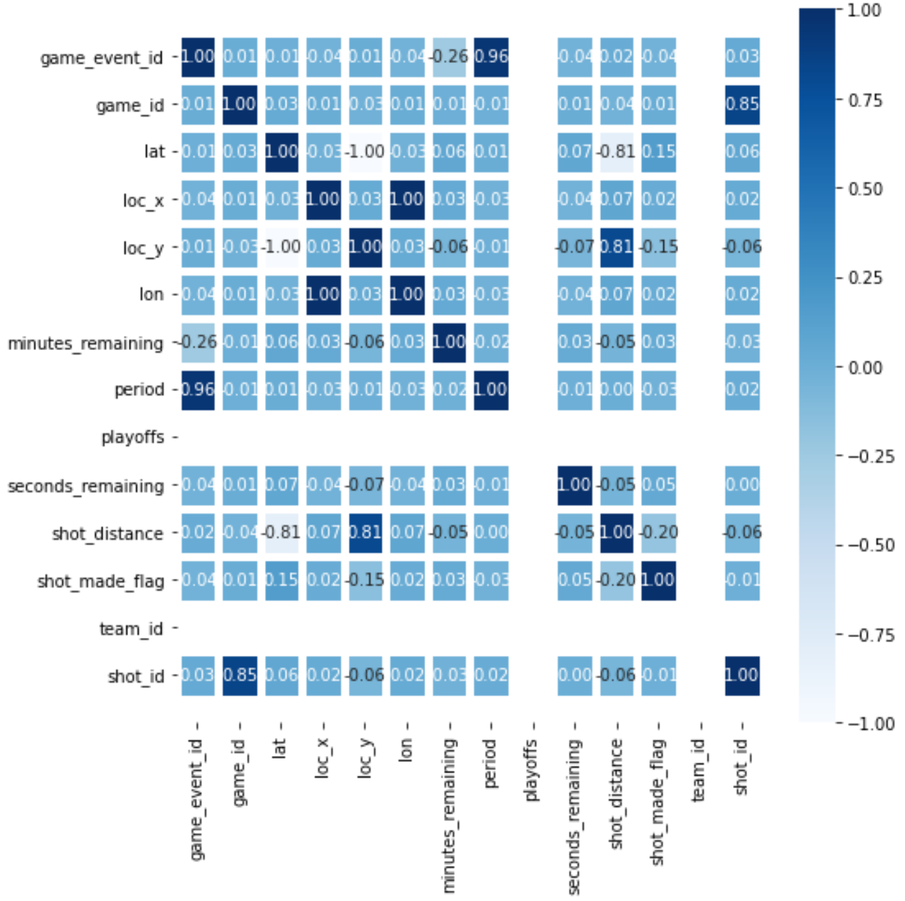
$$p = \frac{6\sum d_i^2}{n(n^2-1)}$$

* d_i= x_i의 순위 - y_i의 순위

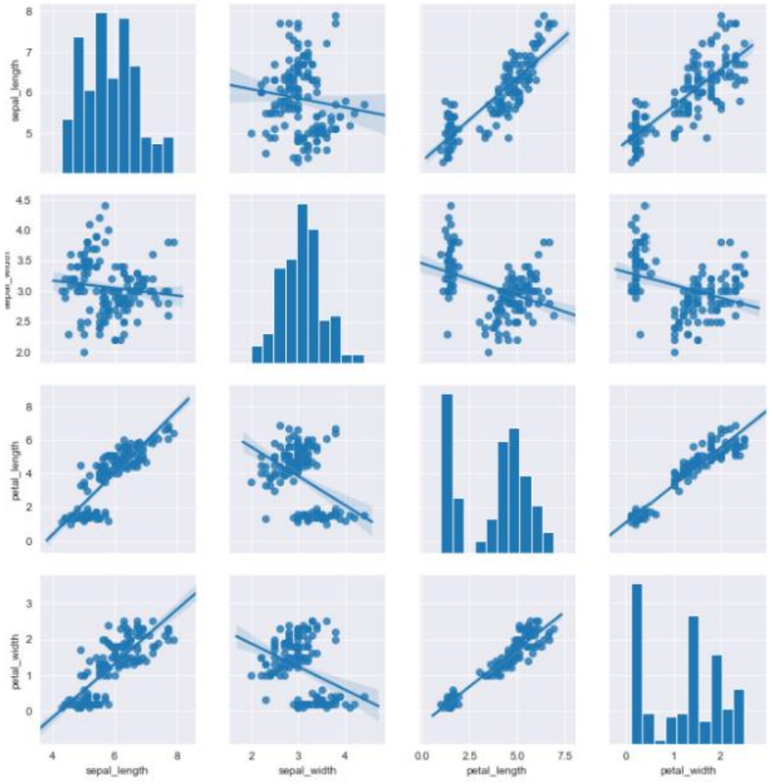
상관분석

피어슨(Pearson)-시각화 (Heatmap, Pairplot)

Heatmap



Pairplot

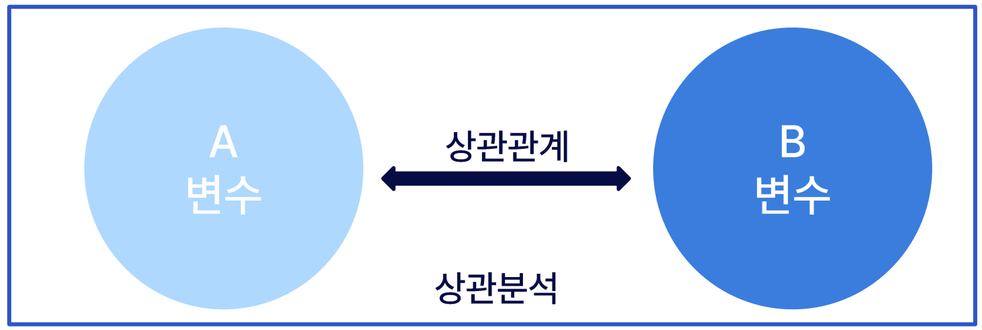


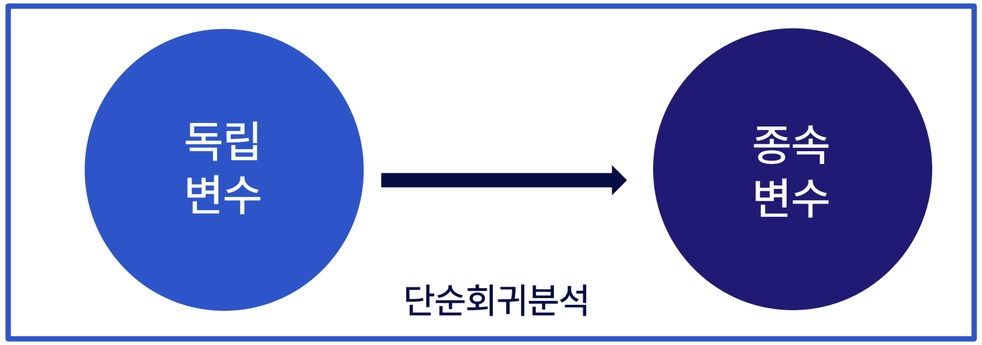
회귀분석

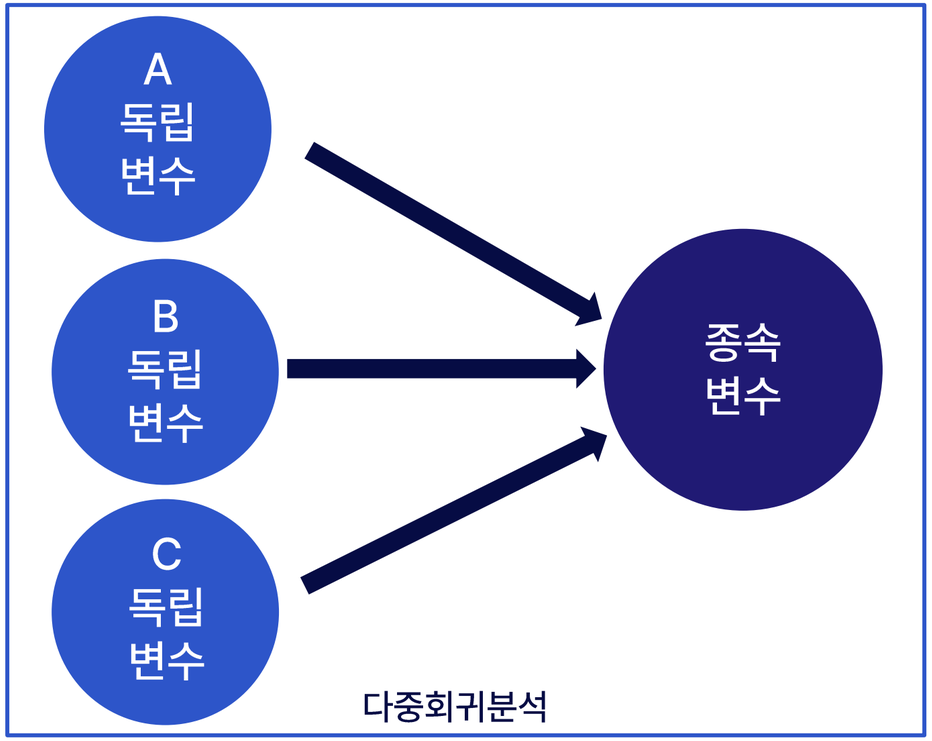
회귀분석의 기본 개념

회귀분석이란?

하나의 변수(종속변수)가 나머지 다른 변수들(독립변수)과의 선형적 관계를 갖는가의 여부를 분석하는 방법





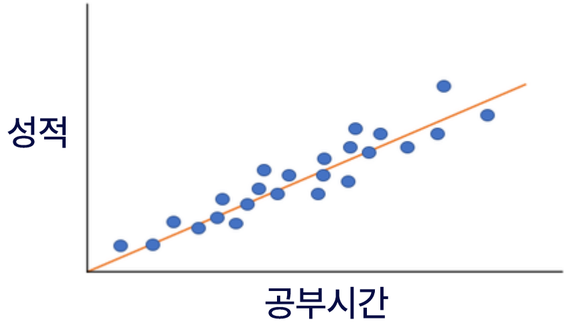


회귀분석

회귀분석의 기본 개념

독립변수와 종속변수(연속형)

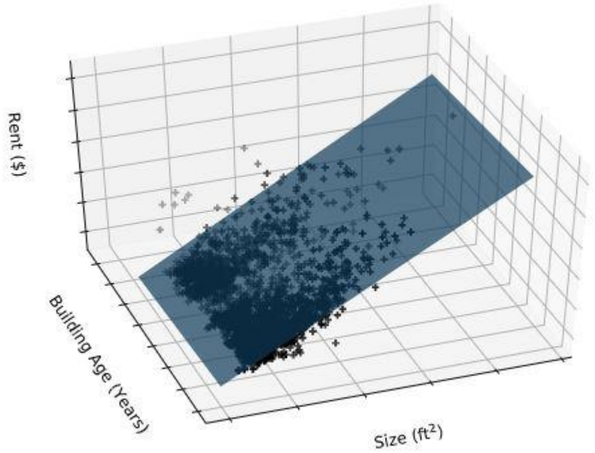
독립변수(원인 변수 or 설명 변수) : 종속변수에 영향을 미치는 변수 ex) 공부시간
종속변수(결과 변수 or 반응 변수) : 독립변수의 영향으로 나타나는 결과 변수 ex) 성적



단순 회귀 vs 다중 회귀

단변량 단순 선형 회귀: 종속변수 1개, 독립변수 1개
단변량 다중 선형 회귀: 종속변수 1개, 독립변수 2개 이상

+ 다변량 단순 선형 회귀: 종속변수 2개 이상, 독립변수 1개
+ 다변량 다중 선형 회귀: 종속변수 2개 이상, 독립변수 2개 이상



회귀분석

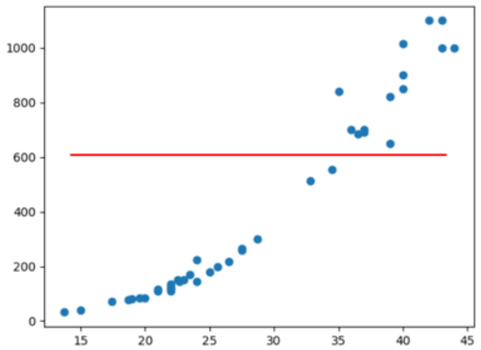
단순/다중 선형회귀분석(연속형)

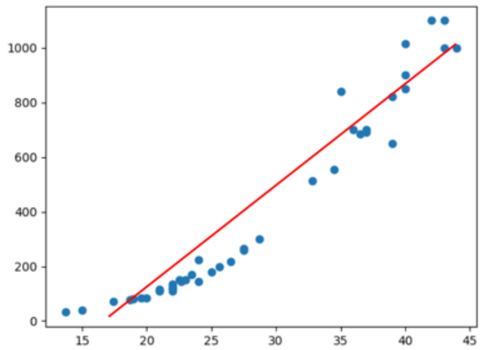
단순 선형 회귀

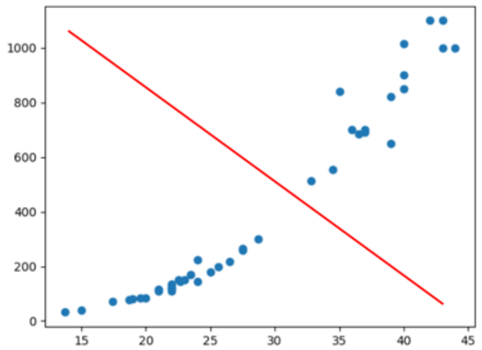
$$Y = Wx + b$$

W : 회귀 계수, 기울기, 가중치(weight)
b : 편향(bias)

Ex) 날씨가 더울수록 맥주가 많이 팔릴까?
- 독립변수(x): 기온
- 종속변수(y): 맥주 매상







다중 선형 회귀

$$Y = W_1x_1 + W_2x_2 + W_3x_3 + \cdots + W_nx_n + b$$

Ex) 여러 개의 변수가 있을 때, 모든 변수들이 유의미한가?
- 독립변수(x): 습도,기온,날씨(맑음 or 비),가격
- 종속변수(y): 매상 예측

회귀분석
다중선형회귀분석 - 다중공선성

다중공선성이란?

회귀 분석 시 사용된 모형의 일부 독립변수(설명변수)가 다른 독립변수(설명변수)와 상관 정도가 높아
데이터 분석 시 부정적인 영향을 미치는 현상

⇒상관관계가 높은 변수들을 독립변수로 사용하지 않음
⇒회귀 분석의 경우, 독립 변수들 끼리는 서로 독립이라는 가정 하에 활용
⇒보통 VIF*>10일 경우, 높다고 판단함 => 해당 변수 제거 or 다중공선성이 발생한 독립변수들을 합침

	VIF_Factor	Feature
0	10.009713	Height
1	411.350120	Weight
2	98.367558	Shucked Weight
3	62.118977	Viscera Weight
4	80.164597	Shell Weight

	VIF_Factor	Feature
0	9.460318	Height
1	9.460318	Shell Weight

Ex) 일평균 음주량 (X1), 혈중 알코올 농도(X2), 학업 성취도(Y) 라고 가정
- X1과 X2의 경우, 상관관계가 높다
- X1 or X2 중 하나는 유의미한 변수로 드러나는 반면, 나머지 계수는 불안정한 계숫값을 도출 => 설명력이 약해짐
- 마치 한 명의 학생 앞에서 선생님 두 명이 동시에 수학을 가르칠 경우 강의 전달력이 약해지는 것과 같음

* 분산 팽창 인수(VIF, Variance Inflation Factor)

회귀분석
활용 예)

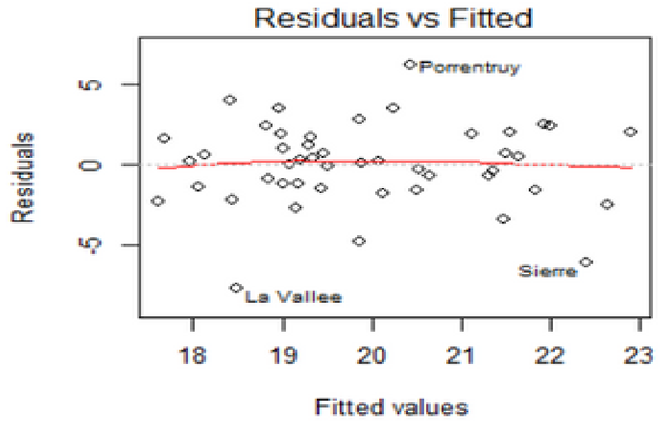
모델 검정

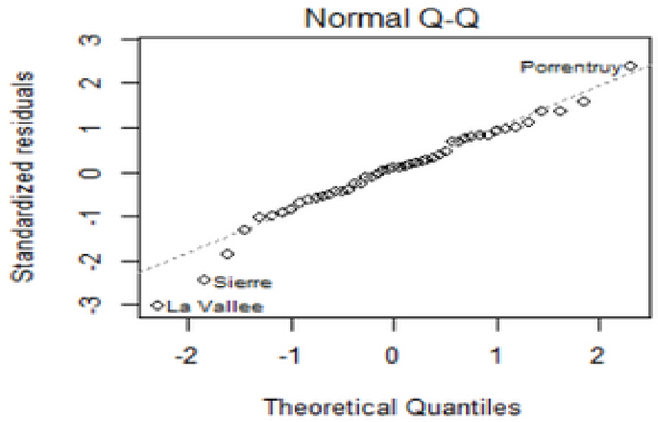
생성한 회귀모델이 적절한 모델인지를 검정하는 과정

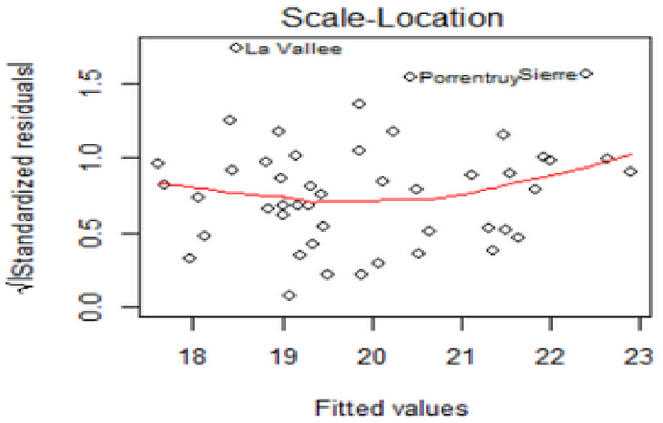
Ex) 주택 가격 데이터를 활용한 지역의 평균 주택 가격 예측
- 종속변수 :
평균 주택 가격(Y)

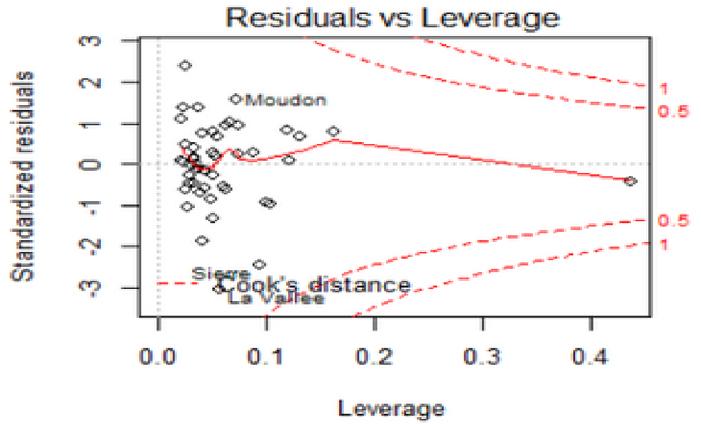
- 독립변수 :
지역 범죄율(X1)
가구당 평균 방 개수(X2)
빈곤층 비율(X3)
고속도로와의 접근성 지수(X4)

Y= W1*X1 + W2*X2 + W3*X3 + b









회귀분석

로지스틱 회귀(logistic regression)(종속 변수-범주형 (binary)

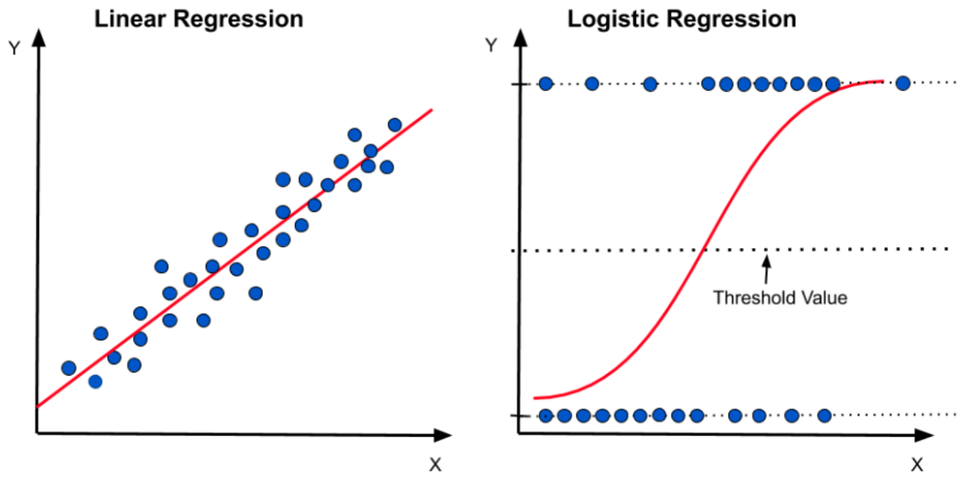
로지스틱 회귀란?

출력 변수를 직접 예측하는 것이 아닌 두 개의 카테고리를 가지는 binary 형태의 출력 변수 (성공 or 실패) , (예 or 아니오)를 예측할 때 사용하는 회귀분석 방법

Ex) 성별, 성공 여부, 합격 및 불합격, 양성/음성

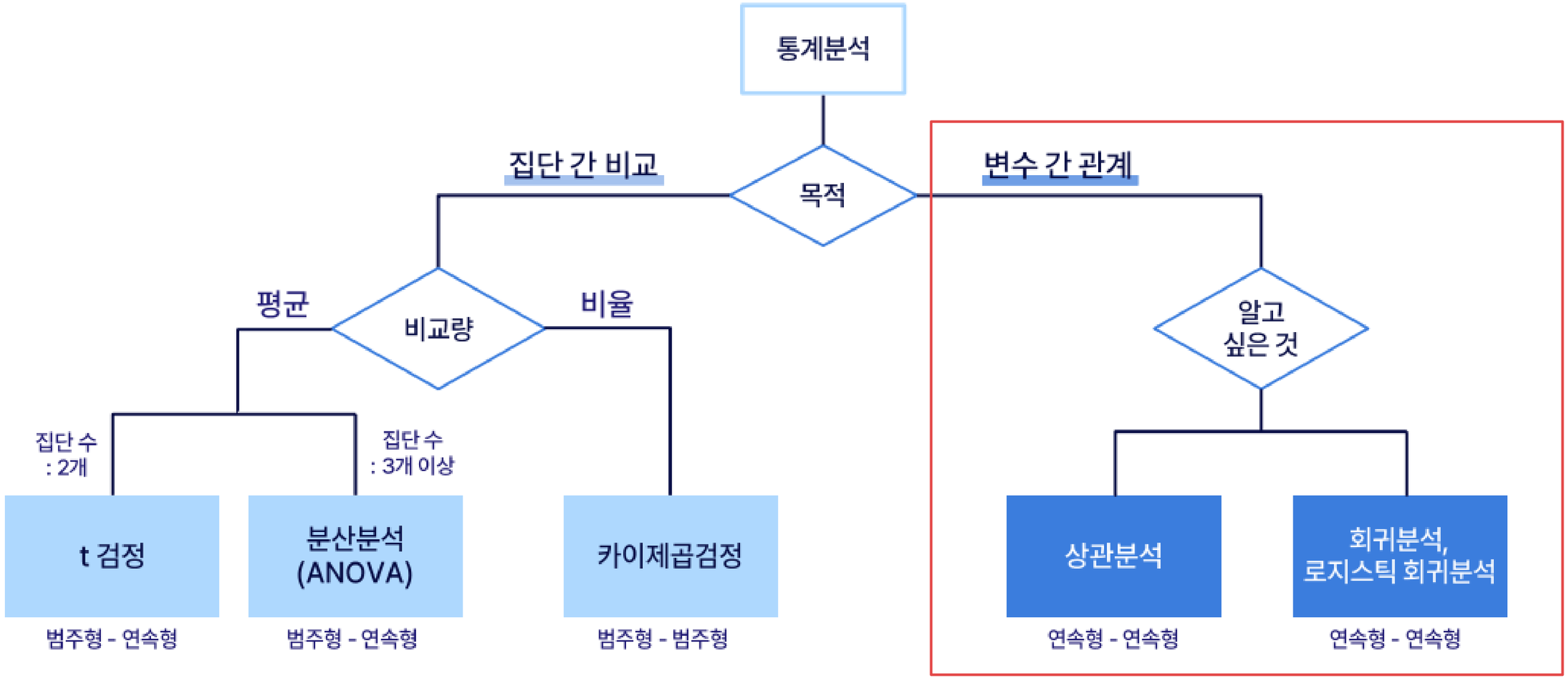
$$\hat{y} = \begin{cases} 1 & \text{if } \theta(x) \geq 0.5 \\ 0 & \text{if } \theta(x) < 0.5 \end{cases}$$

종속 변수의 확률이 0.5보다 크면 1로, 0.5보다 작으면 0으로 분류



Today

통계분석 II – 상관분석, 회귀분석



2023 D&A

Basic Session 8차시

THANK YOU