Round 9

PRESS
START

파뿌리
파이썬 뿌시는 이십대들!!

《 Round 9 》

- 모델링
- 회귀분석
- 회귀분석 실습

New Assignment

파뿌리
파이썬 뿌시는 이십대들!!

《 Round 9 》

- 모델링 《
- 회귀분석
- 회귀분석 실습

파뿌리
파이썬 뿌시는 이십대들!!

Exploratory Data Analysis

- 문제 정의
- 시각화 & 변수탐색
- 결측치, 이상치 탐지

Data Preprocessing

- 적절한 데이터 처리
- 정규화
- 교차검증 설정

Feature Engineering

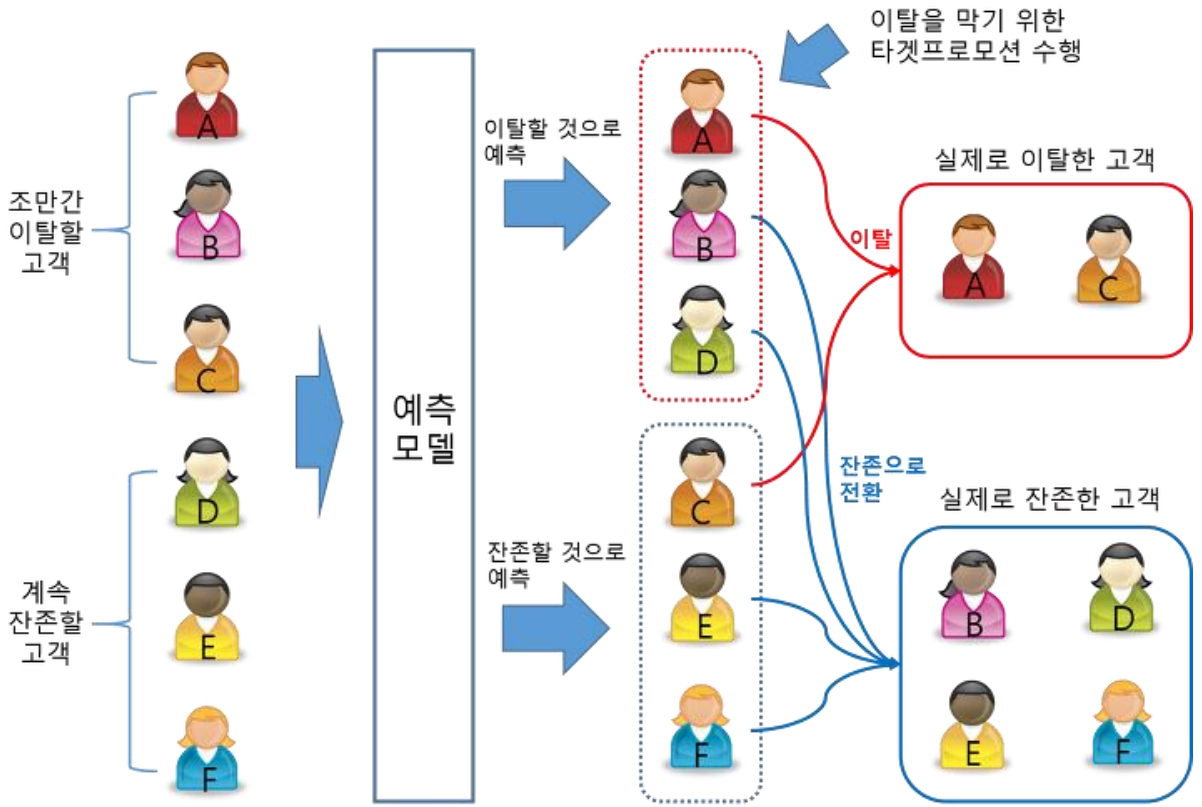
- 변수 생성
- 차원 축소
- 특징 추출

Modeling

- 예측 모델링
- 분류예측 모델링
- 결과 해석

Modeling?

- Model + ing = 모델을 만드는 일
- Model : 특정 조건 하에서 관심의 대상이 되는 변인이 갖게 될 값을 예측



파뿌리
파이썬 뿌시는 이십대들!!

Machine Learning?

- 인공지능의 한 분야

- 어떠한 작업 T에 대해 꾸준한 경험 E를 통하여
  그 T에 대한 성능 P를 높이는 것

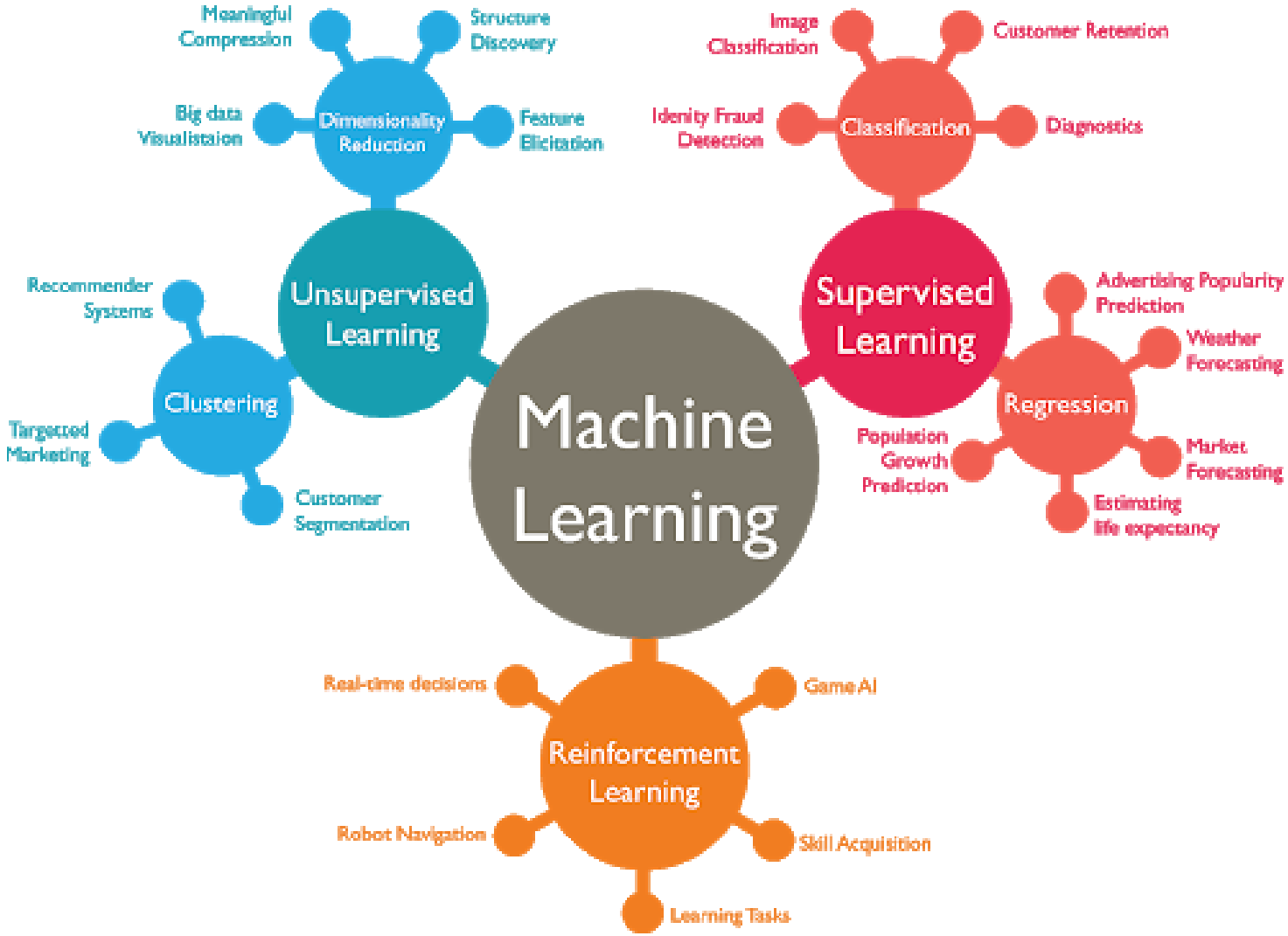
ex)

- 주가, 환율 등 경제지표 예측 (Prediction)

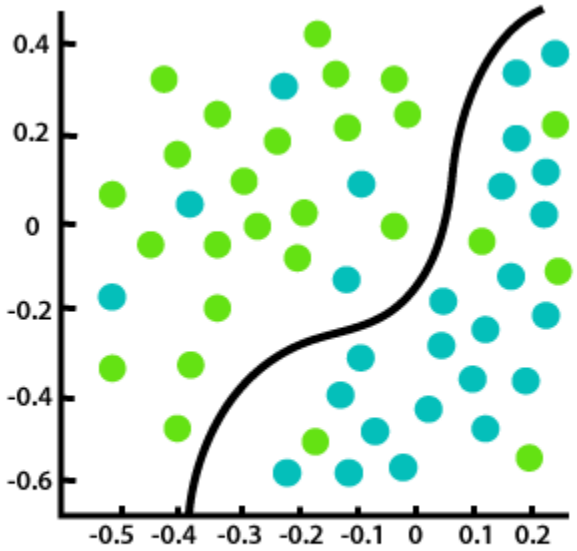
- 은행에서 고객을 분류하여 대출을 승인하거나 거절 (Classification)

- 비슷한 소비패턴을 지닌 고객 유형을 군집으로 묶음 (Clustering)

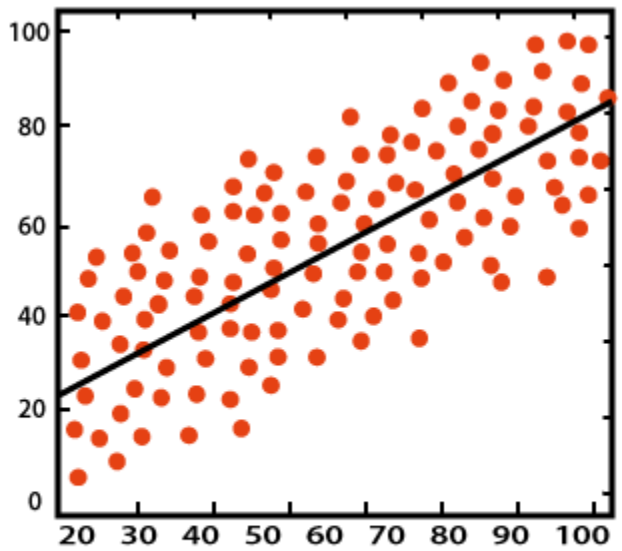
Machine Learning?



Supervised Learning



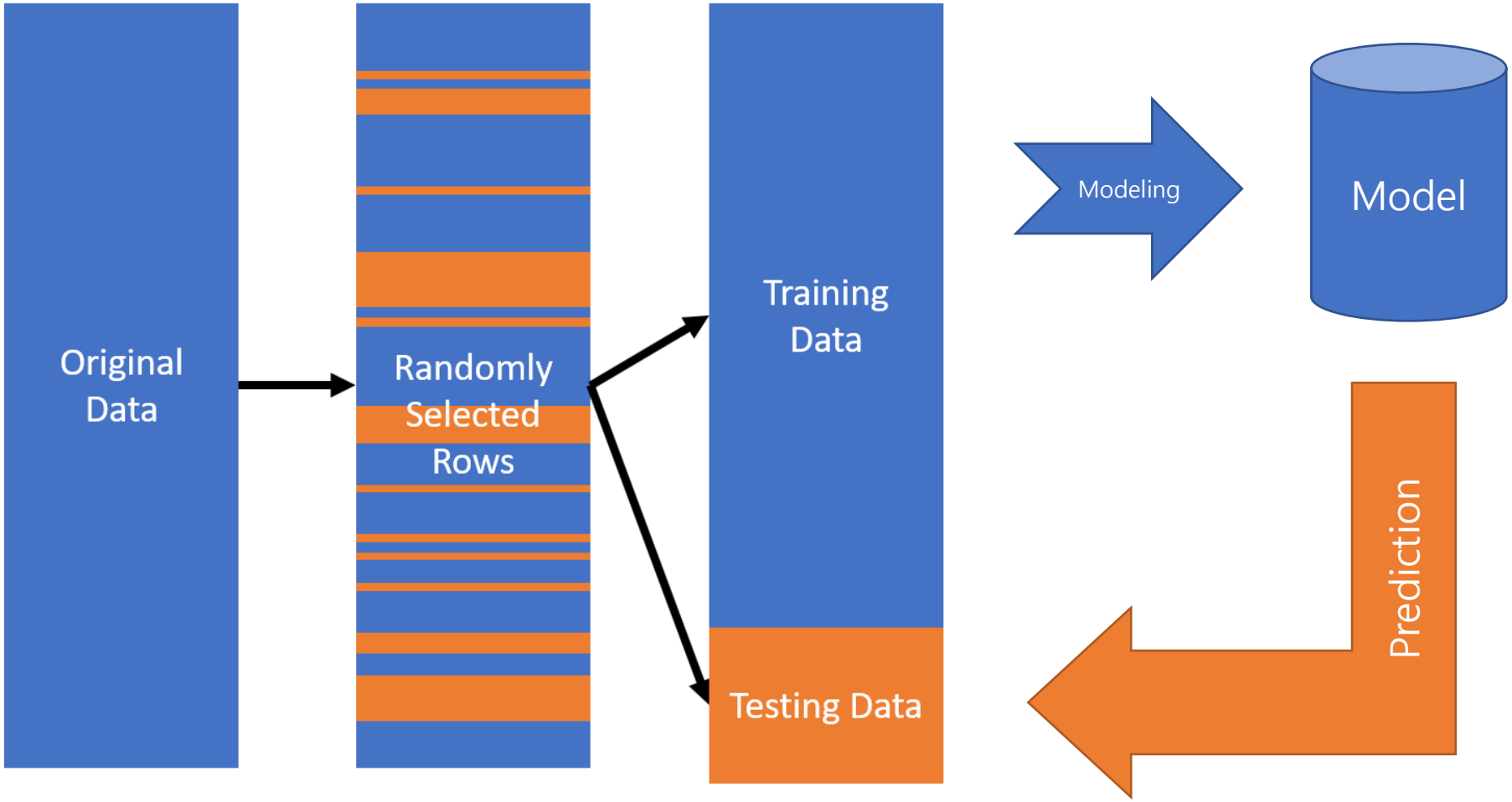
Classification



Regression

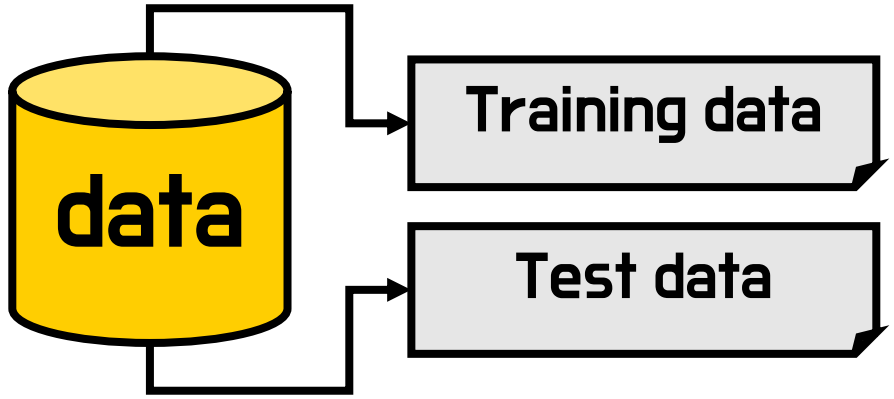
Learning library

Data splitting



Data splitting

```
## dataset을 training data와 test data로 분할
# train_test_split(독립변수, 종속변수, test data 사이즈(%), 랜덤 추출 시드값)
from sklearn.model_selection import train_test_split
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.3, random_state=10)

print('train data 개수: ', len(X_train))
print('test data 개수: ', len(X_test),"\n")
```

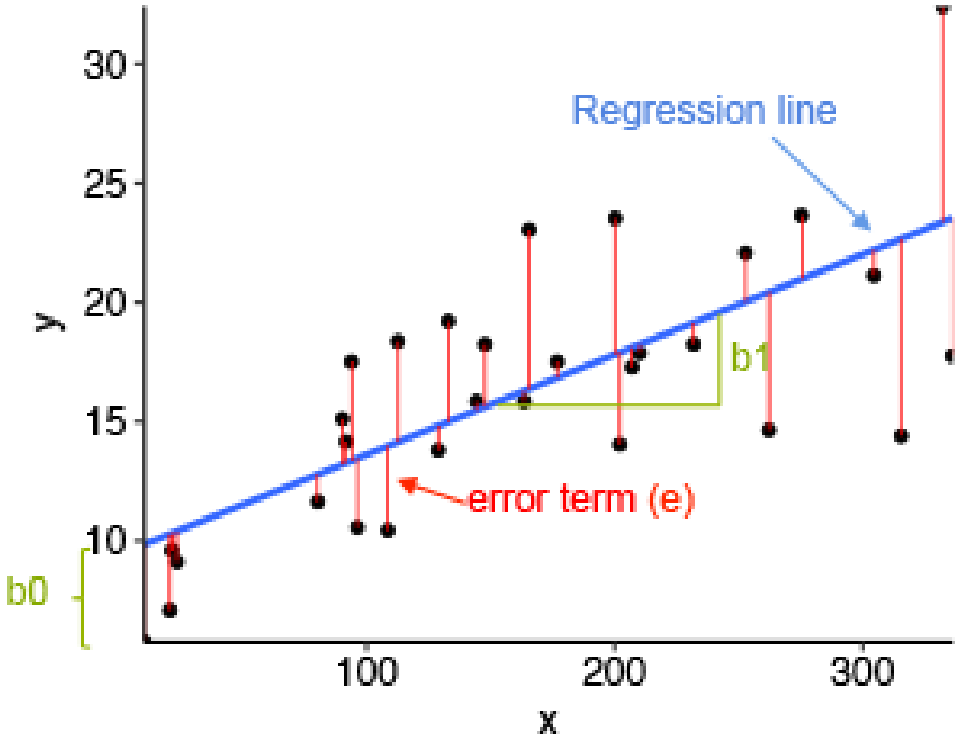


파뿌리
파이썬 뿌시는 이십대들!!

《 Round 9 》

- 모델링
- 회귀분석 《
- 회귀분석 실습

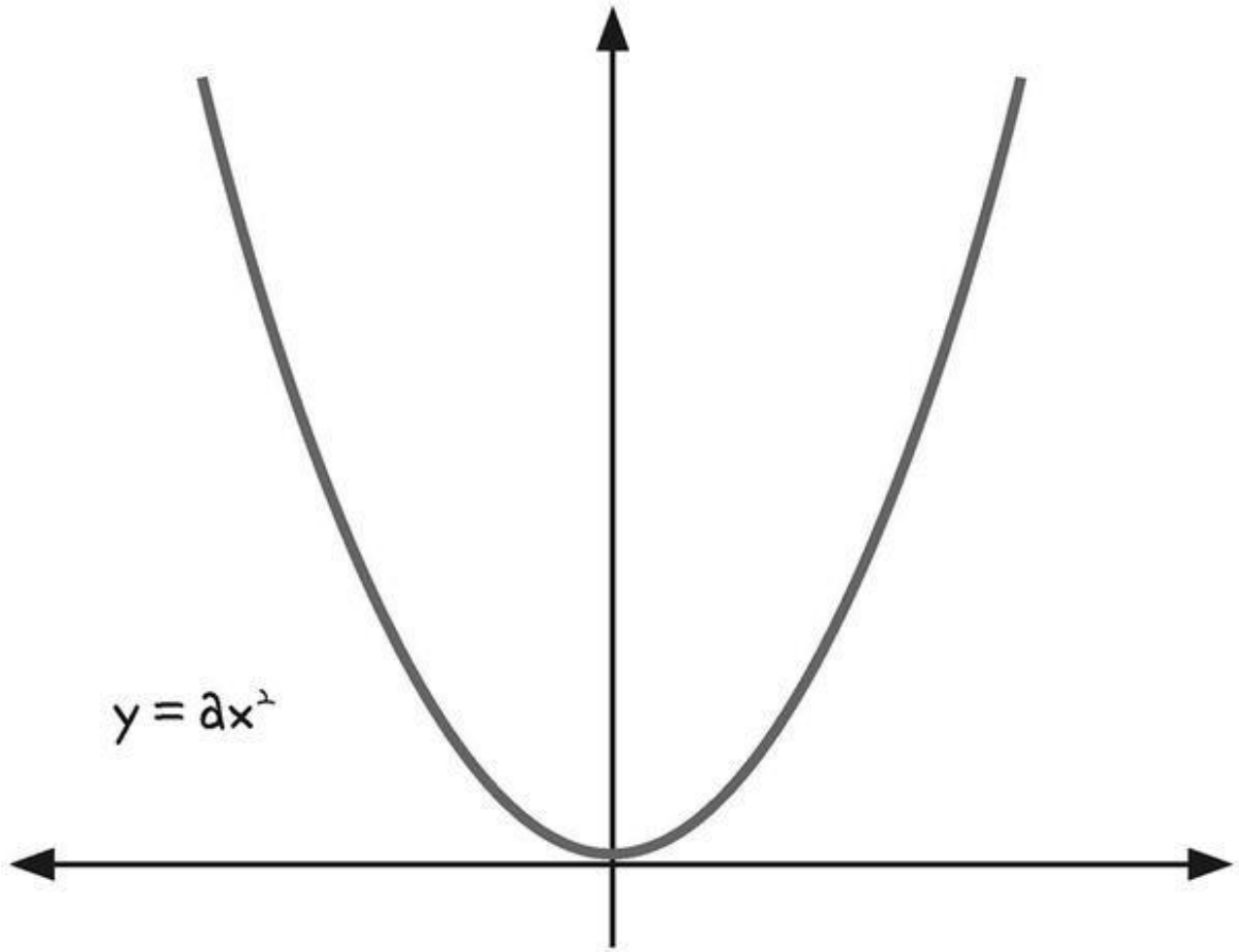
Regression



- 원인변수(x)와 결과변수(y)의 관계를 찾자!
- 즉, $B_0 + B_n X_n \cdots$ 을 찾는 것.
- 에러의 합의 제곱이 최소화 되는 회귀식을 근사
- 관계설명, 추이예측 등에 활용!

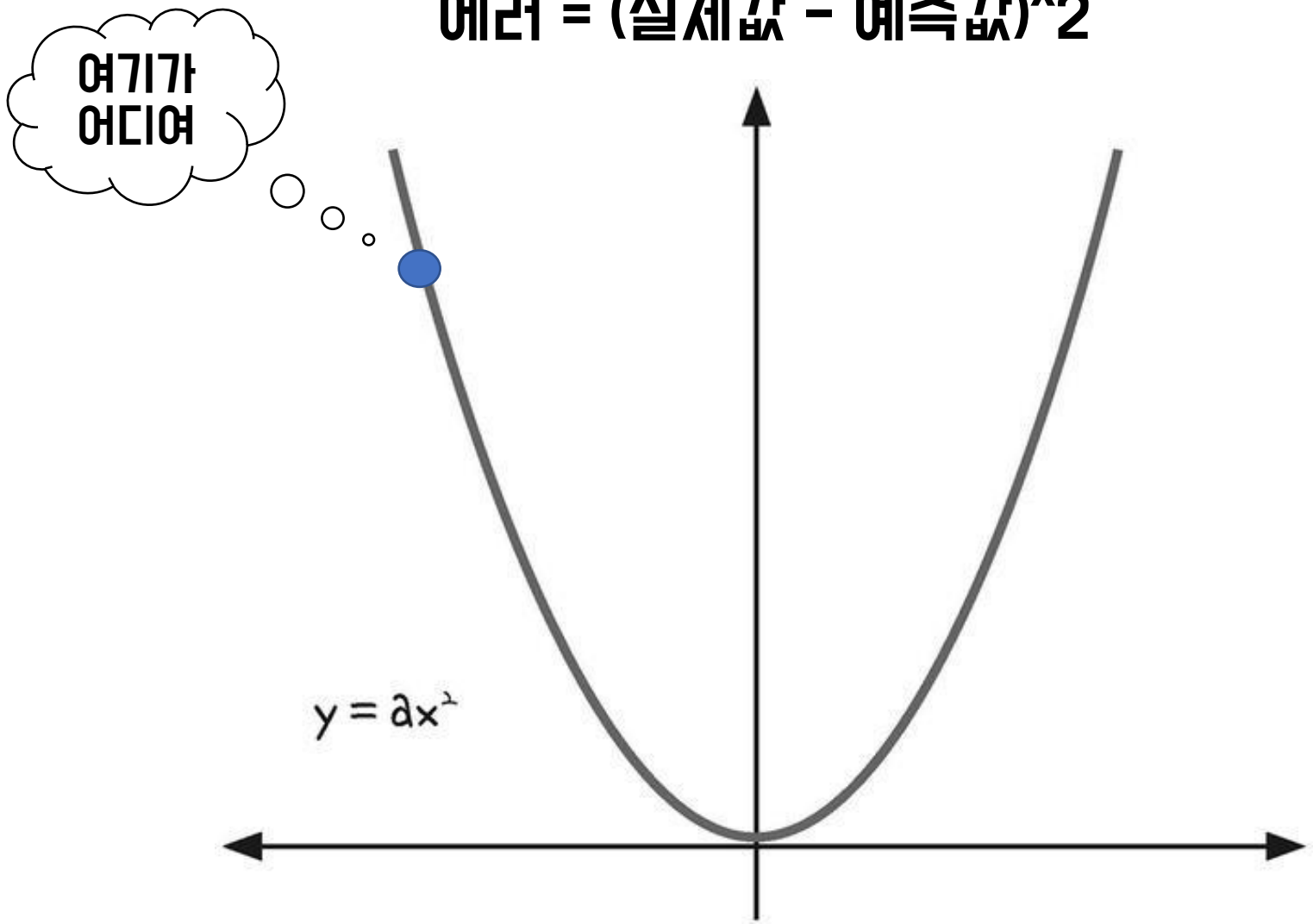
Regression

에러 = (실제값 - 예측값)^2



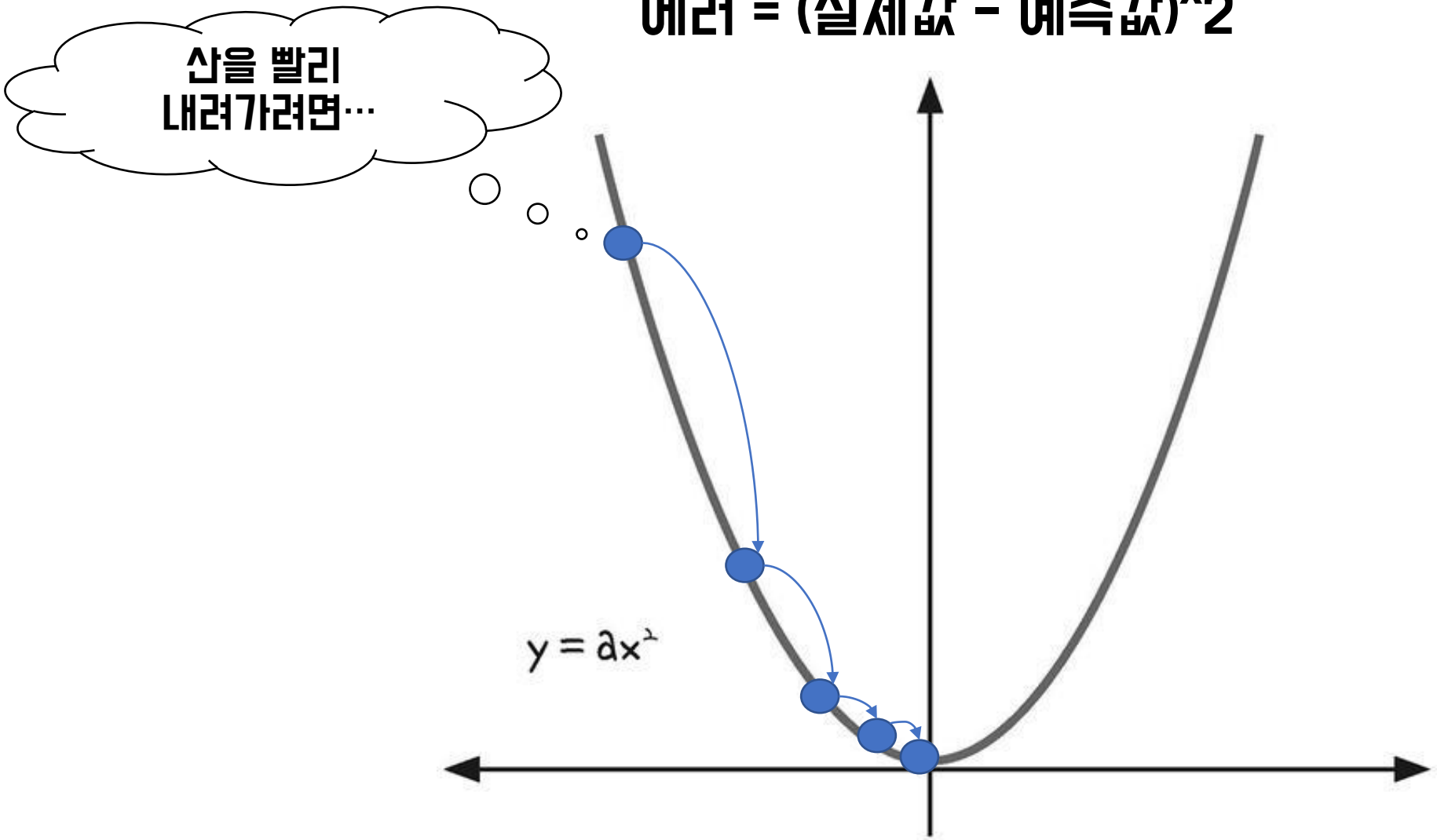
Regression

에러 = (실제값 - 예측값)^2

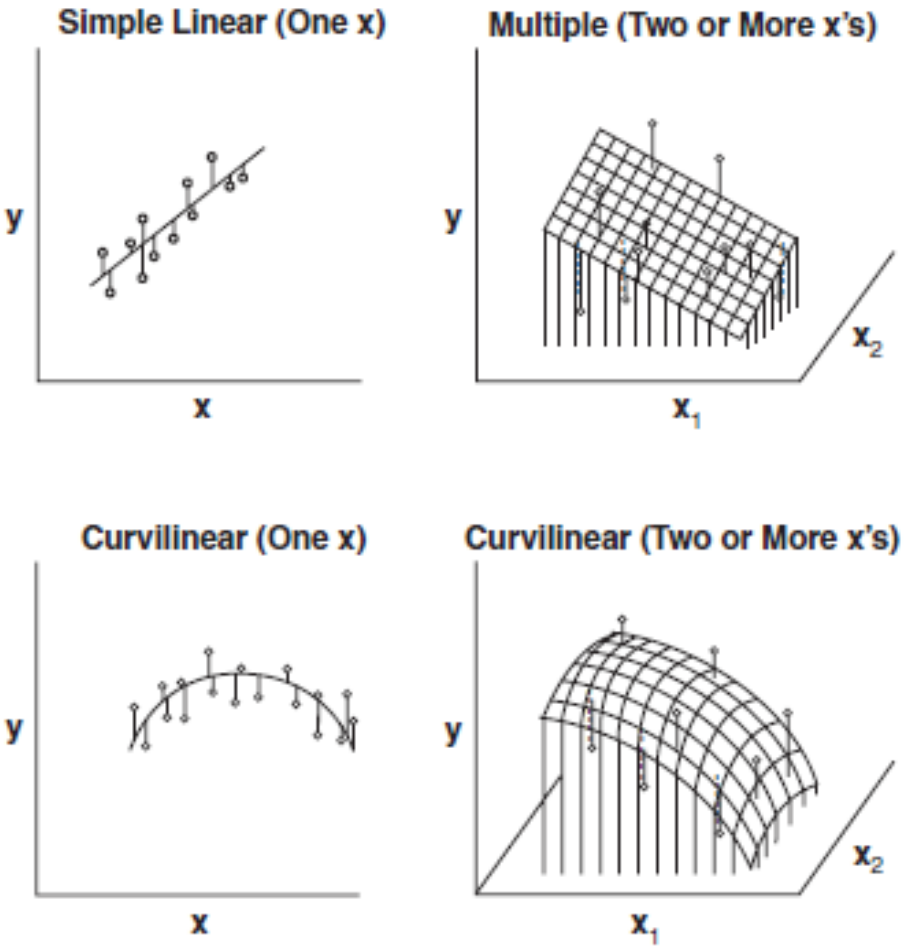


Regression

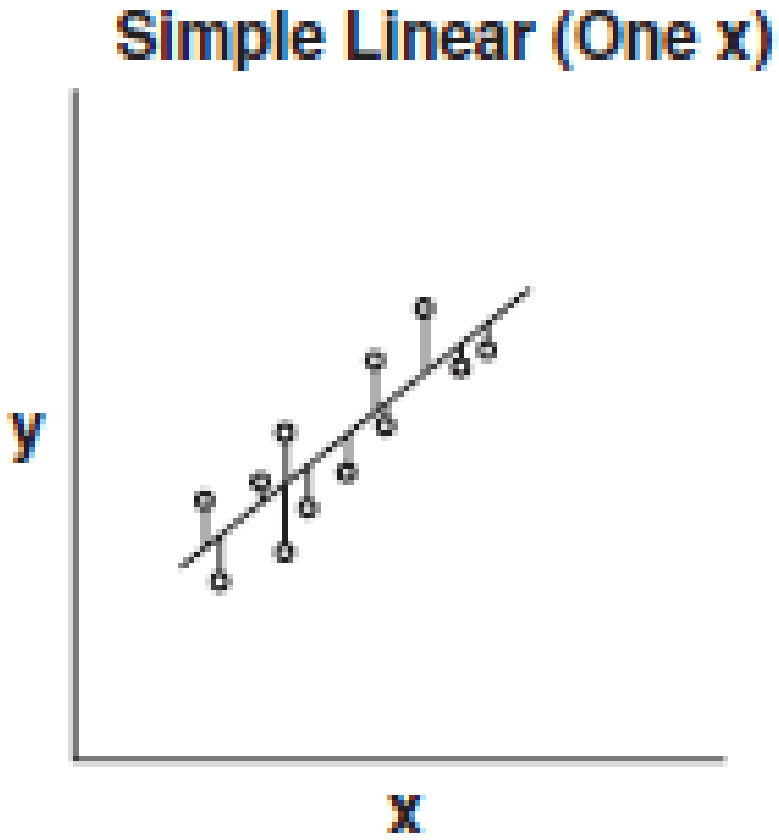
에러 = (실제값 - 예측값)^2



Types of regression

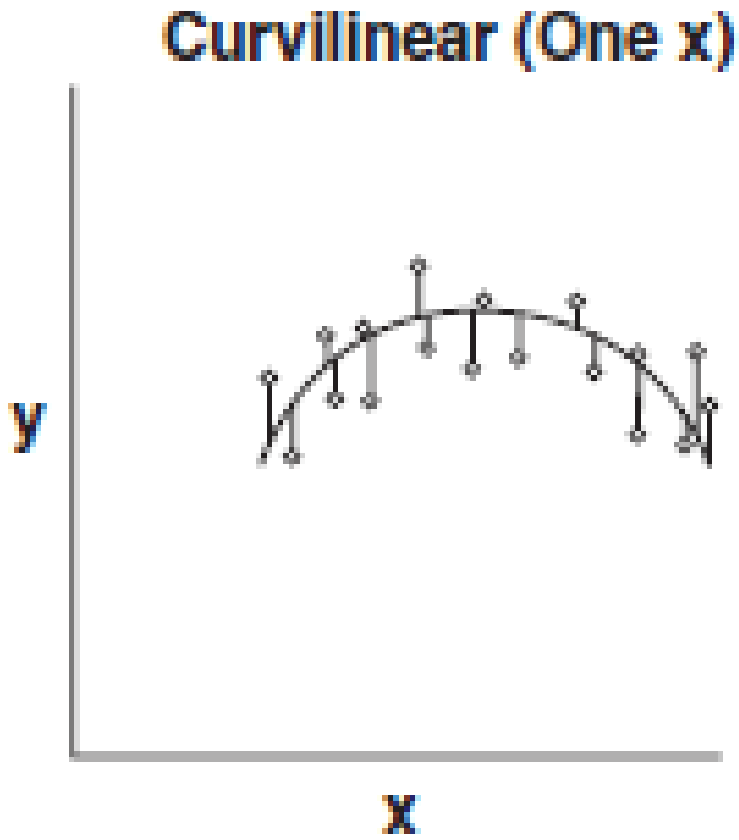


Simple linear regression



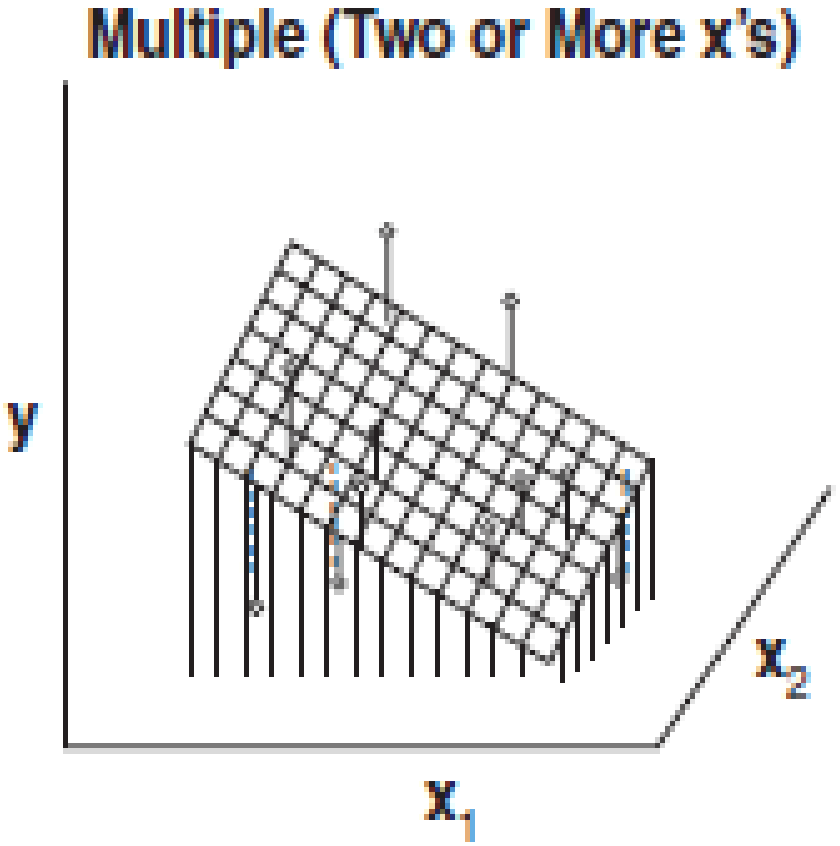
- 단순선형회귀분석
- 단일 X와 단일 Y의 관계를 선형식으로 표현
- 즉, $B_0 + B_1 x$ 을 찾는 것.

Polynomial regression



- 다항회귀분석
- 단일 X와 단일 Y의 관계를 비선형식으로 표현
- 즉, $B_0 + B_1 x^2 + B_2 x$ 을 찾는 것.
- 단순선형회귀분석보다 유연하게 표현가능.

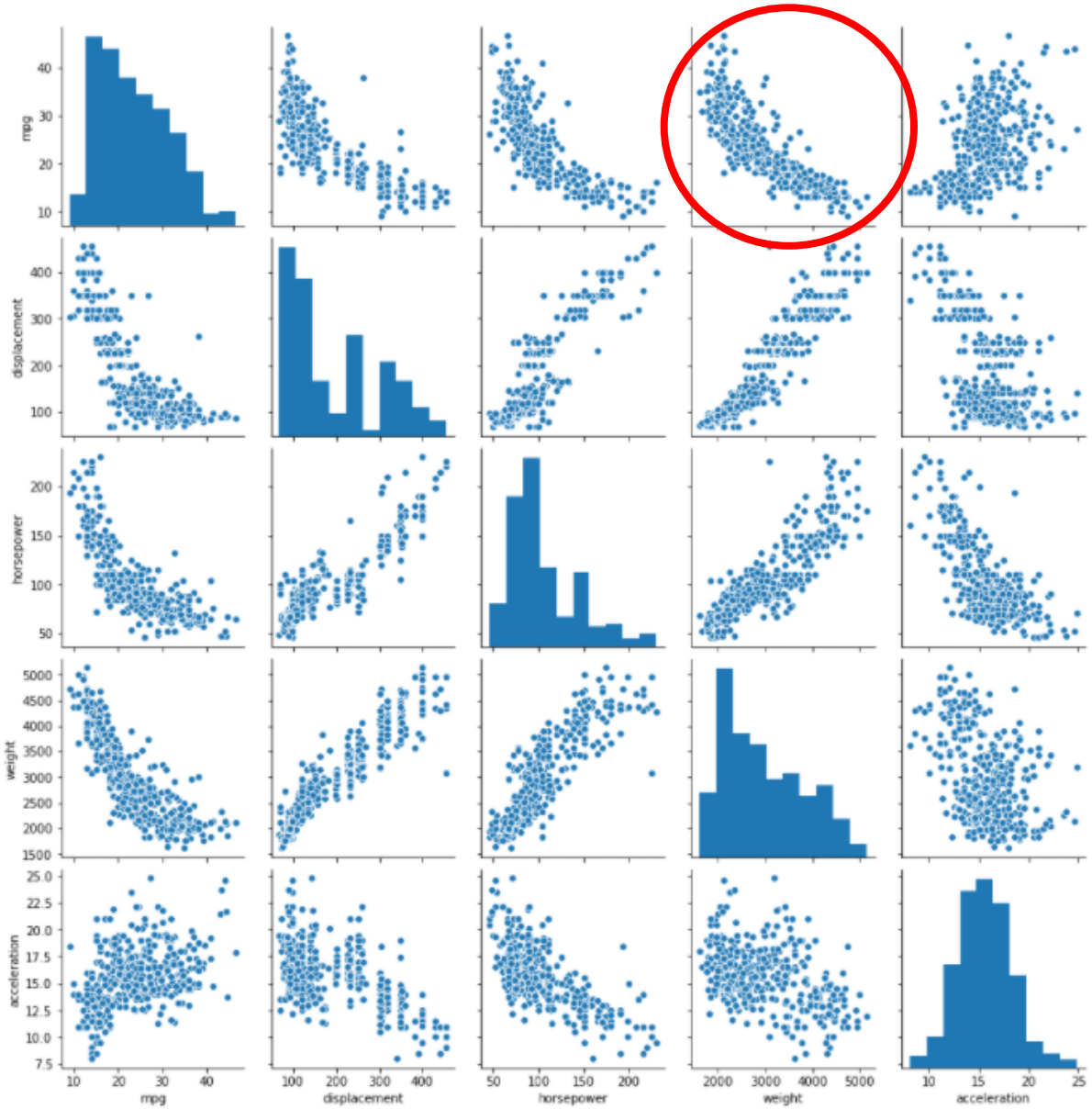
Multiple linear regression



- 다중선형회귀분석
- 다수의 X와 단일 Y의 관계를 선형식으로 표현
- 즉, $B_0 + B_1 x_1 + \cdots + B_n x_n$ 을 찾는 것.
- 실제의 현상을 표현을 하기 좋다.
- 대부분의 회귀분석에서 사용.

《 Round 9 》

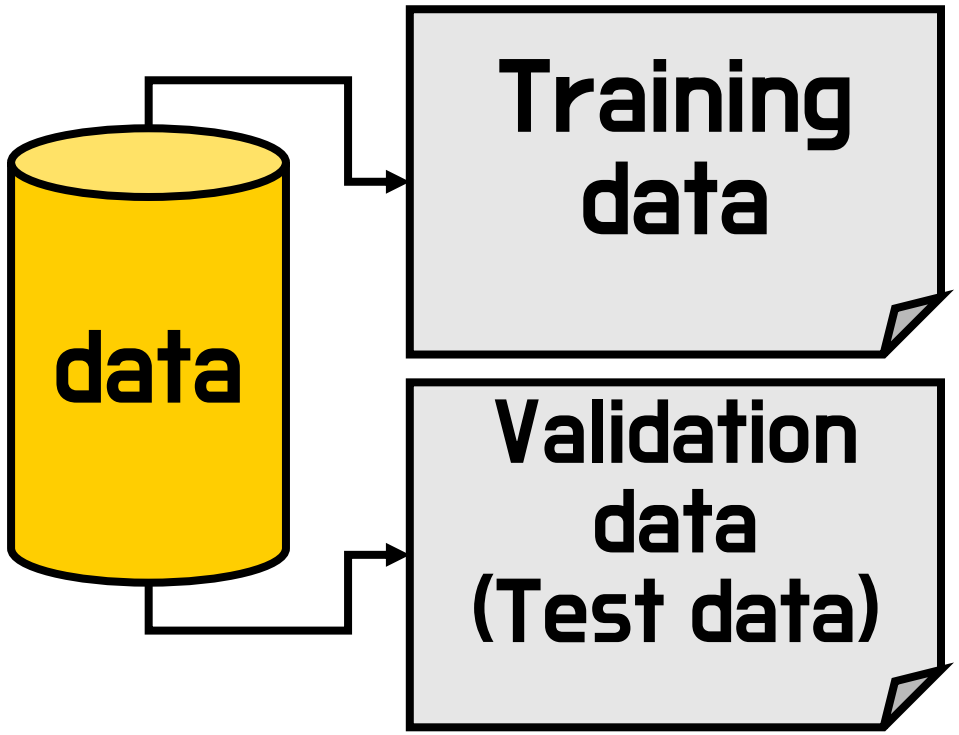
- 모델링
- 회귀분석
- 회귀분석 실습 《



source code :

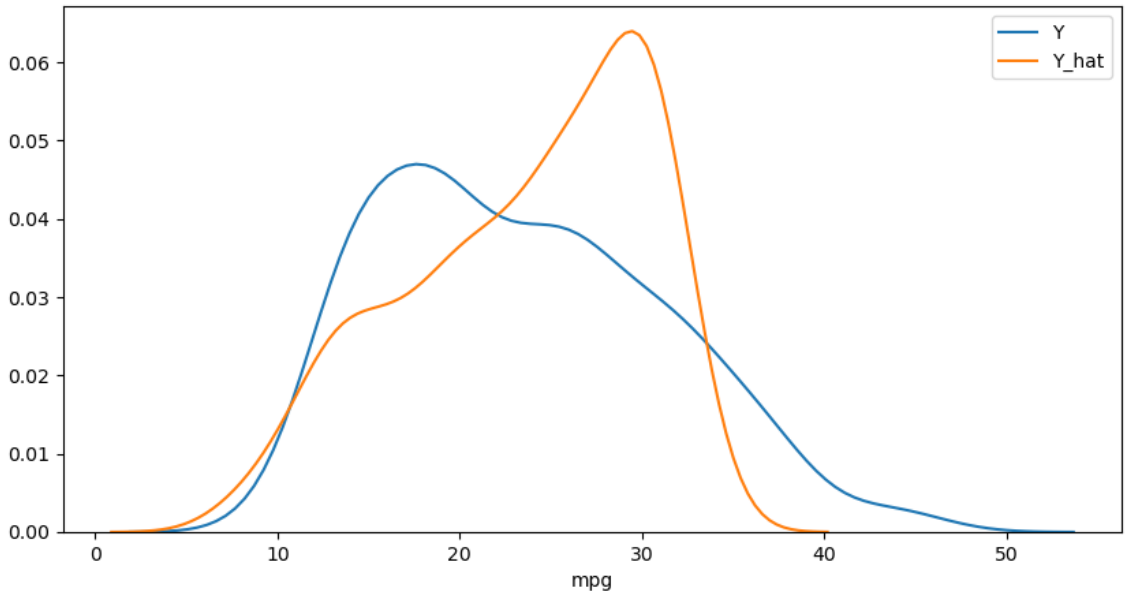
https://github.com/koptimizer/Python_Breakers/blob/master/season2/code/regSample.ipynb

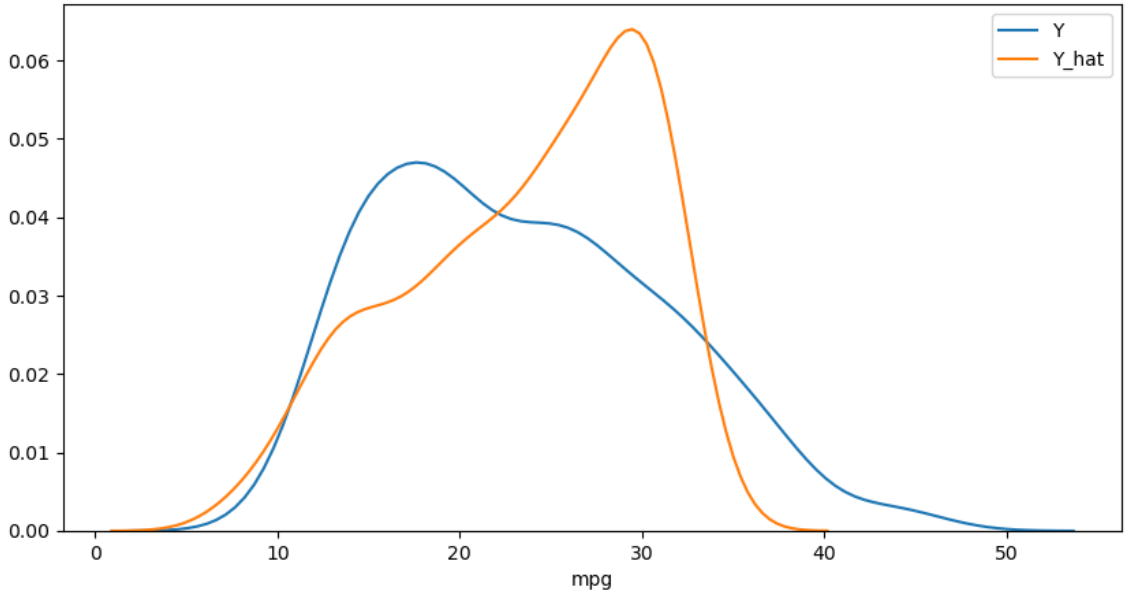
```python
## mpg와 horsepower, weight가 선형관계임이 확인 weight을 x, mpg를 y로 선택
# 속성(변수) 선택
X = ndf[['weight']]  #독립 변수 X
Y = ndf['mpg']       #종속 변수 Y


## dataset을 training data와 test data로 분할
# train_test_split(독립변수, 종속변수, test data 사이즈(%), 랜덤 추출 시드값)
from sklearn.model_selection import train_test_split
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.3, random_state=10)

print('train data 개수: ', len(X_train))
print('test data 개수: ', len(X_test),"\n")
```



```python
# sklearn 라이브러리에서 선형회귀분석 모듈 가져오기
from sklearn.linear_model import LinearRegression

# 단순회귀분석 모형 객체 생성
lr = LinearRegression()

## 학습 시작
# train data를 가지고 모형 학습
lr.fit(X_train, Y_train)

# 학습을 마친 모형에 test data를 적용하여 결정계수(R^2) 계산
r_square = lr.score(X_test, Y_test)

# 회귀식과 결정계수(R^2) 산출
print('회귀식 :', float(lr.coef_),'X +',lr.intercept_)
print('결정계수(R^2) :', r_square)
print('\n')

# 모형에 전체 X 데이터를 입력하여 예측한 값 y_hat을 실제 값 y와 비교
Y_hat = lr.predict(X)

plt.figure(figsize=(10, 5))
ax1 = sns.distplot(Y, hist=False, label="Y")
ax2 = sns.distplot(Y_hat, hist=False, label="Y_hat", ax=ax1)
plt.show()
plt.close()
```

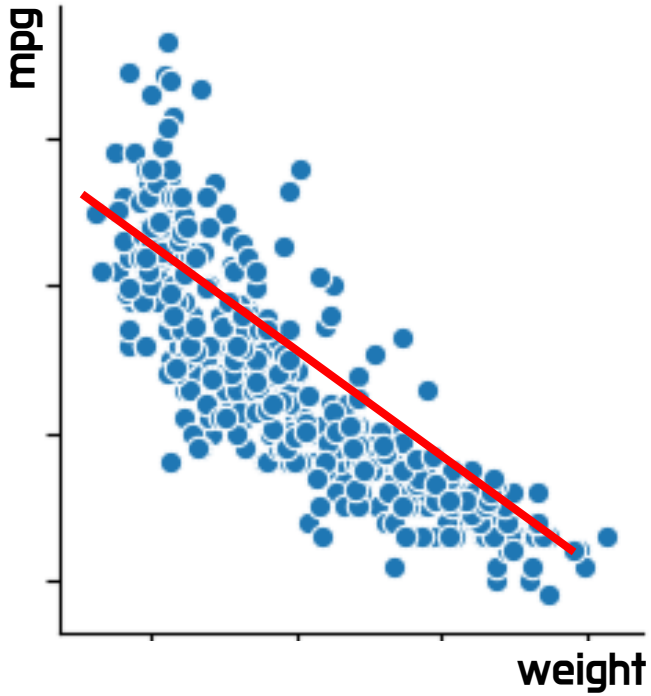
```
train data 개수:  274
test data 개수:  118

회귀식 : -0.007753431671236769 X + 46.7103662572801
결정계수(R^2) : 0.6822458558299325
```





뭐야??? 머신러닝 좋다면서!!

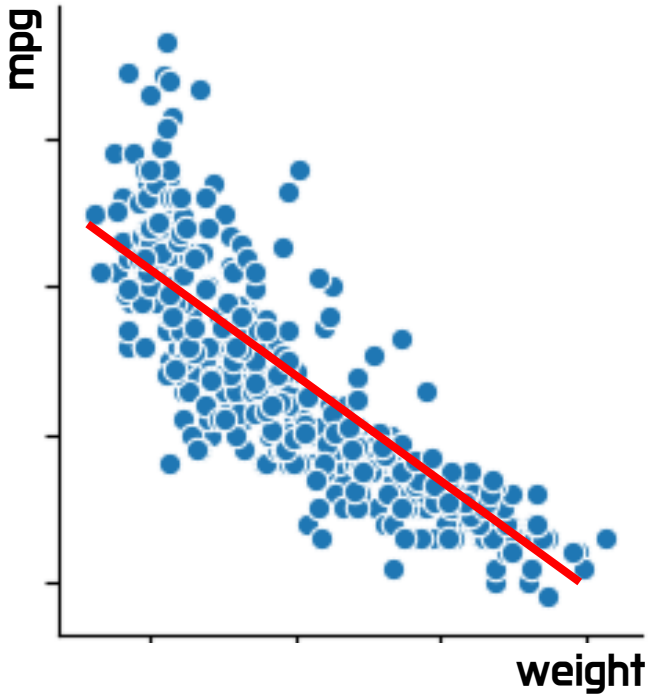
단순선형회귀분석은 정말 단순하다



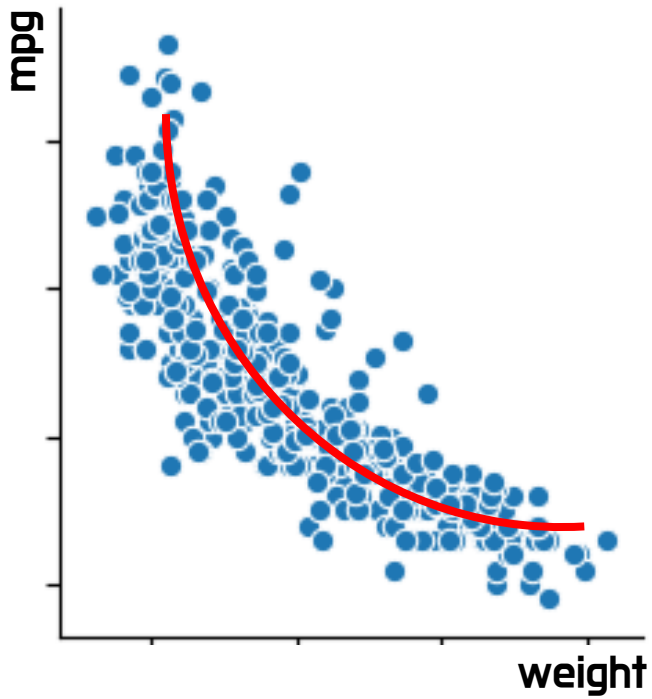
$Y = \beta 0 + \beta 1X$

두 변수 간의 관계를 직선으로 설명했기 때문에 정확도가 낮을 수 밖에 없다.

다항회귀분석



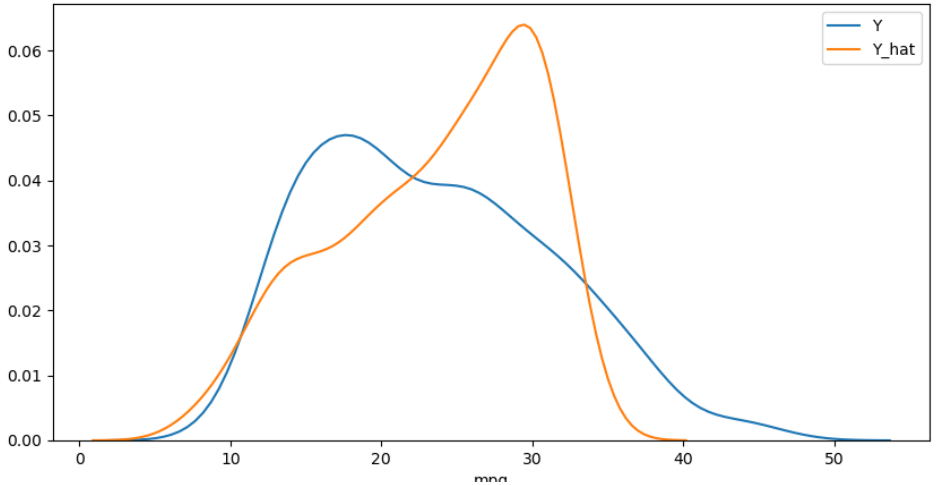
$Y = \beta 0 + \beta 1X$



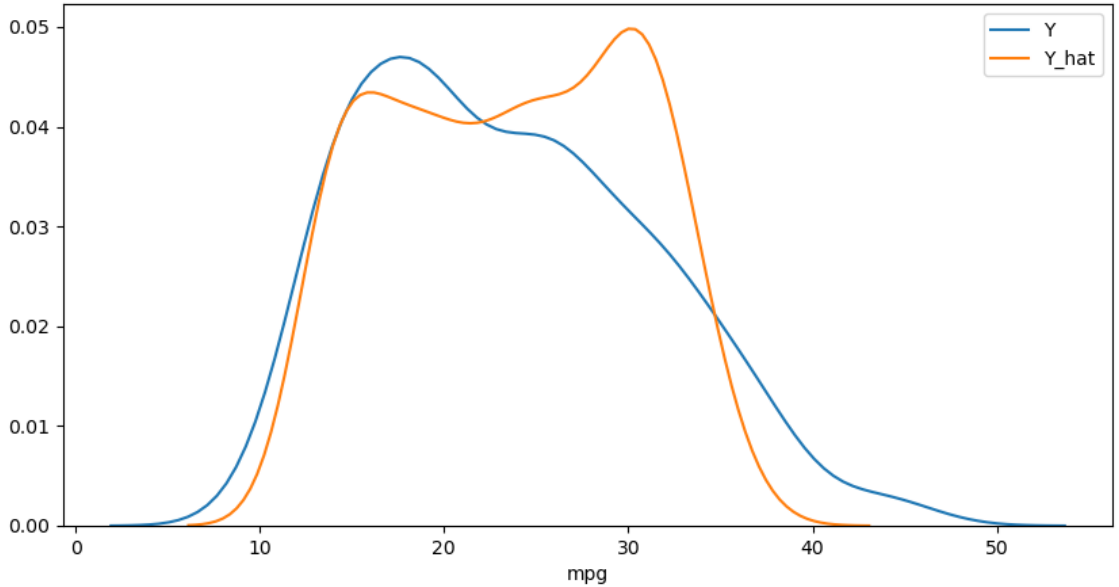
$Y = \beta 0 + \beta 1X^2 + B2X$

두 변수 간의 관계를 곡선으로 설명한다면 정확도를 높일 수 있지 않을까?

단순회귀분석과 다항회귀분석 결과비교

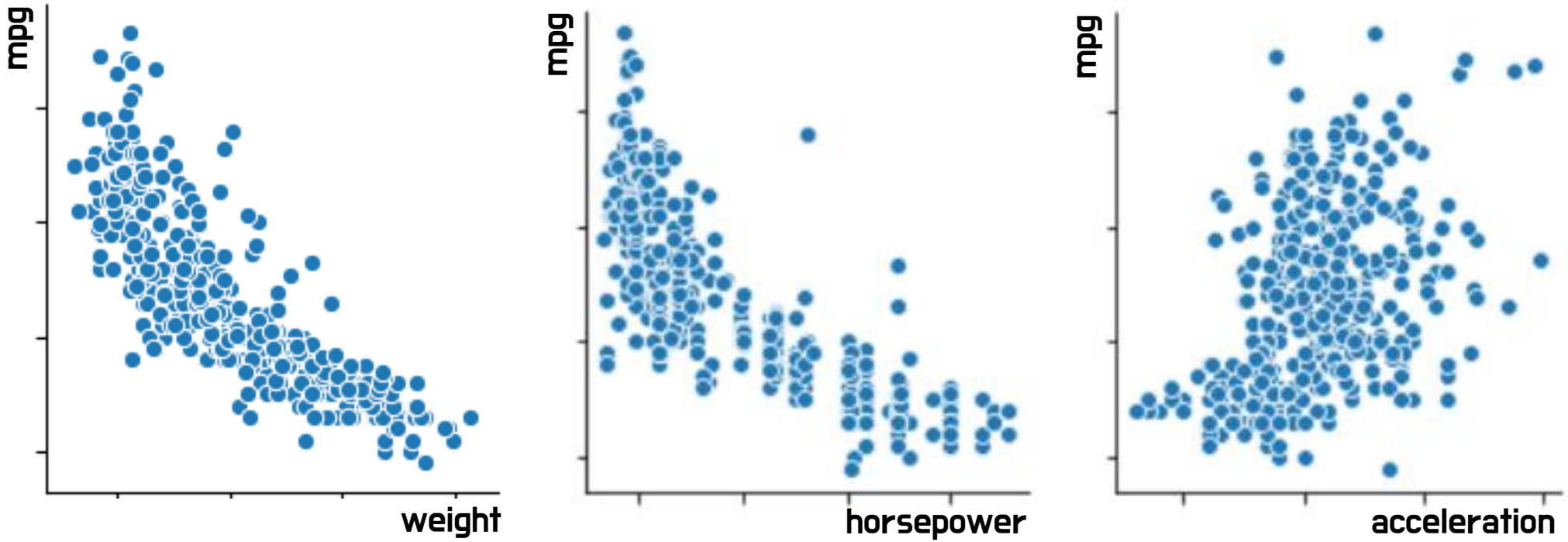


회귀식 : -0.007655397189267713 X + 46.60365052224634
결정계수(R^2) : 0.689363809315209



회귀식 : -0.016911418141332478 X^2 + 1.4345111388654186e-06 X + 60.405921782601645
결정계수(R^2) : 0.72554701541758

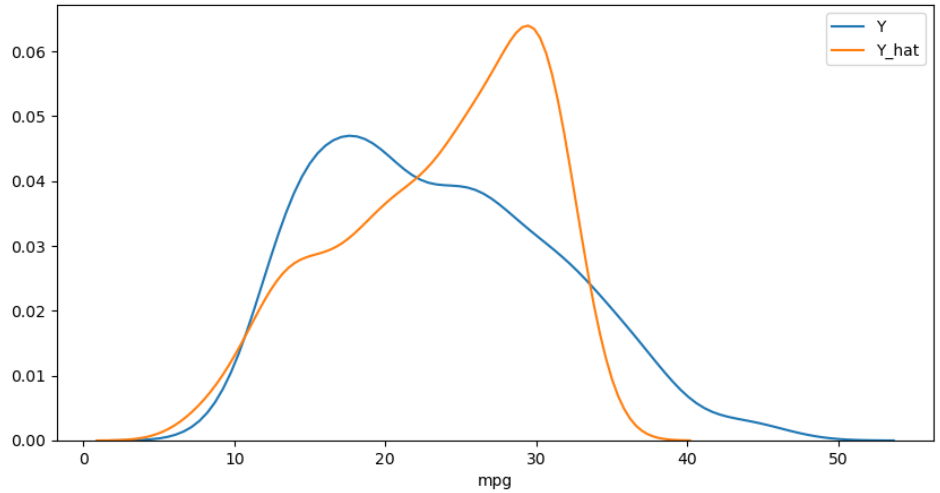
다중회귀분석

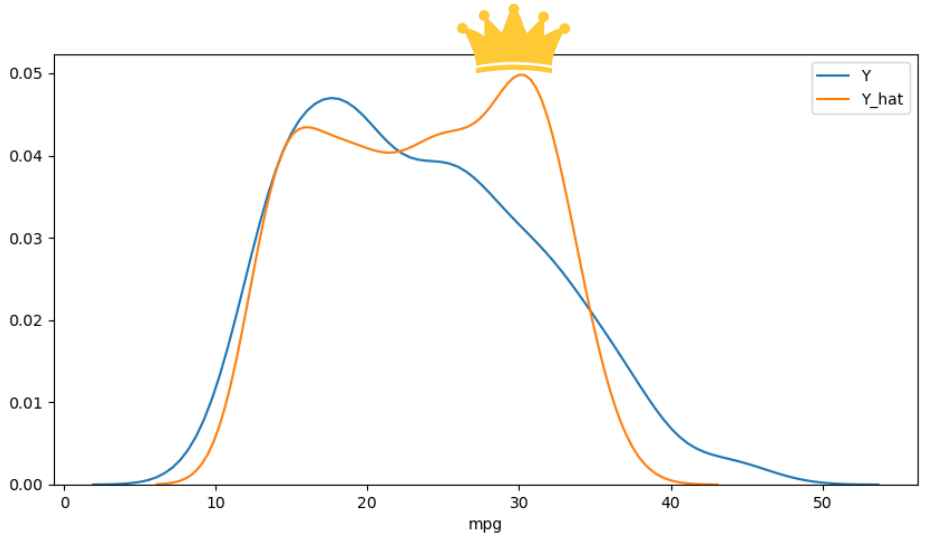


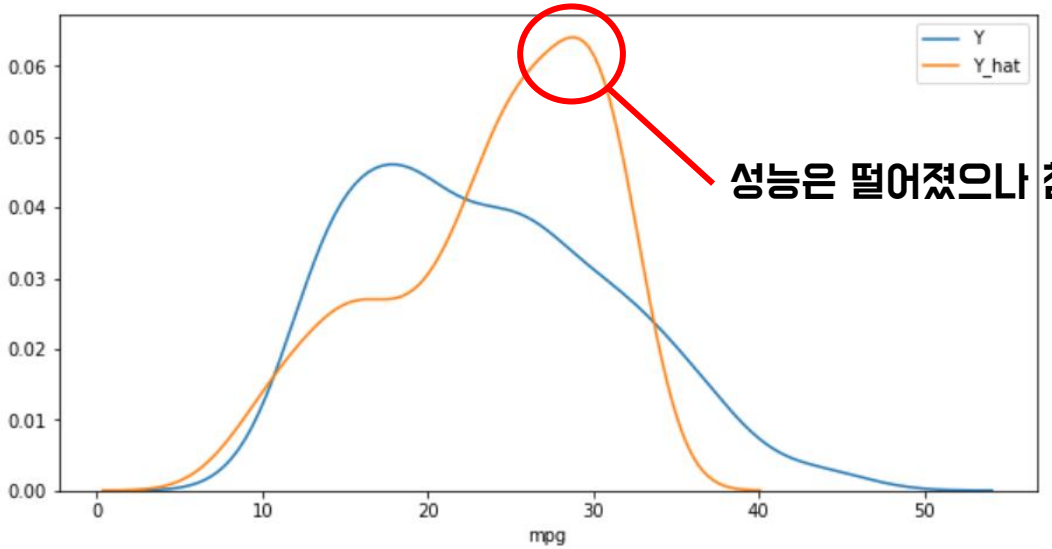
$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3 \ldots$$

여러 독립변수로 현상을 설명하려 한다면 정확도를 더 높일 수 있지 않을까?

단순회귀분석 VS 다항회귀분석 VS 다중회귀분석

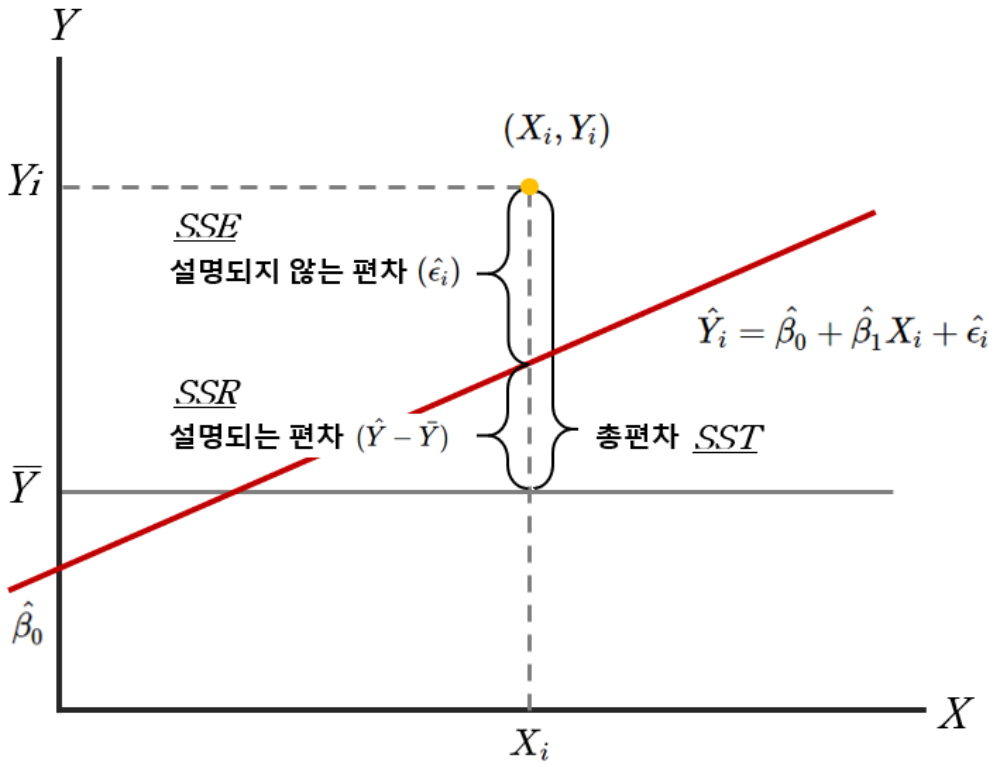






성능은 떨어졌으나 첨도가 개선되었다

결정계수(R Squere) 해석



- Model의 설명력, R^2 = 0.7 = 70%만큼 설명한다.
- 설명 가능 편차 / 총 편차 = SSR/SST
- 0 ~ 1사이의 값을 가지며 높을수록 강한 설명력
- Model의 목적에 따라 값이 높더라도 유의미하지 않을 수 있으며, 다른 지표들도 함께 검토해야 한다.

파뿌리
파이썬 뿌시는 이십대들!!

NEXT STAGE