Simple Linear Regression

1. Linear Regression

연속형 변수 y를 x를 통해 예측

1. Logistic Regression (카테고리 예측)

ex) 답변 (yes, no) 질병 예측

공통점: x를 통해 Y 예측

→ 머신러닝을 통한 학습, 분석할 때 사용되는 기술

(Computer Vision, Neural Network etc…)

예측에 이용되는 x값이 하나면, 2차원 평면에 나타냄

x값이 여러 가지 라면?)

multiple linear regression

선형의 경향이 아니라 포물선과 같은 형태?)

polynomial regression

회귀분석?)

x를 이용하여 y를 예측하는 과정

x (독립 변수)

y (종속 변수)

- 종속변수과 독립변수는 인과 관계가 없다!
- 경향성만 확인함 (다른 요인들이 있을 수 있음)
- 상관 관계가 존재한다!

→ 머신러닝에서도 똑같이 상관 관계를 확인하는 용도로 사용됨

<용어 정리>

Terms

- X = independent / predictor variable(s) / feature / 독립변수
- Y = dependent / response variable / target / 종속변수
- β_0 = intercept; value of Y when $X = 0$. * β_0 and β_1 are also called regression coefficient.
- β_1 = slope; change in Y when X changes 1 unit.
- ε_i = random error * Assume that the errors follow normal distribution with mean 0 and unknown σ^2.
- e_i = residual
- n = number of observations
- i = i-th observation

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Terms (Cont'd)

- When there is hat sign, it is our **estimated** value.
  - $\hat{y}_i$ = estimation
  - $\hat{\beta}_0$ = estimated intercept
  - $\hat{\beta}_1$ = estimated slope

- Finding a regression line means finding the parameters β_0 and β_1.

<Simple Linear Regression>

- A model with a single regressor x that has a relationship with a response y that is a **straight line**.
- Linear Model: $Y = \beta_0 + \beta_1 X + \varepsilon$
- Expectation:

$$E(Y|X) = E(\beta_0 + \beta_1 x + \epsilon) = \beta_0 + \beta_1 x \ (\because E(\epsilon) = 0)$$

- Variance

$$Var(Y|X) = Var(\beta_0 + \beta_1 x + \epsilon) = Var(\epsilon) = \sigma^2$$

- When we calculate the expectation and variance, β_0, β_1 and x are regarded as **constants**.

선형 방정식을 만들 때, 관측값 y와 예측값 y hat과의 오차 범위를 최소화하는 예측값을 잡아야 한다.

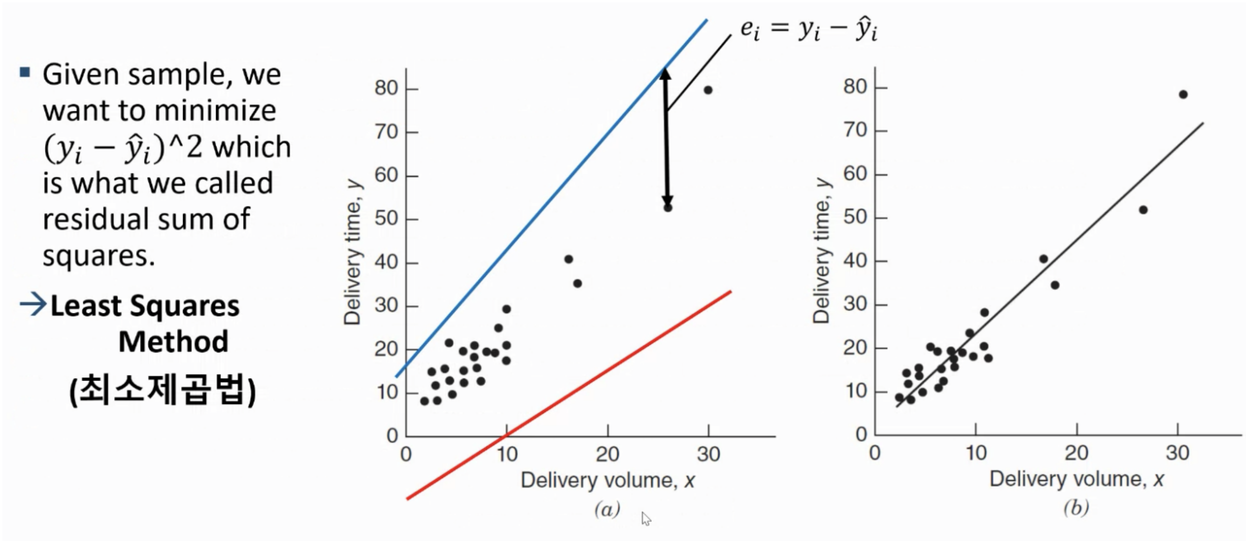
(x1, y1) → (x1, y1 hat) …

(y1 - y1 hat) - (y2 - y2 hat) … (yn - yn hat) 오차

→ 음수와 양수가 나올 수 있음

(y1 - y1 hat)^2 + (y2 - y2 hat)^2 … (yn - yn hat)^2 (최소 제곱법)

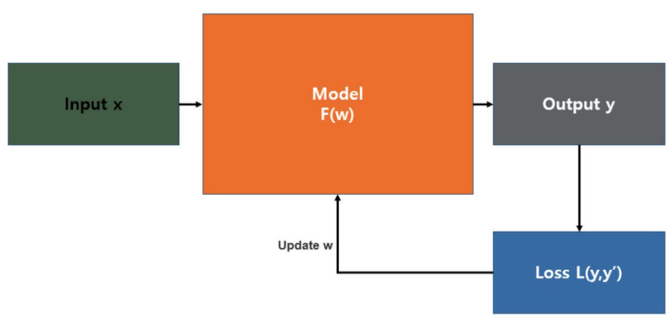
값들을 최소화해서 베타0, 베타1 찾기 위해서 사용 → 다양한 식들이 존재하지만 이 식이 가장 기본적



Least Squares Method (최소제곱법)

- We find the **<u>cost function (or loss function)</u>** using residual sum of squares.

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x))^2$$



cost function을 최소화 하는 것이 가장 중요 ! (최소화하는 베타0와 베타1)

Least Squares Method (최소제곱법)

- Using partial derivative,
  - $\frac{\partial}{\partial \beta_0} S(\beta_0, \beta_1) = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0$
  - $\frac{\partial}{\partial \beta_1} S(\beta_0, \beta_1) = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) x_i = 0$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{1}{n} \left(\sum_{i=1}^{n} y_i \right) \left(\sum_{i=1}^{n} x_i \right)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right)^2}$$

- Therefore, $\hat{\beta}_0$ and $\hat{\beta}_1$ are the least-squares estimators of the intercept and the slope, respectively.
- Special case of **<u>gradient descent</u> (we will study later on)**.

<Applications>

Applications 1. Data Analysis (survey, experiments)

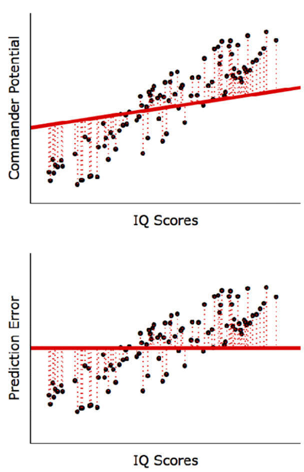
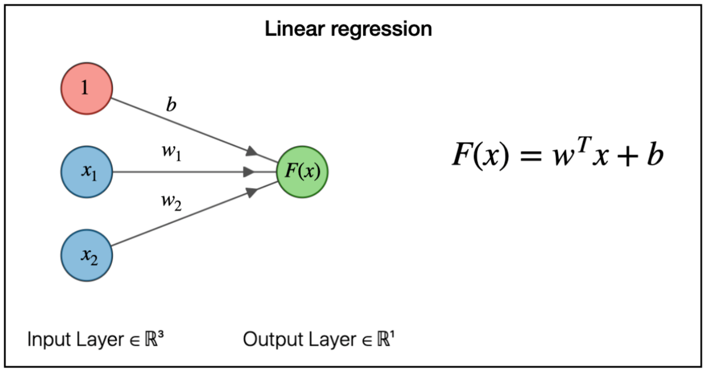


표 22. 스포츠중계시청에 미치는 영향요인

	비표준화 계수		표준화 계수	t	유의확률
	B	표준오차	베타		
상수항	1.634	0.295		5.536	0.000
배구토토 참여경험	0.175	0.048	0.244	3.678	0.000
동료친지와 즐김	0.179	0.064	0.180	2.771	0.006
취미와 여가생활	0.165	0.057	0.191	2.872	0.005
씨름관람경험	-0.191	0.048	-0.247	-3.962	0.000
주변사람의 긍정적시선	0.165	0.068	0.162	2.434	0.016
배당률에 의한 종목선택	0.115	0.053	0.138	2.155	0.032
종속변수: 문항 8-6					

R^2=0.865, 수정된 R^2= 0.343, F값 변화량: 4.642(p-0.032)

Applications 2. Neural Network



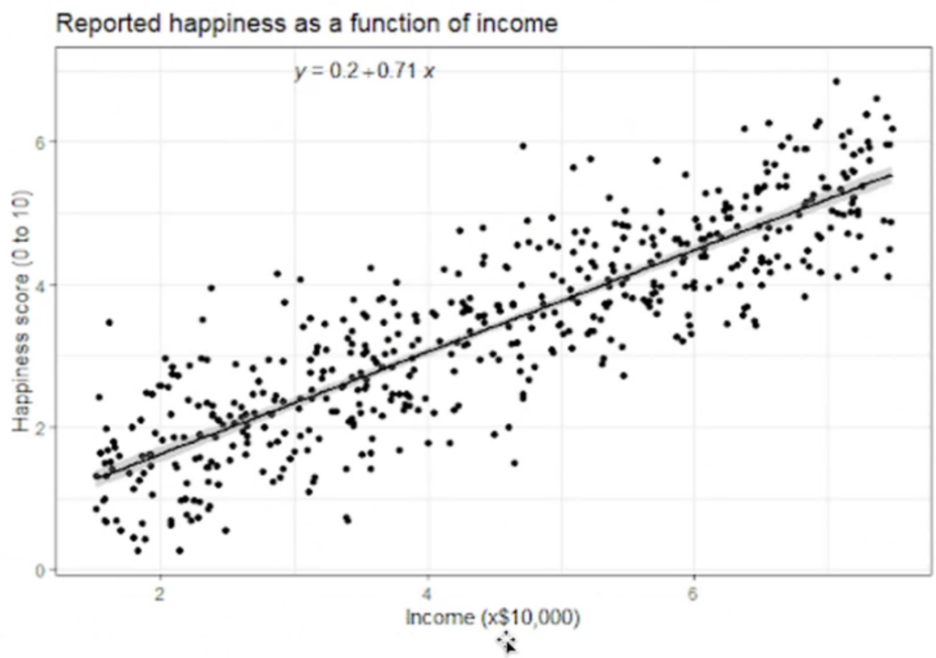
$$F(x) = w_1 \cdot x_1 + w_2 \cdot x_2 + 1 \cdot b$$

<Data Science 에서의 활용>

python library (sklearn)

```
from sklearn.linear_model import LinearRegression

line_fitter = LinearRegression()
line_fitter.fit(X, y)
y_predicted = line_fitter.predict(X)
```



오차 (평균 제곱근 오차)

⇒ RMSE (Root Mean Square) 값이 작을수록 Ideal한 데이터임을 평가