특성 공학과 규제
(1) 다중 회귀

비타민 10기 2조
조장 이주석

목차

복습 과제 풀이

저번 주차의 복습 과제를 풀이하며, 내용을 복습하는 시간을 가집니다.

01

다중 회귀

다중 회귀의 이론에 대해서 설명합니다.

02

다중 회귀 정의

- 2개 이상의 설명변수(독립변수, 원인변수)로 종속변수(반응변수, 결과변수)를 추정하는 회귀분석
- 회귀방정식을 기반으로 여러 원인 x를 사용하여 하나의 결과 y를 설명

Simple Linear Regression

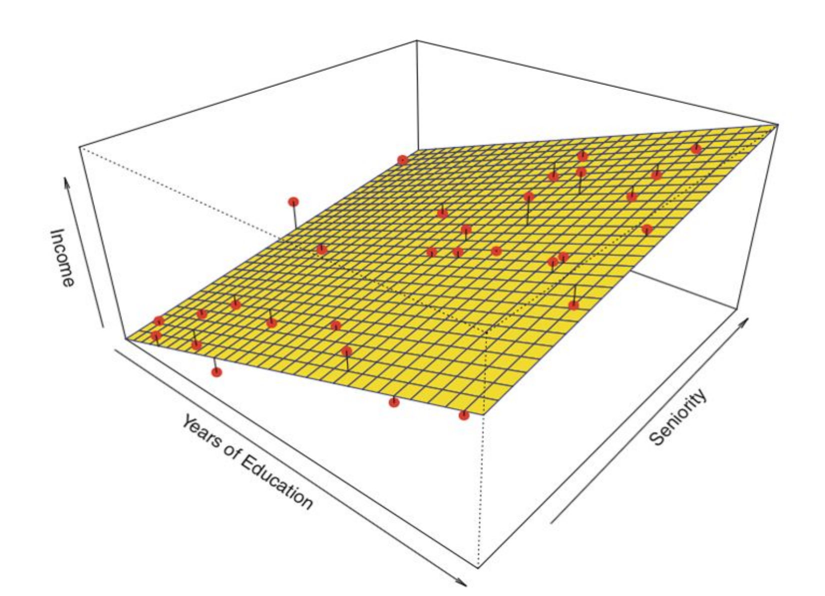
$$y = b_0 + b_1{}^*x_1$$

Dependent variable (DV)　　Independent variables (IVs)

Multiple Linear Regression

$$y = b_0 + b_1{}^*x_1 + b_2{}^*x_2 + \ldots + b_n{}^*x_n$$

다중 선형 회귀 그림 예시



- 예시) 지능(x1), 학업 동기(x2), 가정의 사회경제적 지위(x3)가 성적(y)에 미치는 영향

다중 회귀 정의

- 다중 선형 회귀는 독립변수의 숫자가 많아 행렬을 이용하여 표현하는 것이 편리
- y_i : i번째 종속변수 값
- x_{ij} : i,j번째 독립변수 값
- β_j : j번째 독립변수의 회귀계수
- ϵ_i : i번째 값의 오차

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

다중 회귀 정의

- 행렬을 이용하여 표현할 때, 행렬곱을 아는 것이 필수적
- 두 행렬 A의 열의 개수와 행렬 B의 행의 개수가 같을 때,
  행렬 A의 제i행의 각 성분과 행렬 B의 제j열의 각 성분을 그 순서대로 곱하여 더한 것을
  (i, j)성분으로 하는 행렬을 두 행렬 A와 B의 곱이라 함

$$\underbrace{\begin{bmatrix} a & b \\ c & d \end{bmatrix}}_{A} \times \underbrace{\begin{bmatrix} e & f \\ g & h \end{bmatrix}}_{B} = \underbrace{\begin{bmatrix} ae + bg & af + bh \\ ce + dg & cf + dh \end{bmatrix}}_{C}$$

$$\boxed{\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}}$$

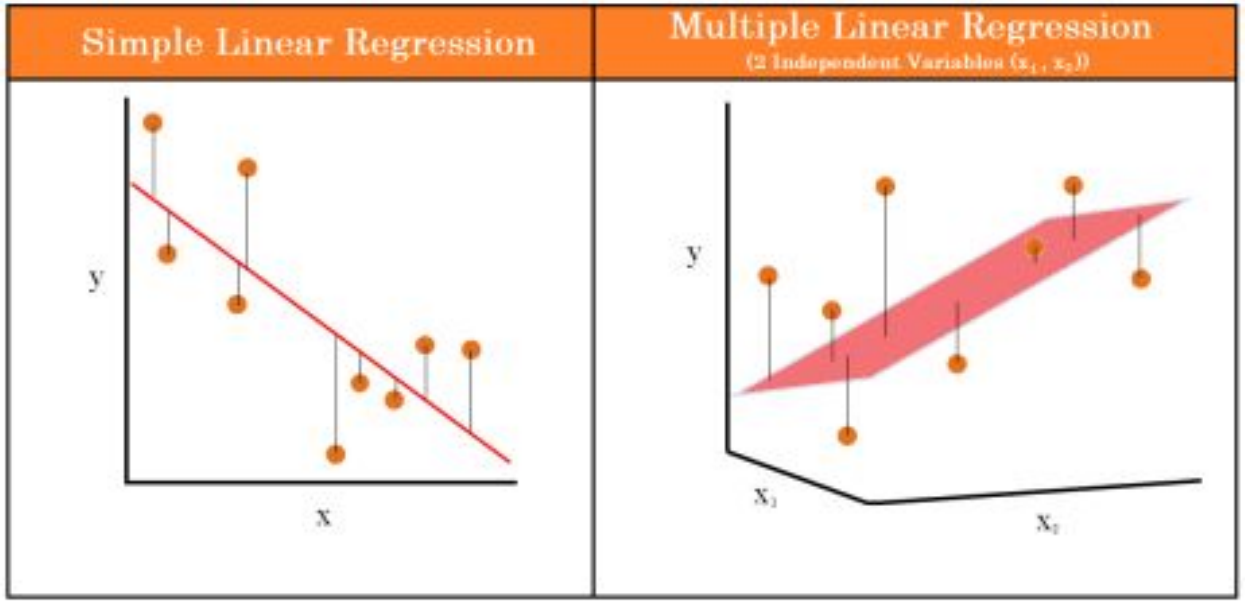
다중 회귀 정의

성적 : y 학업 동기 : x2
지능 : x1 가정의 사회경제적 지위 : x3

- 단순 회귀가 아닌, 다중 회귀를 써야 할 때?

- 단순 선형 회귀 : y = b0 + b1 * x1 VS. • 다중 선형 회귀 : y = b0 + b1 * x1 + b2 * x2 + b3 * x3

- 단순 회귀는 영향을 미치는 변수 생략할 위험이 큼



따라서, 다중 회귀를 통해 영향을 미치는 변수
모두 포함하여 더 정확한 추정 가능

다중 회귀 정의

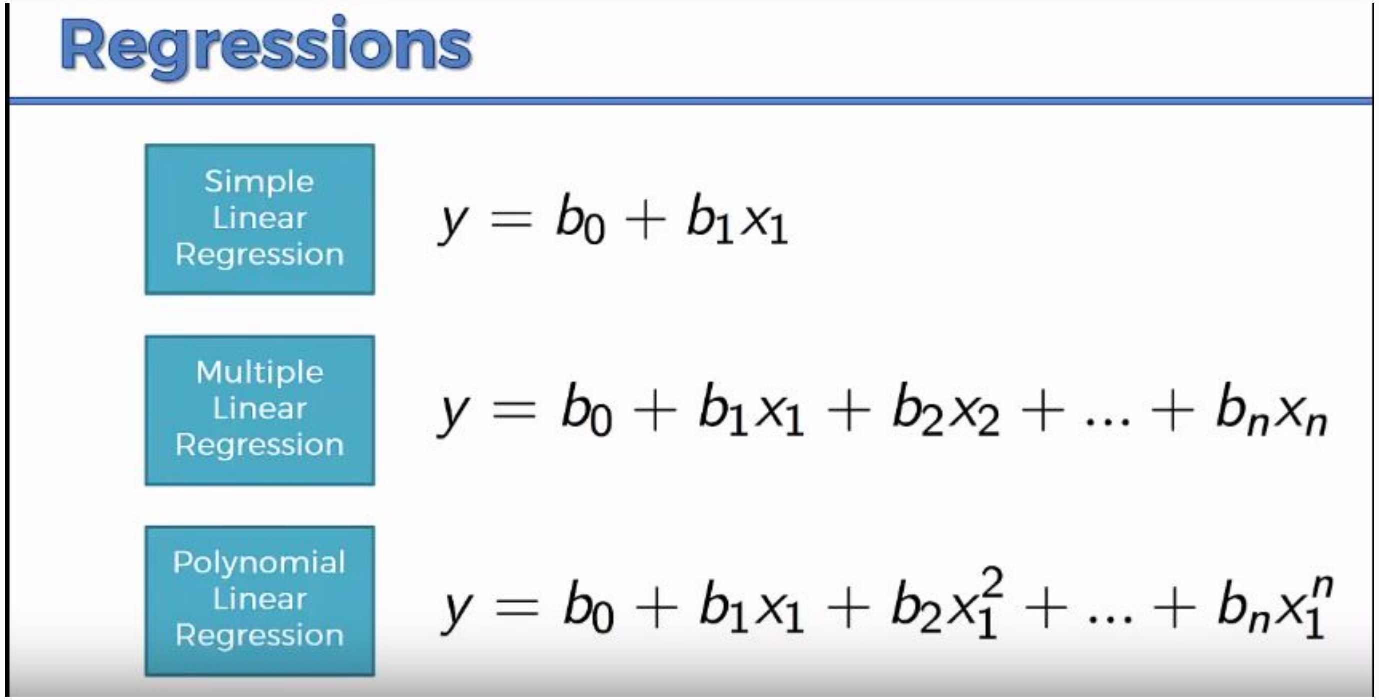
- 다중 선형 회귀분석의 기본 가정 4가지
- 선형성(Linearity) : 종속 변수와 독립 변수 사이에는 선형 관계가 있다.
- 독립성(Independency) : 독립 변수는 서로 linearly independent(선형 독립)이다.
- 정규성(Multivariate Normality) : residual(잔차)가 정규분포를 따른다.
- 등분산성(Homoscedasticity) : 분석하는 집단의 분산이 같다.

만약, 위 기본 가정 4가지가 충족되지 않는다면,
회귀분석 수행 시 왜곡된 결과 도출 가능

다중 선형 회귀 VS. 다항 회귀

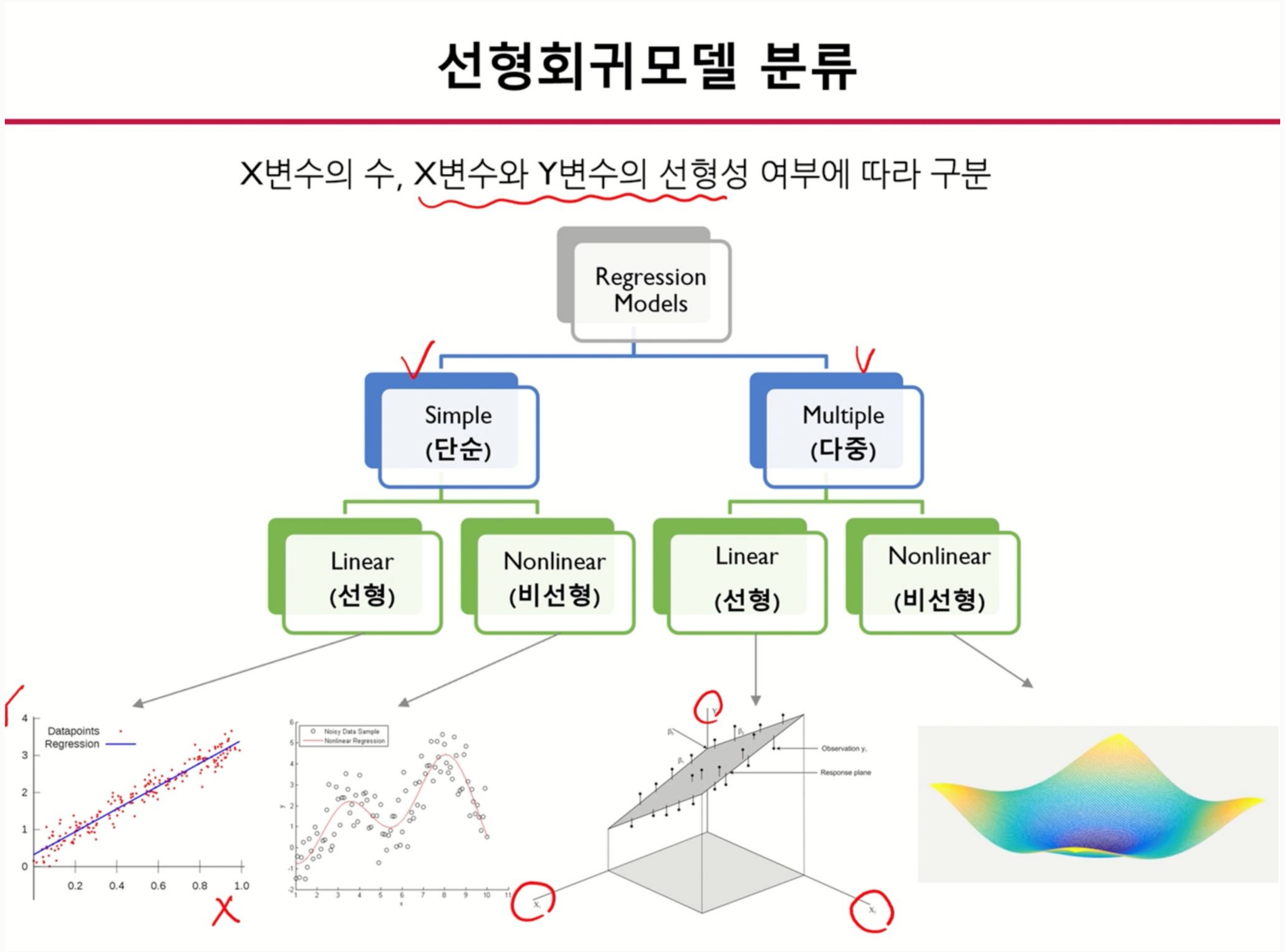
- 다중 선형 회귀는 독립변수들과 종속변수가 선형 관계
- 다항 회귀는 독립변수들과 종속변수가 선형 관계가 아님!
- 다항 회귀는 독립변수의 숫자가 선형 회귀처럼 한 개여도 성립

  예시) y = b0 + b1 * x1**2



선형 회귀 분류

- 회귀 모델을 상황에 따라 분류하여 적절히 사용하는 것이 중요



공분산과 상관계수

다중 회귀분석에서 가장 중요한 개념은 공분산과 상관계수입니다.
이것들의 개념을 설명합니다.

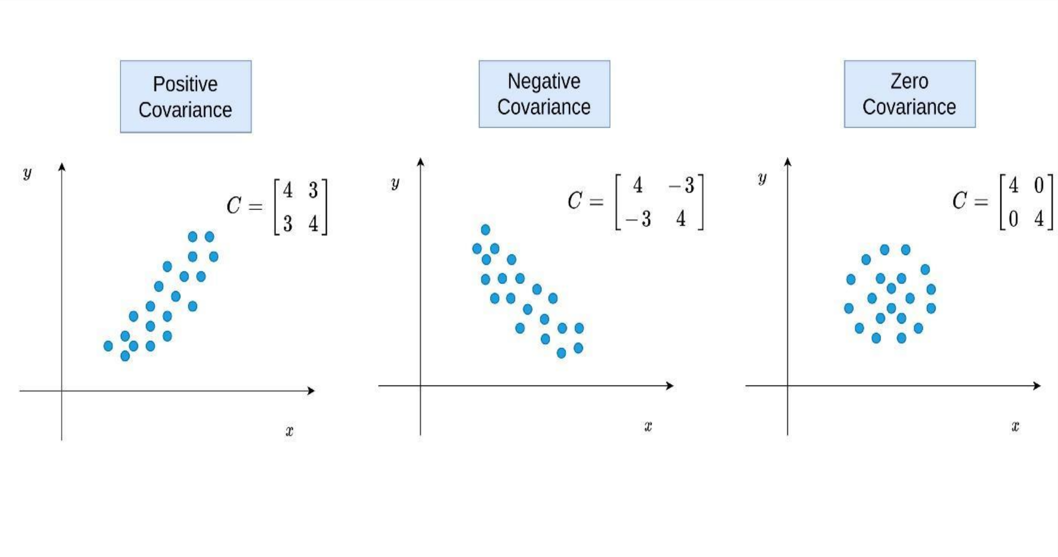
03

공분산과 상관계수

- 공분산
- 2개의 확률변수의 선형 관계를 나타내는 값

$$\mathrm{cov}\,(X,\,Y) = \sum_{i=1}^{N} \frac{(x_i - \overline{x})\,(y_i - \overline{y})}{N}.$$

공분산 공식



공분산 값에 따른 분포

$$\Sigma = \begin{bmatrix} Var(X_1) & Cov(X_1, X_2) & \cdots & Cov(X_1, X_p) \\ Cov(X_1, X_2) & Var(X_2) & \cdots & Cov(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_1, X_p) & Cov(X_2, X_p) & \cdots & Var(X_p) \end{bmatrix}$$

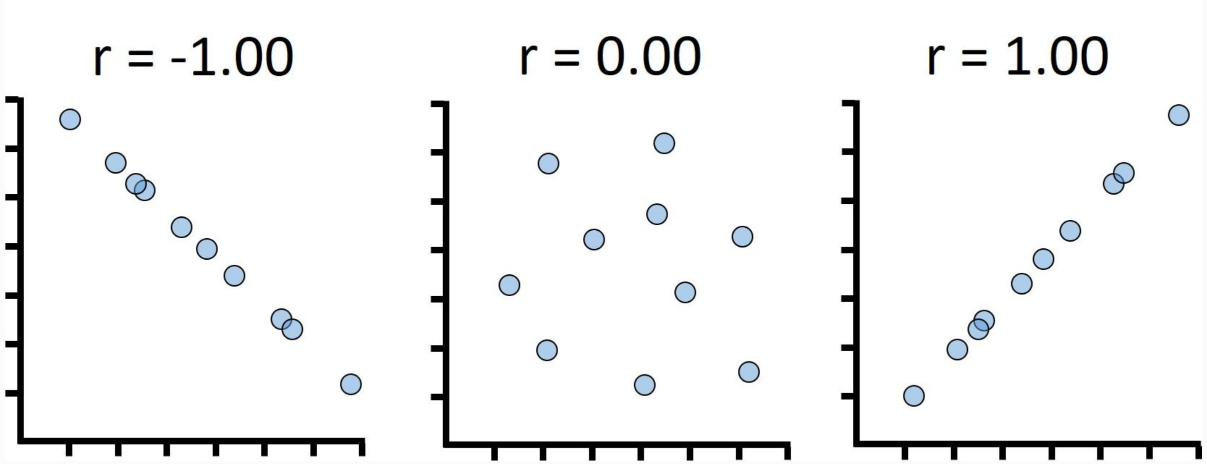
공분산 행렬

공분산과 상관계수

- 상관계수
- 두 변수 x와 y 간의 선형 상관관계를 계량화한 수치

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \times \Sigma(y_i - \bar{y})^2}}$$

상관계수 공식



상관계수 값에 따른 분포

다중공선성

다중공선성은 다중회귀분석에서 자주 다뤄지는 문제이며,
과적합과 함께 모델의 성능을 결정하는 중요한 개념입니다.

04

다중공선성(Multicollinearity)

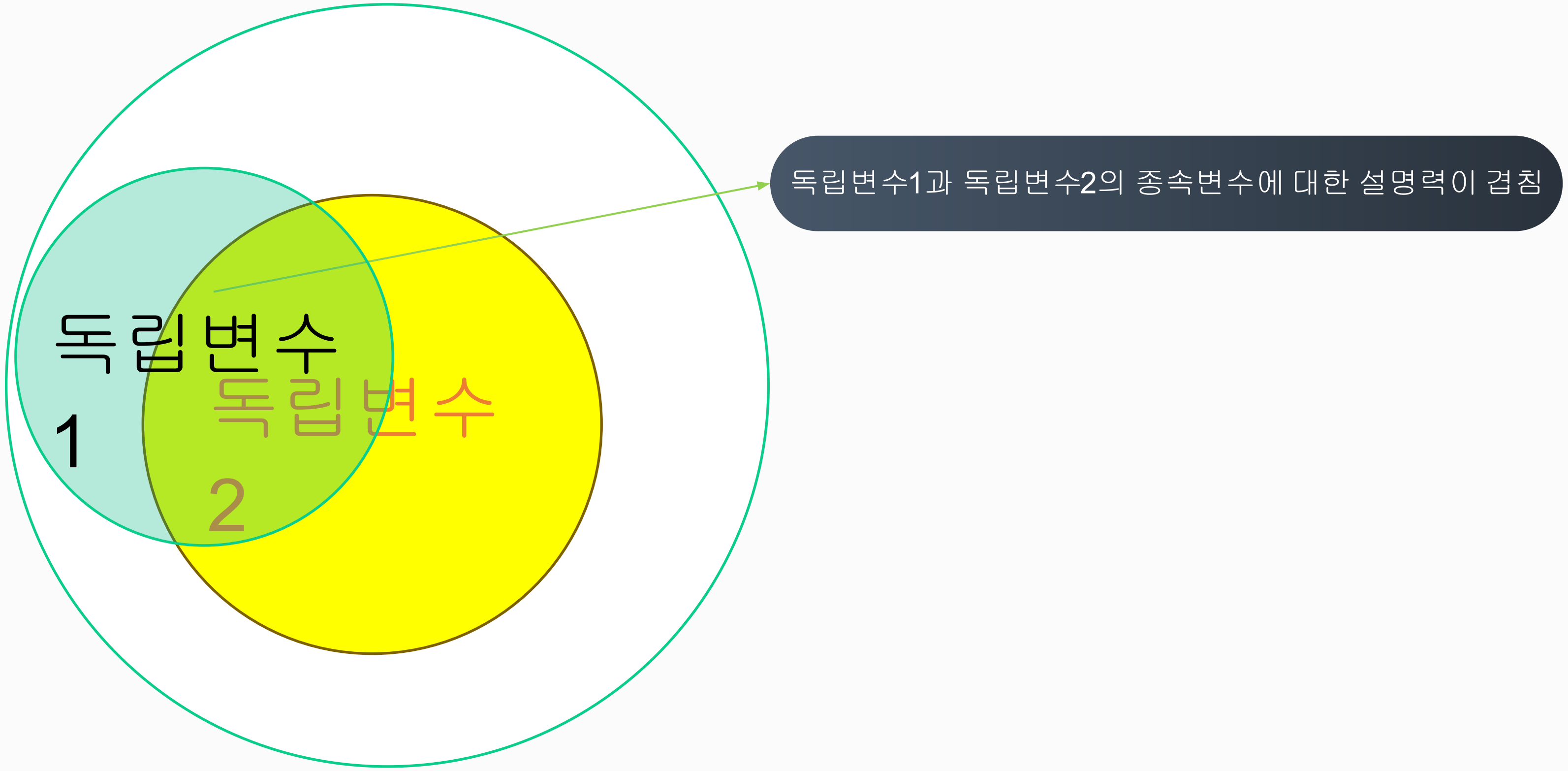
- 다중 회귀분석에서 독립변수들 간에
  높은 상관관계가 나타나는 현상

EX) 종속변수 = 독립변수1 + 독립변수2 + 독립변수3 + …

다중공선성은 다중 회귀분석의 기본 가정인 독립성
즉, 독립변수들끼리는 선형 독립이어야 한다는 가정을 위배하게 됩니다.
이는 회귀분석에 오류를 일으키게 되고, 모델의 성능을 저하시키는 효과를 낳게 됩니다.

다중공선성(Multicollinearity)

종속변수의 변동성(분산)



결과에 왜곡이 생길 수 있음
이를 해결하는 방법으로는 변수 제거
또는 새로운 변수 생성 등이 있음

다중공선성(Multicollinearity)

- VIF(Variance Inflation Factor)
- 분산 팽창 인수

- VIF(Variance inflation factor)

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}$$

VIF가 10 이상인 경우 다중공선성이 있는 변수라고 판단

$$x_1 = \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_p X_p + \varepsilon$$

X1을 종속변수, 나머지 변수를 독립변수로 하여 회귀 모델($f1$) 적합

$$\downarrow$$

$$R_1{}^2$$

$f1$ 의 R^2를 이용하여 VIF_1계산

$$\downarrow$$

$$VIF_1 = \frac{1}{1-R_1{}^2}$$

VIF_1의 의미 : 다른 변수의 선형결합으로 X1을 설명할 수 있는 정도

$R^2 > 0.9$ 이상인 경우, $\mathrm{VIF} > 10$

VIF값이 높을수록(보통 10 이상), 나머지 독립변수의 해당 독립변수에 대한 종속성이 높다고 판단될 수 있으므로,
그 변수에 대한 다중공선성이 높다고 판단할 수 있습니다.

특성 공학

데이터의 특성을 다루는 것은 매우 중요합니다.

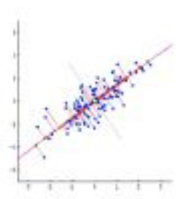
05

특성 공학(Feature Engineering)

- 모델의 성능을 향상시키기 위해
  데이터의 특성을 생성, 선택하는 작업

생성

PCA(주성분 분석)



PolyNomial Features

제곱	곱
방의 한변의 길이	길이 X 높이, 방 개수 X 방 면적

특성 생성

데이터로부터 새로운 특성을 만드는 과정이며, 해당 데이터에 대한 도메인지식을 바탕으로 데이터에 여러 조작을 하는 것

선택

변수 선택법

Forward selection　Backward Elminiation　Stepwise selection

특성 선택

데이터에서 불필요한 특성을 도메인 지식 바탕으로 지우는 과정, 혹은 여러 알고리즘을 바탕으로 최적의 특성을 선택하는 과정

데이터 분석

데이터 소개와 간단한 분석을 해보려 합니다.

06

보스턴 집값 데이터

- Boston House Prices
- Kaggle에서 주로 다뤄지는 데이터이며, 1978년 당시에 보스턴에서의 주택가격(종속변수)에 영향을 미치는 요소들(독립변수)를 정리해놓은 정형 데이터
- 데이터가 sklearn.datasets.load_boston에 저장되어 있으므로, 언제 어디서든 손쉽게 연습할 수 있다는 장점이 있음

14개 속성의 의미

CRIM : 지역별 범죄 발생률
ZN : 25,000평방피트를 초과하는 거주 지역 비율
INDUS : 비상업 지역의 넓이 비율
CHAS : 찰스강의 더미변수(1은 강의 경계, 0은 경계 아님)
NOX : 일산화질소 농도
RM : 거주할 수 있는 방 개수
AGE : 1940년 이전에 건축된 주택 비율

DIS : 5개 주요 고용센터까지 가중 거리
RAD : 고속도로 접근 용이도
TAX : 10,000달러당 재산세 비율
PTRATIO : 지역의 교사와 학생 수 비율
B : 지역의 흑인 거주 비율
LSTAT : 하위 계층의 비율
PRICE(MEDV) : 본인 소유 주택 가격의 중앙값

보스턴 집값 데이터

데이터 불러오기

```
from sklearn.datasets import load_boston
import pandas as pd

df = pd.DataFrame(data = load_boston().data, columns = load_boston().feature_names)

df["Price"] = load_boston().target
```

데이터프레임 형태로 불러옴

보스턴 집값 데이터

train - test 분리 & 다항회귀 준비

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import PolynomialFeatures

feature_list = df.columns.difference(['Price'])
X = df[feature_list]
y = df["Price"]

train_input, test_input, train_target, test_target = train_test_split(X, y, random_state = 42)

poly = PolynomialFeatures(include_bias = False)
poly.fit(train_input)
train_poly = poly.transform(train_input)
print(train_poly.shape)
```

```
(379, 104)
```

다중 선형 회귀 및 다항 회귀를 위한 훈련, 평가 데이터셋 분리

보스턴 집값 데이터

모델 평가

```
from sklearn.linear_model import LinearRegression

lr = LinearRegression()
lr.fit(train_poly, train_target)

print(lr.score(train_poly, train_target))
```

```
0.944831397521159
```

```
from sklearn.linear_model import LinearRegression

lr2 = LinearRegression()
lr2.fit(train_input, train_target)

print(lr2.score(train_input, train_target))
```

```
0.748087259862344
```

평가점수 상으로는, 다항 회귀가 더 적합한 모델

보스턴 집값 데이터

다중공선성 확인

```
from statsmodels.stats.outliers_influence import variance_inflation_factor

def vif_df(X):
  vif = pd.DataFrame()
  vif["VIF_Factor"] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
  vif["Feature"] = X.columns
  return vif
```

statsmodel 라이브러리의 다중공선성 확인 함수 이용

보스턴 집값 데이터

다중공선성 확인

	VIF_Factor	Feature
0	2.100373	CRIM
1	2.844013	ZN
2	14.485758	INDUS
3	1.152952	CHAS
4	73.894947	NOX
5	77.948283	RM
6	21.386850	AGE
7	14.699652	DIS
8	15.167725	RAD
9	61.227274	TAX
10	85.029547	PTRATIO
11	20.104943	B
12	11.102025	LSTAT

VIF 값이 10 이상인 변수가 너무 많음
=> 다중공선성 존재 가능성 매우 높음