Regression

Regression in Machine Learning

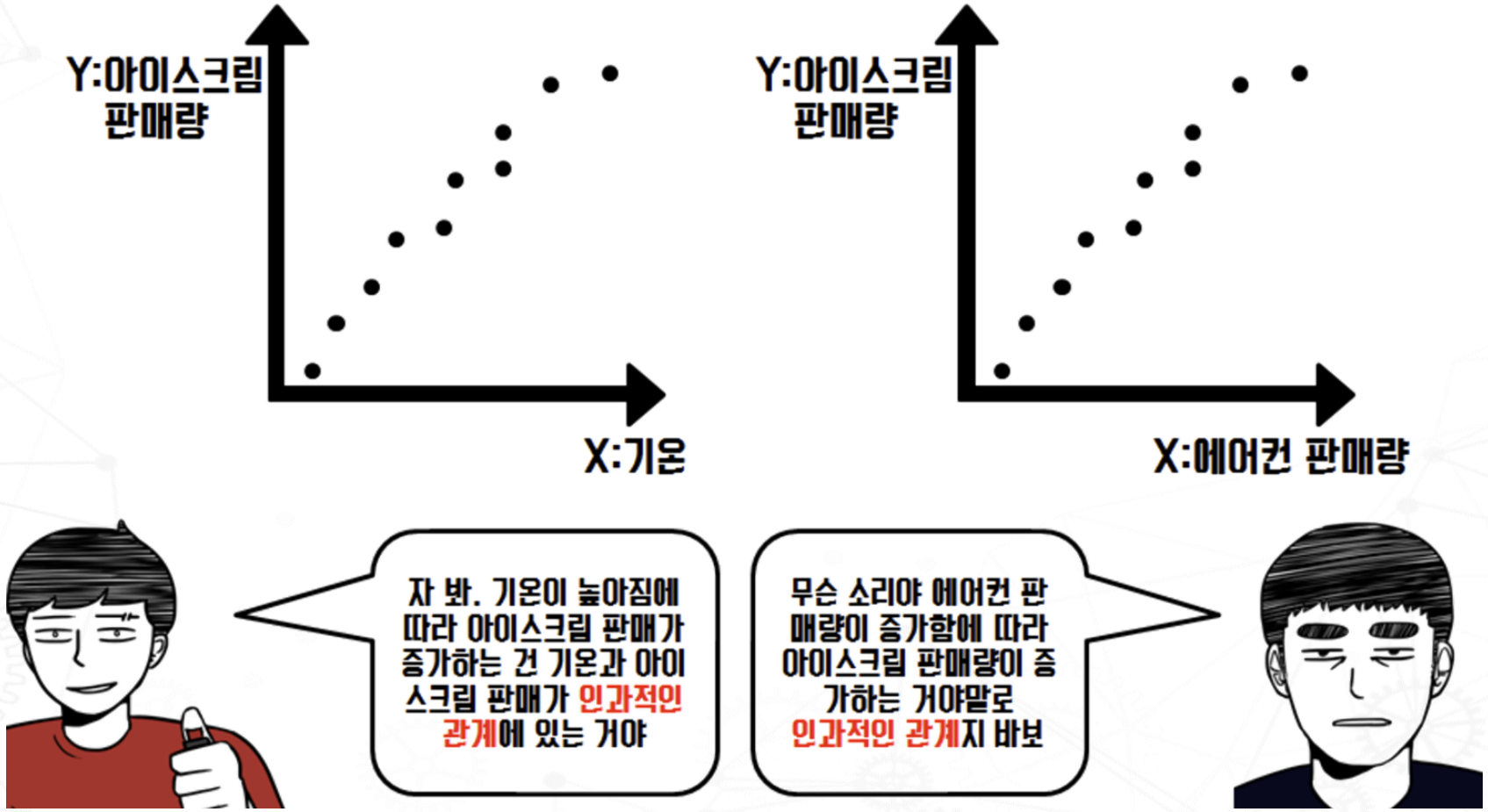
온도	판매량
20	40
21	42
22	44
23	46

양적 → 회귀 regression

공부시간	시험결과
20	불합격
21	불합격
22	합격
23	합격

범주형 → 분류 classification

Causal Relationship vs. Correlation



Covariance

❖ Covariance

- It is a measure of how two probability variables x and y are related to each other and change, indicating the direction of the relationship between the two probability variables

$$Cov(X,Y) = E(X-\mu_X)(Y-\mu_Y)$$
$$= E(XY) - \mu_X\mu_Y$$

$$Cov(aX, bY) = abCox(X, Y)$$

X	1	2	3	4	5
Y	10000	20000	30000	40000	50000

Cov(X,Y)=E{(1−3)(10000−30000)+(2−3)(20000−30000)+(4−3)(40000−30000)+5
−3(50000−30000)} =E(40000+10000+0+10000+40000)=20000

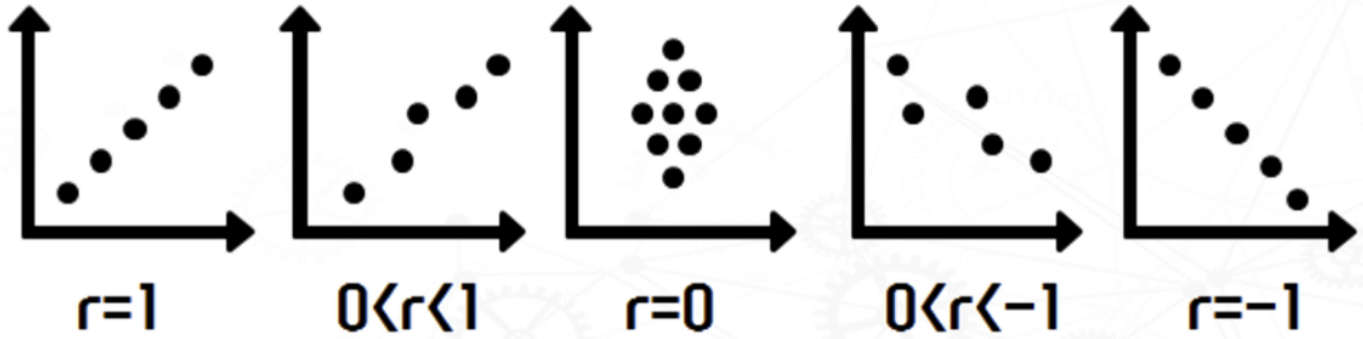
Correlation coefficient

❖ Correlation coefficient

- A numerical value (in coefficients) indicating the degree of correlation between two variables X and Y
- It has a value between -1 and 1, and the closer the absolute value is to 1, the higher the degree of correlation between the two variables

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

$$Corr(aX, bY) = \frac{ab}{|ab|} Corr(X, Y)$$



Regression

❖ Regression

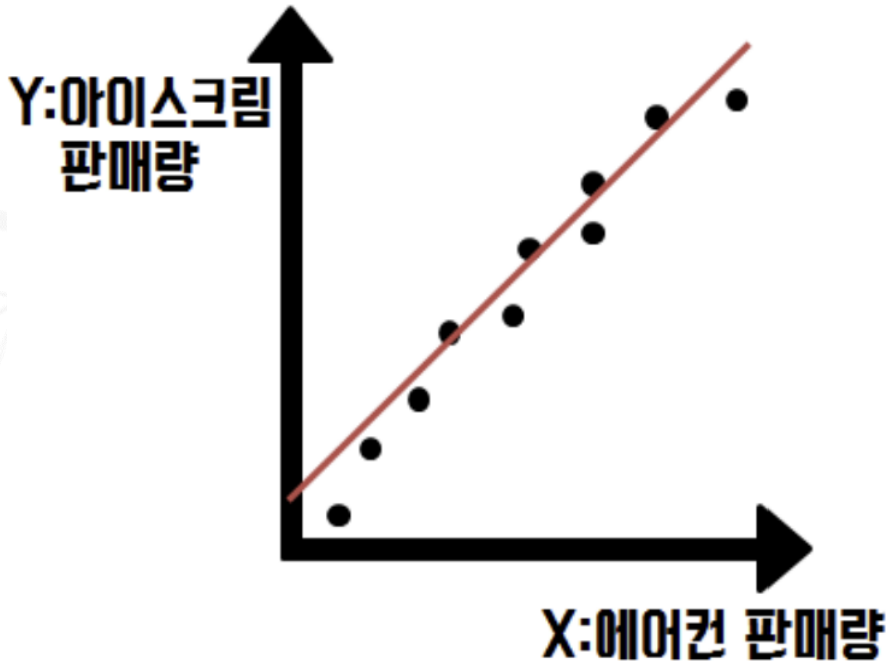
- Analysis method for testing causality
- One or more cause variables (independent variables) affect other variables (dependent variables)
- Independent Variables: Variables that affect dependent variables
- Dependent Variables: Variables affected by other variables

$$X(IV) \rightarrow f(process) \rightarrow Y(DV)$$

Regression

❖ Regression

- A statistical method of assuming a mathematical model and estimating this model from the data of measured variables to determine the relevance of variables
- Predicting the value of a dependent variable based on the value of an independent variable

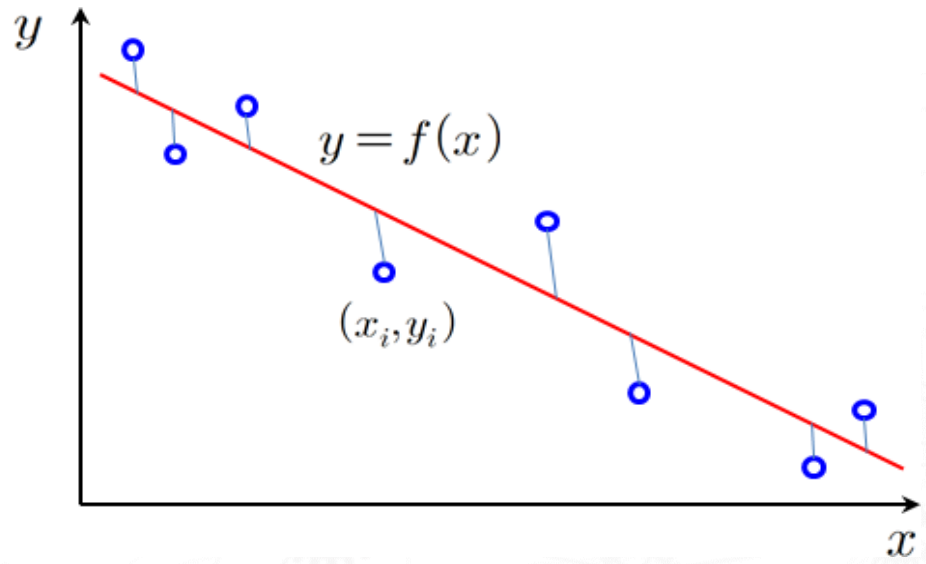


Cov(x, y)=2000000
r=0.8467
Y=0.02+10.67x

Regression

❖ Regression

- An analysis technique that explains the change of dependent variables as a linear combination of independent variables
- A technique for estimating the statistical relationship between dependent and independent variables
- Analyze interrelationships between variables and predict changes in other variables from changes in certain variables

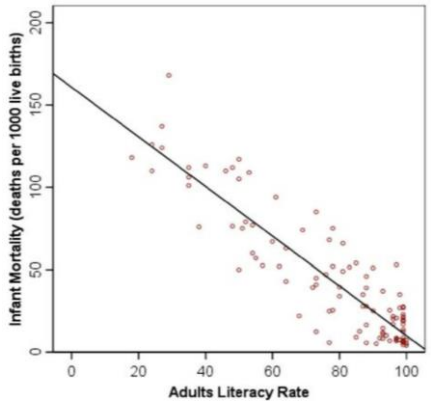


Regression

❖ Simple Regression vs. Multiple Regression

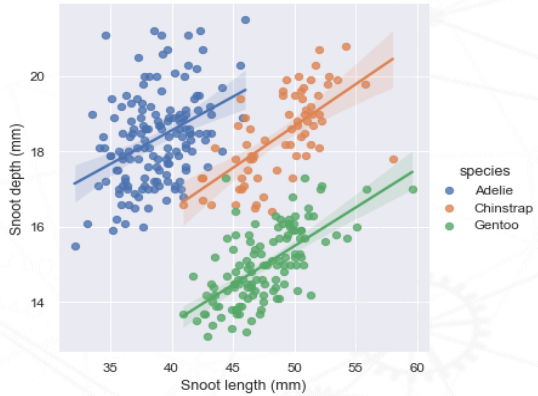
- **Simple Regression :** one independent variable

$$y = a + b x + e$$



- **Multiple Regression :** Two or more independent variables

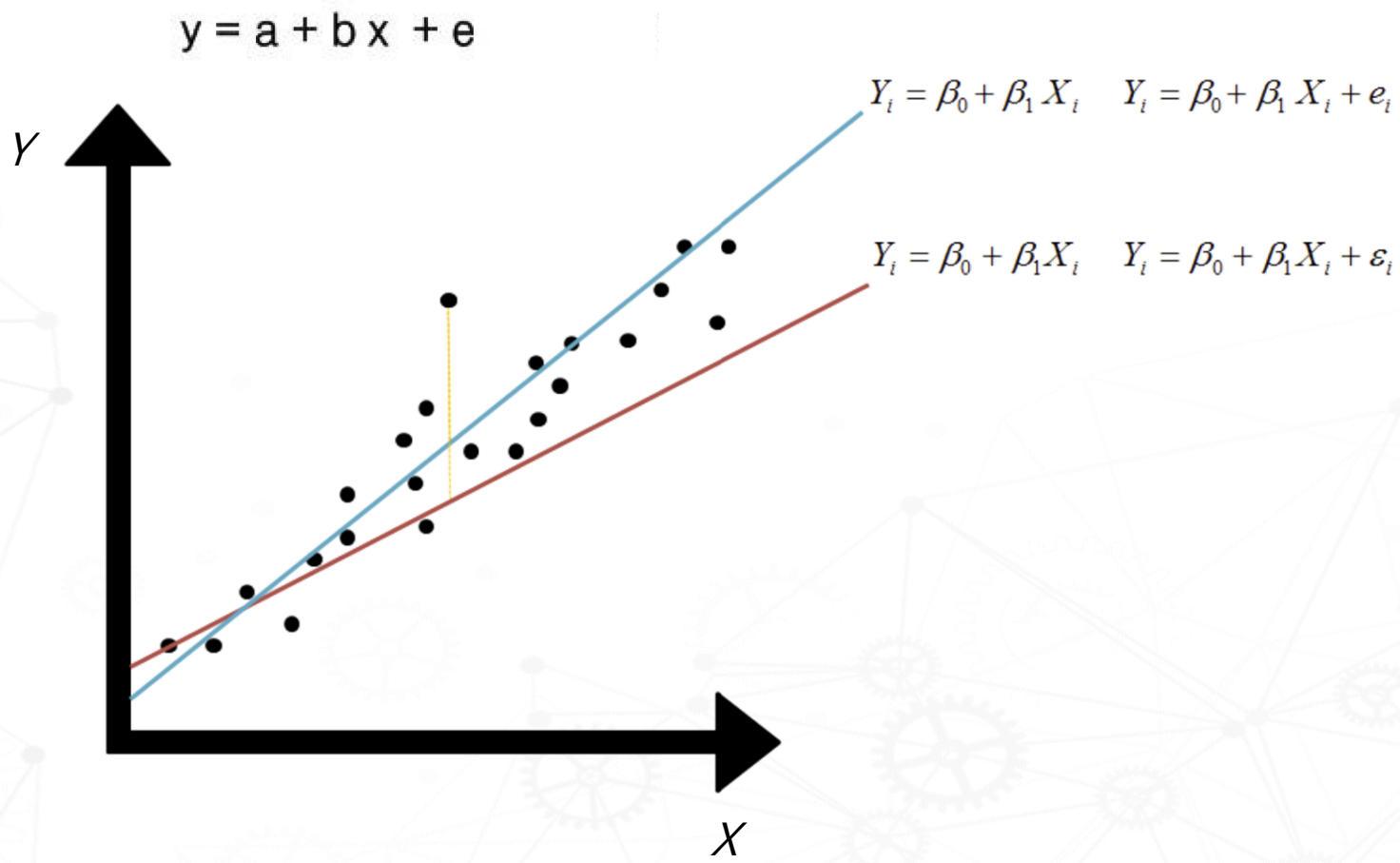
$$y = a + b_1 x_1 + b_2 x_2 + b_n x_n + e$$



- **e (error term) :** Effects of variables other than independent variables on dependent variables

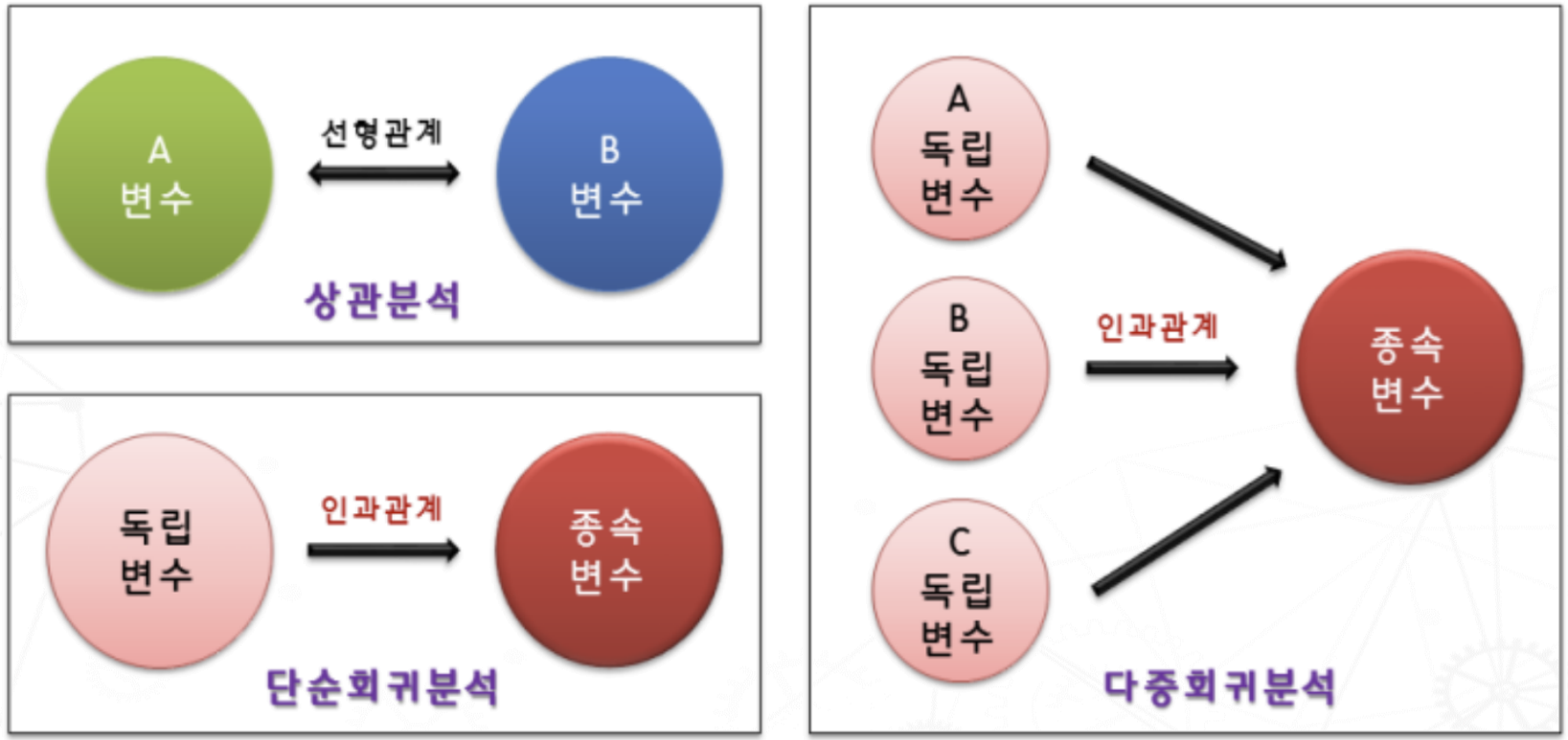
Regression

- Simple Regression vs. Multiple Regression
  - **Simple Regression :** one independent variable



Regression

❖ Simple Regression vs. Multiple Regression



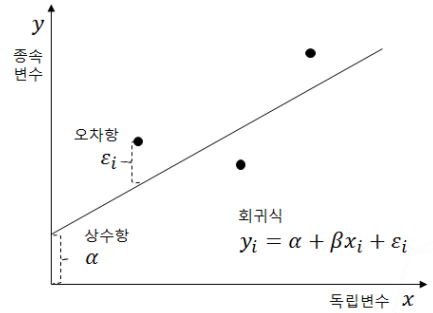
Regression

- The purpose of regression
  - Estimate the value of a dependent variable to the values of an independent variable
  - Review relationships between dependent and independent variables
  - Review the suitability of regression applications
  - Verification of the statistical significance of predictions using regression analysis

Regression

❖ Prerequisites for regression assumptions

- Linearity between independent and dependent variables
- The change in the value of the dependent variable according to the change in the value of the independent variable is constant



❖ Normality and homoskedasticity error term

- 오차항: 종속변수의 관측치와 예측치 간의 차이
- 오차항의 기대값은 0이며, 일정한 분산을 갖는 정규분포를 이룬다고 가정

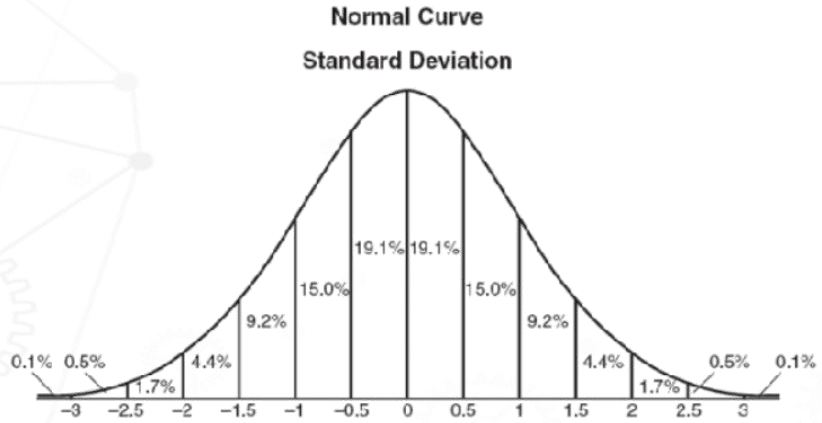
❖ Independency of error term

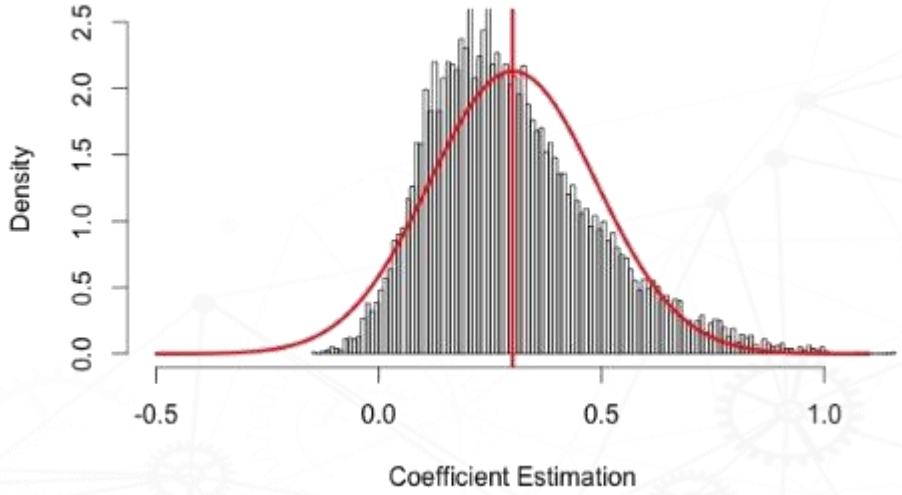
- 예측의 오차값들은 서로 독립적이라는 가정이 필요
- Y의 변화에 따라 오차항이 어떤 패턴을 가져서는 안됨
- Y가 커짐에 따라 오차값이 커지면 가정에 위배

Regression

❖ Normality

- Assume that the distribution of a population in a continuous variable should be established as a normal distribution
- Normal distribution: Symmetric form relative to the center (average) when plotting the percentage of appearance of a particular value

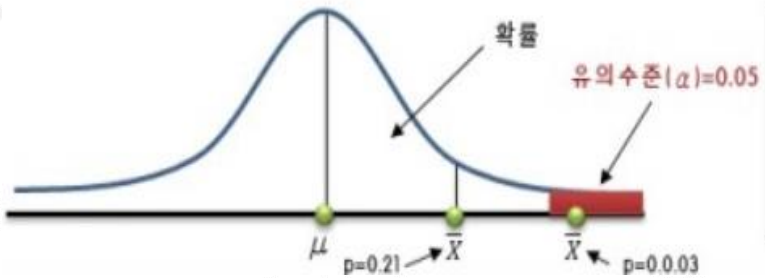




Regression

❖ Normality test

- H0 (**귀무가설**): **표본의 모집단이 정규분포를 이룸**
- H1 (**대립가설**): **표본의 모집단이 정규분포를 이루지 않음**
- **정규성을 만족하기 위해서는 귀무가설을 채택해야 하며, 대립가설을 기각해야함**
  - 유의수준 95% 신뢰구간에서 유의확률 p값이 0.05보다 크게 나타나야 귀무가설을 채택
  - 유의확률(p-value): 귀무가설 하에서 검정통계량의 값이 나타날 가능성을 측정하는 확률값
  - 유의수준(α): 귀무가설이 참인데, 대립가설을 선택하는 오류의 최대 허용 단계를 의미



✓ 유의확률 < 유의수준 → 귀무가설 기각

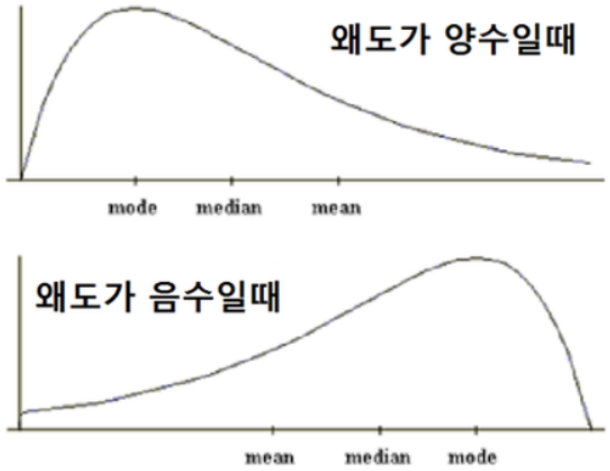
✓ 유의확률 > 유의수준 → 귀무가설 채택

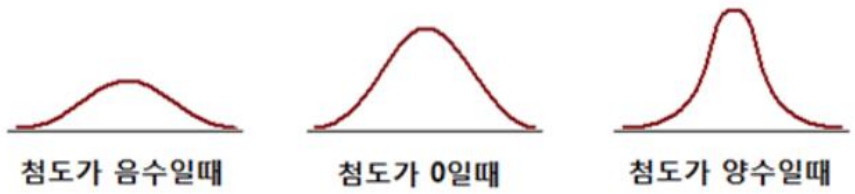
Regression

❖ Normality test

- Central limit theorem
- Kolmogorov-Smirnov & Shapiro-Wilk
- Skewness and kurtosis by Snedecor & Cochran

-

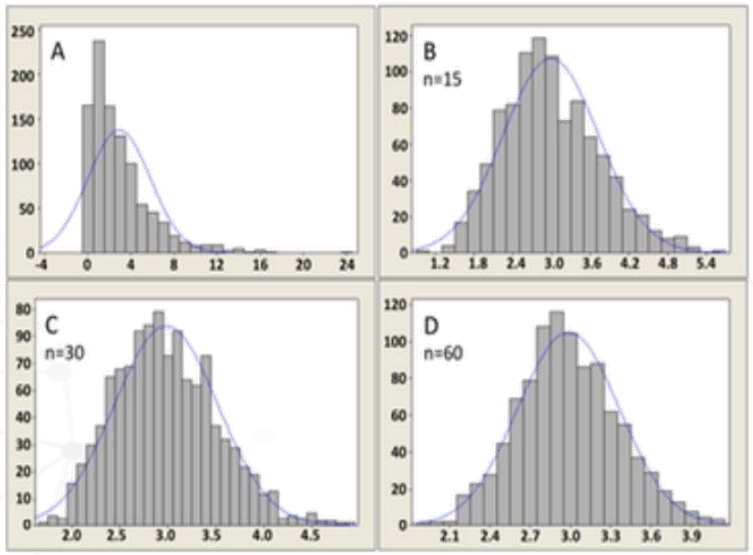




Regression

❖ Normality test

- Central limit theorem

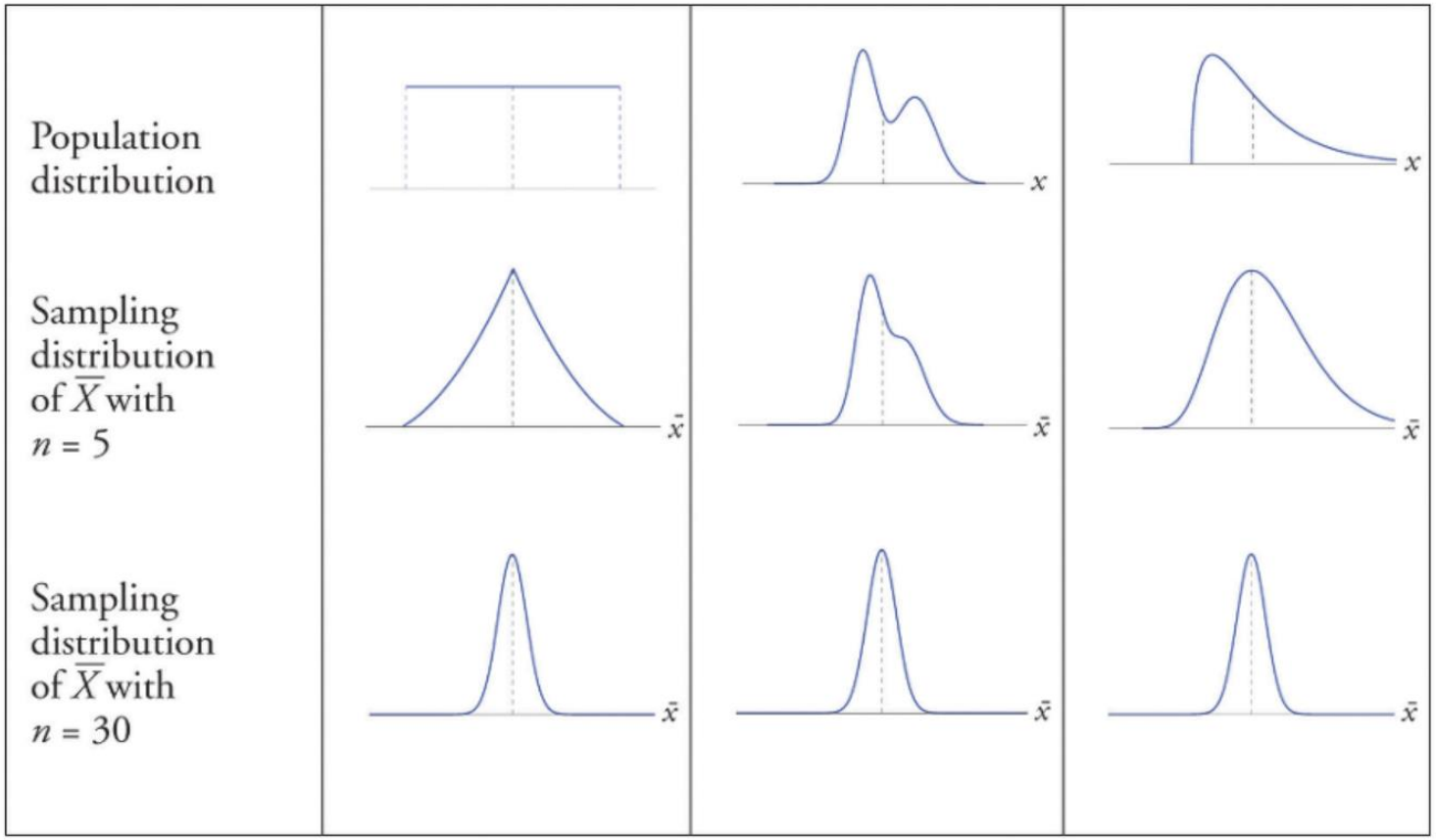


- A 는 모집단의 분포를 나타냄(양의 왜도)
- B-C-D 는 샘플링 접근법 통해 계산된 그래프로서 표본 수가 증가할 수록 정규 분포에 수렴됨

- 표본의 크기가 커질수록 표본 평균의 분포는 모집단의 분포 모양과는 관계 없이 정규 분포에 근접
- 표본 평균의 평균은 모집단의 모 평균과 동일
- 표본 평균의 표준 편차는 모집단의 모 표준 편차를 표본 크기의 제곱근으로 나눈 것
- Data가 대표본(일반적으로, 범주 별 **30개 이상**)이면 대표본 근사(중심극한정리)에 의해 정규성을 만족

Regression

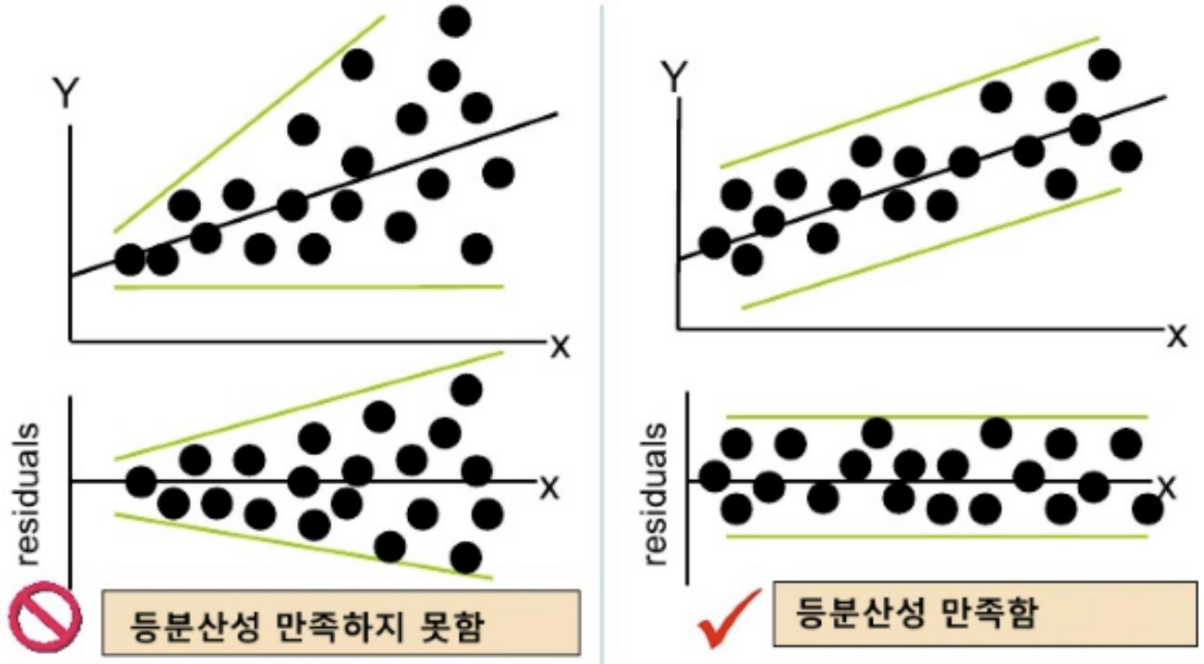
- Normality test
  - Central limit theorem



Regression

❖ Homoskedasticity

- The condition that two or more different groups must satisfy the same variance through ANOVA
  - The data should evaluate for normality beforehand



Regression

❖ Basic principle

$y = a + bx + e$ (회귀식)

$\hat{y} = \hat{a} + \hat{b}x$ (추정식) → $\hat{a}$과 $\hat{b}$ 추정

$e = y - \hat{y}$ (관찰치와 예측치의 차이)

❖ Definition of e (error)

TSS (Total Sum of Square) : $\sum(y - \hat{y})^2$
실제치(y) 와 추정치($\hat{y}$)의 차이의 제곱의 합

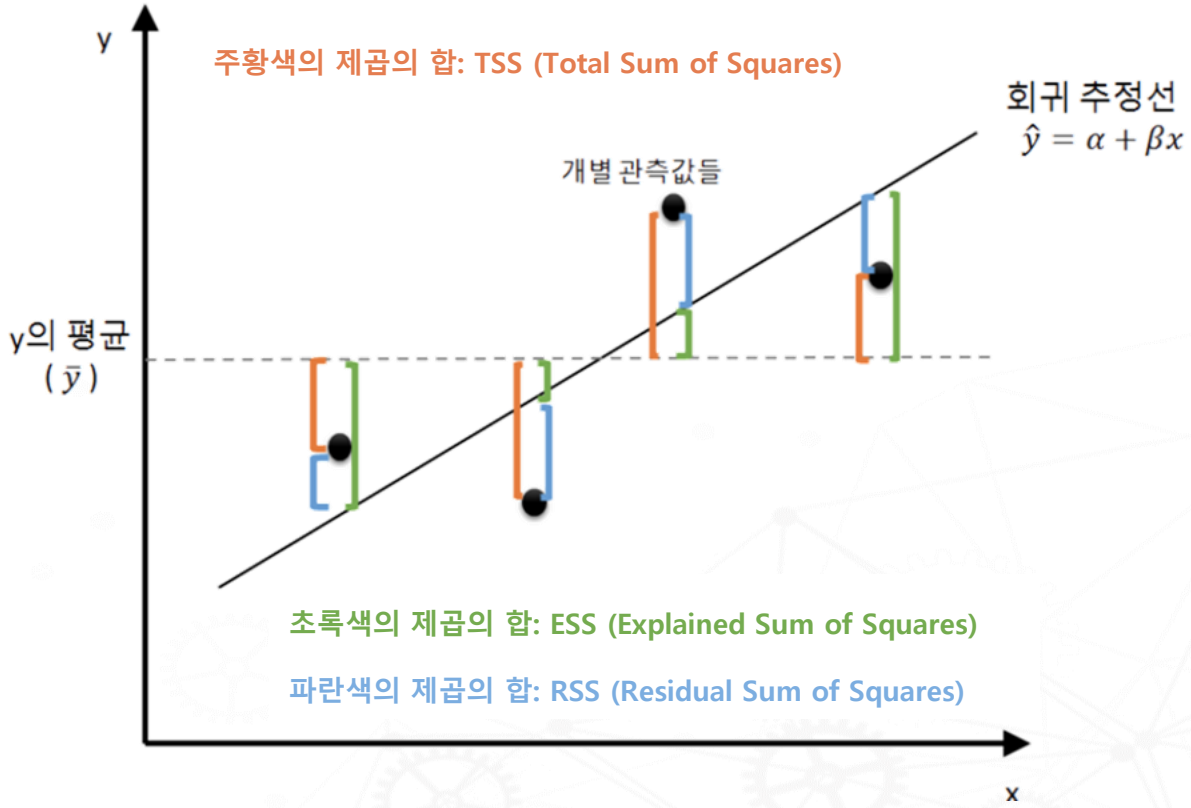
RSS (Residual Sum of Square) : $\sum(y - \bar{y})^2$
실제치(y)와 y의 평균($\bar{y}$)와의 차이의 제곱의 합
Unexplained Error (회귀선으로 설명이 안 되는 분산)

ESS (Explained Sum of Square) : $\sum(\bar{y} - \hat{y})^2$
y의 평균($\bar{y}$)와 추정치($\hat{y}$)의 차이의 제곱의 합

Regression

❖ Least square method (Ordinary Least Square: OLS)

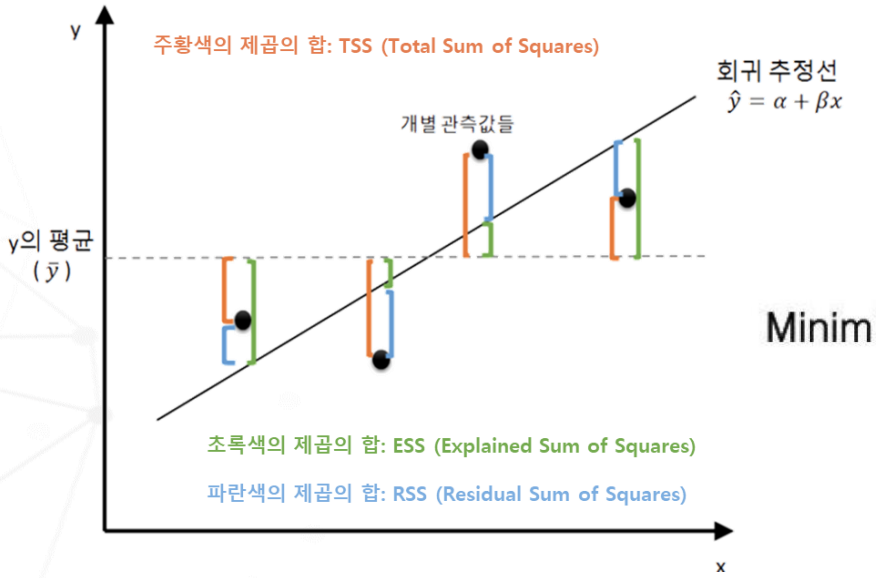
- The least squares method that minimizes the sum of the squares of the errors



Regression

❖ Least square method (Ordinary Least Square: OLS)

- The least squares method that minimizes the sum of the squares of the errors



Minimize $\sum e_i = \sum [y_i - (a\hat{} + b\hat{} x)]^2$

$e = y - y\hat{} = (y - y) + (y - y\hat{})$

$\sum (y - y\hat{})^2 = \sum (y - y)^2 + \sum (y - y\hat{})^2$

TSS = RSS+ESS

총 변동 = 설명된 변동 + 설명 안된 변동

Regression

❖ Approach

- Simultaneous (or direct) method: “Enter”
  - Derive a regression model from a complete set of independent variables

- Stepwise Method: “Stepwise”
  - A method of sequentially including independent variables one by one in the regression model based on the explanatory power of each independent variable

❖ Goodness of Fit

- Multiple R (correlation coefficient)
  - 종속변수와 독립변수의 상관관계
  - 두 변수의 상관성을 나타내는 척도

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

회귀선에 의해 설명되는 변동

$$= \frac{\sum(\bar{y} - \hat{y})^2}{\sum(y - \hat{y})^2} = 1 - \frac{\sum(\bar{y} - \hat{y})^2}{\sum(\hat{y} - \hat{y})^2}$$

전체 변동

- 결정계수 R^2 (coefficient of determination)
  - 상관계수의 제곱
  - X와 Y간의 상관관계가 클수록 R^2는 1에 가까워짐
  - 회귀식이 자료를 얼마나 잘 설명하고 있는가를 나타냄
  - 일반적으로 $R^2 > 0.65$일 경우 회귀식이 잘 설명한다고 판단

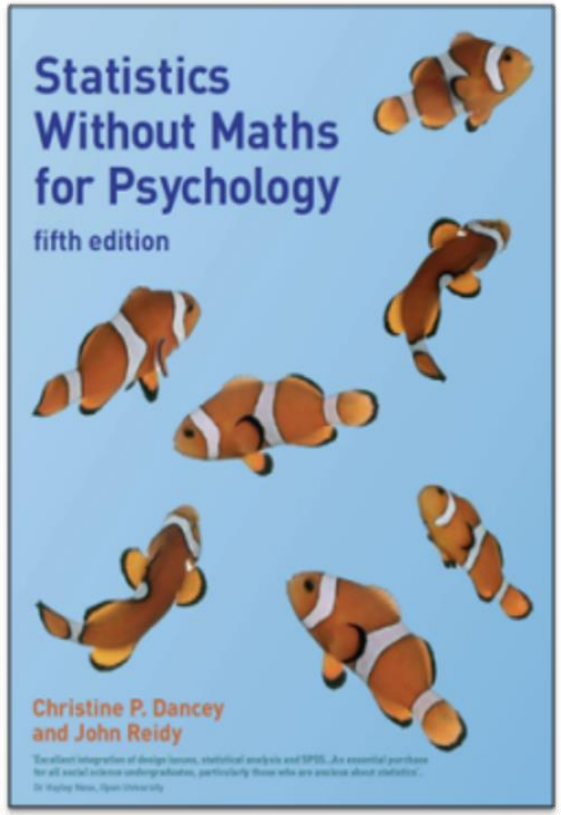
- 수정된 결정계수 (Adjust R^2)
  - 독립변수의 수와 데이터 수를 고려한 결정계수
  - 변수의 수가 증가할수록 결정계수가 높아지는 단점 존재
  - 다중회귀분석에서는 주로 사용하고, 표본의 크기와 독립변수의 수를 고려하여 계산

Regression

❖ Goodness of Fit

Perfect	+1	-1
Strong	+0.9, +0.8, +0.7	-0.9, -0.8, -0.7
Moderate	+0.6, +0.5, +0.4	-0.6, -0.5, -0.4
Weak	+0.3, +0.2, +0.1	-0.3, -0.2, -0.1
Zero	0	

Christine Dancey and John Reidy, Statistics Without Maths for Psychology, p. 175. Prentice Hall, 5th edition, 2011.



Linear Regression

❖ Linear regression

- Model the linear correlation between dependent variable Y and one or more independent variable (or explanatory variable) X
- When the parameters of the regression are linear
- Regression coefficients are optimized to minimize RSS (Residual Sum of Squares) between predicted and actual values, and regulations are not applied