# ADsP 특강

## - 단순선형회귀분석 -

조상구
빅데이터과 경복대학교

# 회귀분석(Regression analysis)


https://www.kaggle.com/code/ryanholbrook/stochastic-gradient-descent

# 회귀분석

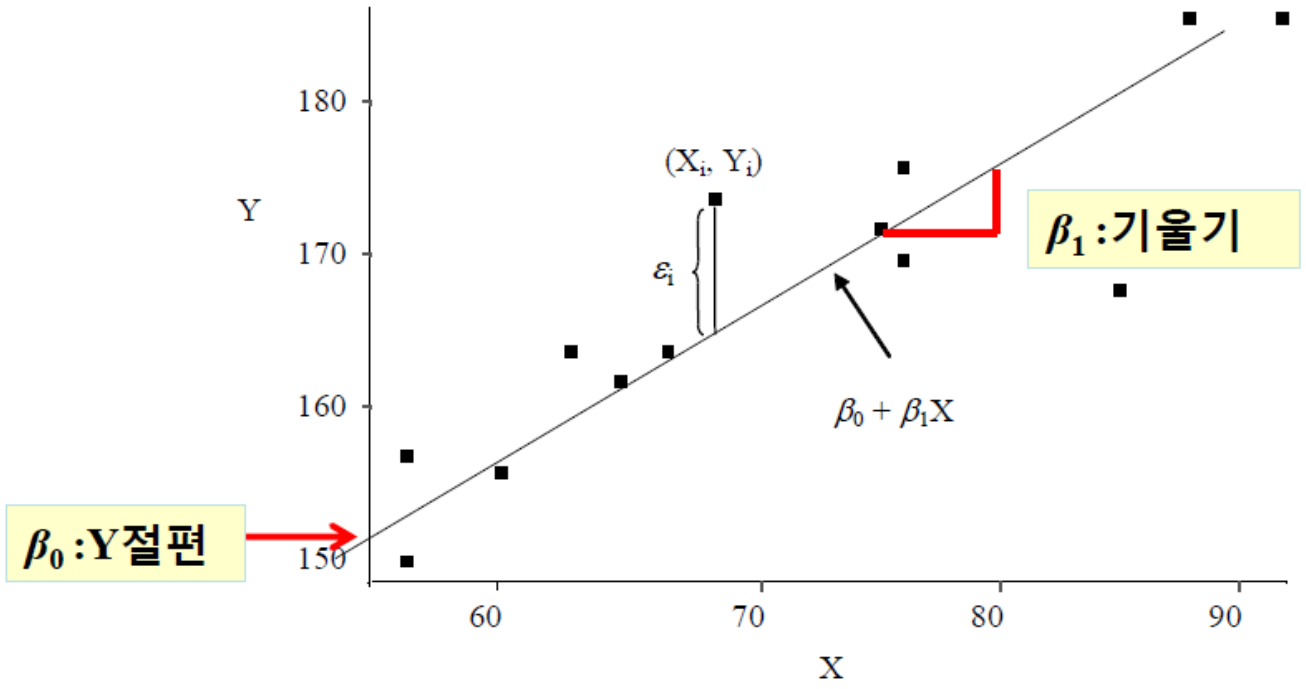
- **어떤 변수가 다른 변수에 영향을 받는 경우, 아래와 같이 정의**
  - 종속변수 (=y): 영향을 받는 변수
  - 독립변수 (=x): 영향을 주는 변수

- **회귀분석 (Regression Analysis)**
  - 독립변수의 변화에 따른 종속변수의 변화를 예측하는 분석방법

- **단순선형회귀분석**
  - 단순: 독립변수가 하나임
  - 선형: 독립변수와 종속변수의 관계가 선형 (직선) 으로 표현됨

# 단순선형회귀모형

- **관측치:** $(X_1,Y_1), (X_2,Y_2), \ldots, (X_n, Y_n)$
- **모형:** $Y_i=\beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, 2, \ldots, n$
- **회귀계수:** $\beta_0, \beta_1$



# 단순선형회귀모형

- **가정**
  - 독립변수는 결정되어 있음. 즉, 확률변수가 아님
  - $\varepsilon_i$ 는 오차를 나타내는 확률변수로 $N(0, \sigma^2)$ 를 따름
  - Y는 X로 인해 예측되는 값에 오차가 더해진 값

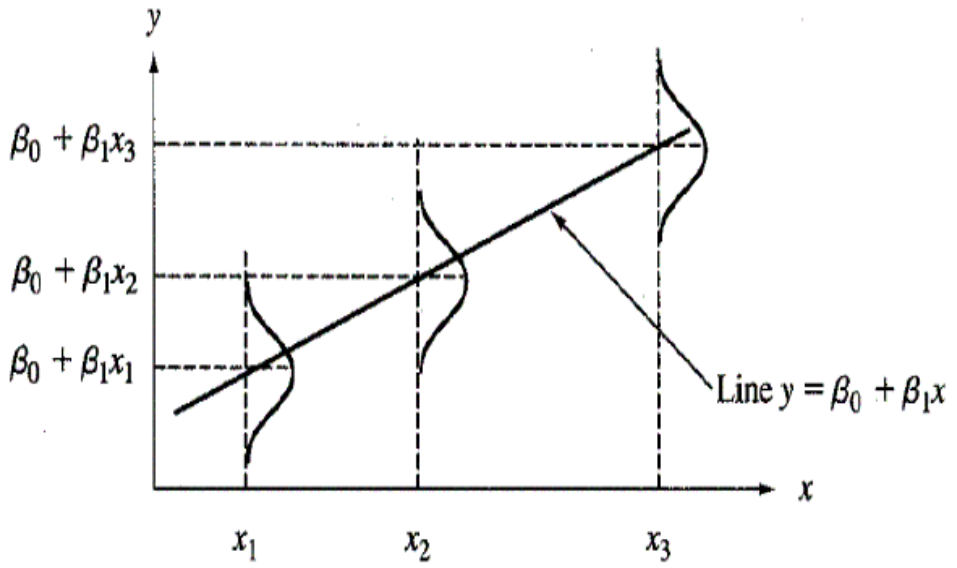
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

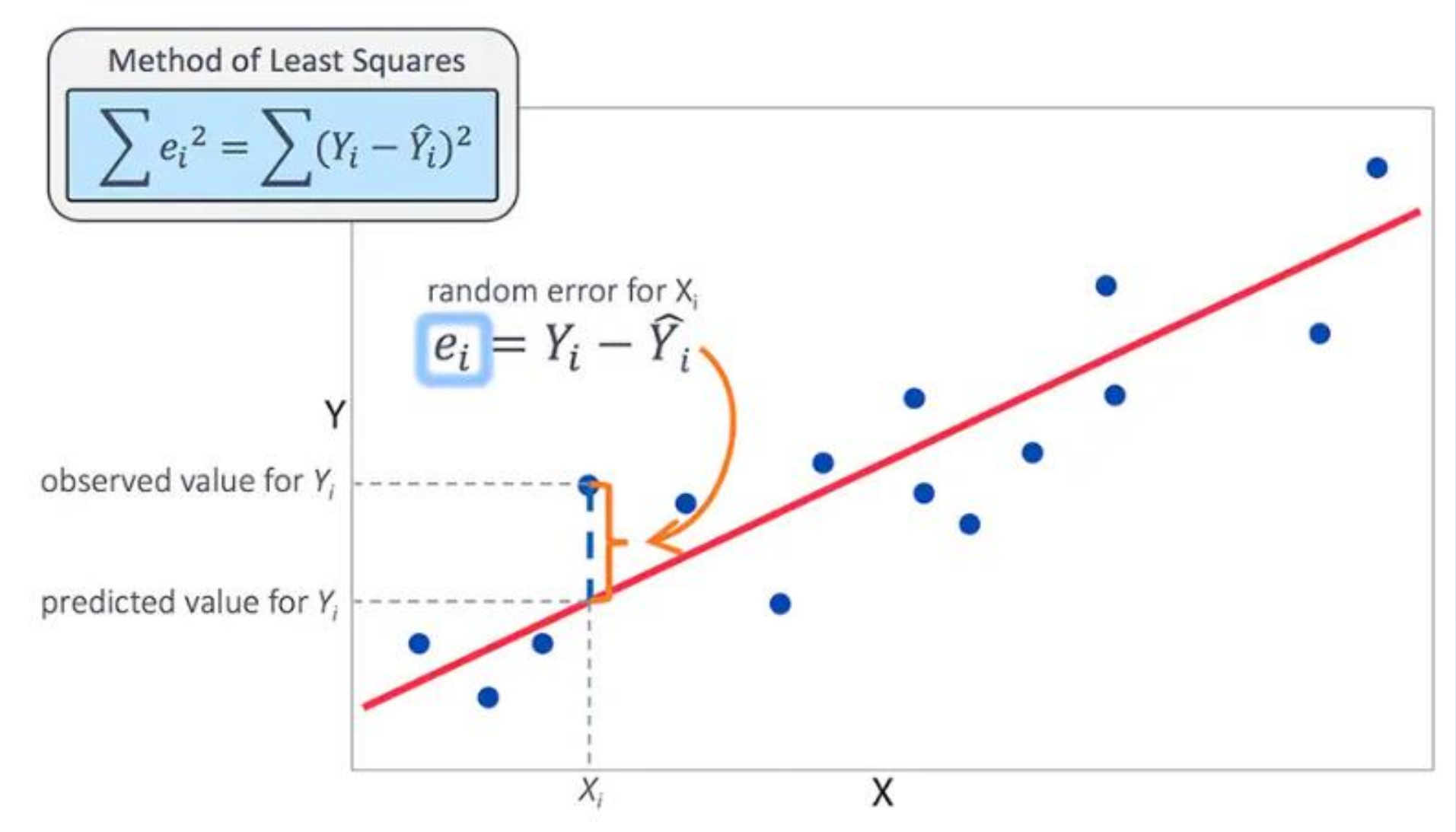
$$Cov(\varepsilon_i, \varepsilon_j) = 0, \quad i \neq j$$

$$Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$$

**즉, $Y_i$ 도 정규분포를 따르는 확률변수**



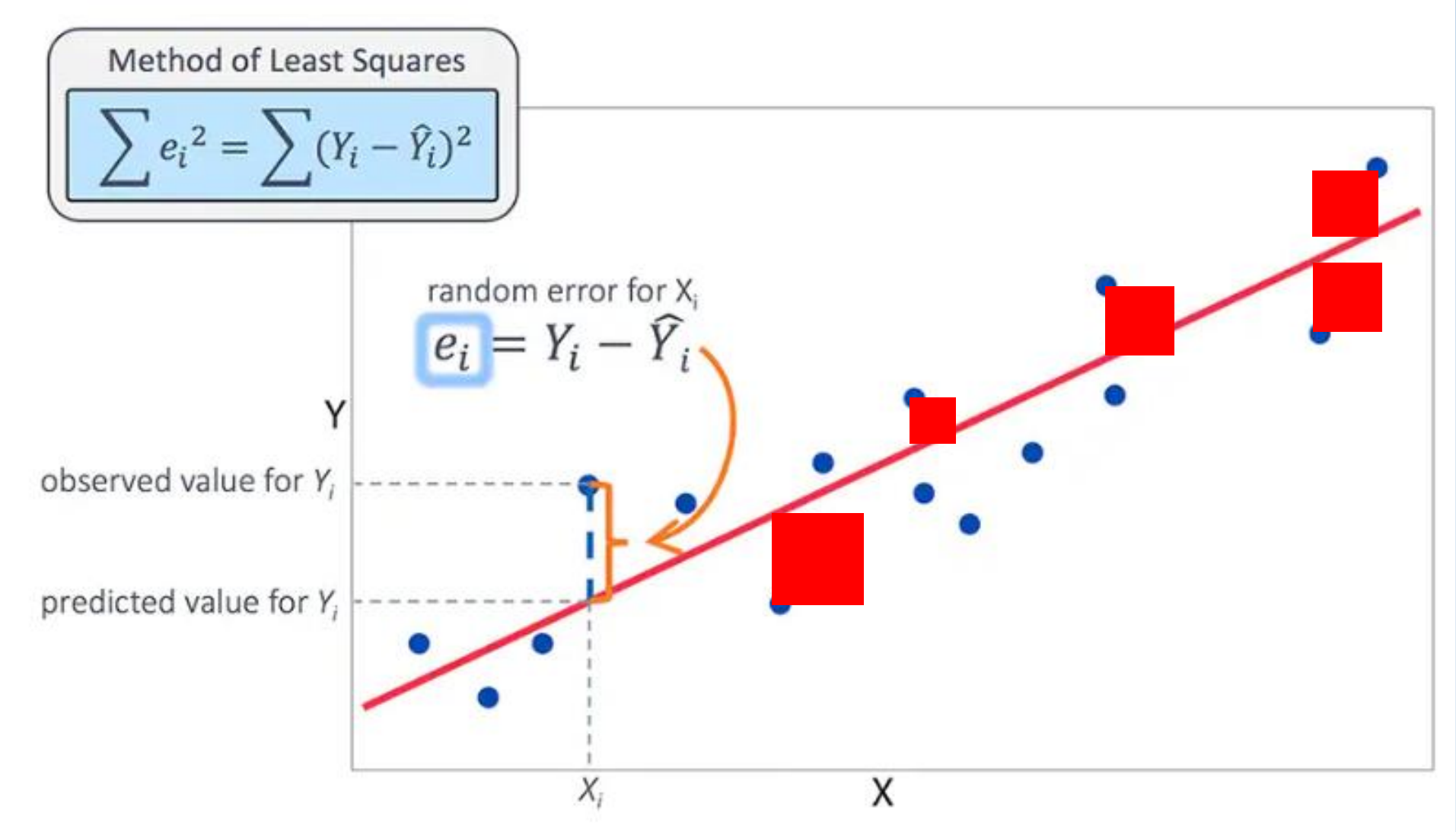
# 회귀분석 계수(coefficients) 구하기 (1/2)




$$y = \alpha + \beta x,$$

$$y_i = \alpha + \beta x_i + \varepsilon_i.$$

# 회귀분석 계수(coefficients) 구하기 (1/2)




$$y = \alpha + \beta x,$$

$$y_i = \alpha + \beta x_i + \varepsilon_i.$$

# 회귀계수 추정 (1/2)

- **최소자승법 (method of least squares)**
  - 종속변수의 관측된 값과 모형에 의한 예측된 값 사이의 오차의 제곱합을 최소화 시키는 회귀계수를 추정함이 목적
  - 오차의 제곱을 최소화시키는 $\beta_0, \beta_1$ 를 찾는 과정

$$\min_{\beta_0,\beta_1} Q = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

$$\begin{cases} \dfrac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) = 0 \\ \dfrac{\partial (Q)}{\partial \beta_1} = -2\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) X_i = 0 \end{cases}$$

$$\begin{cases} n\hat{\beta}_0 + \hat{\beta}_1 \sum X_i = \sum Y_i \\ \hat{\beta}_0 \sum X_i + \hat{\beta}_1 \sum X_i^2 = \sum X_i Y_i \end{cases}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

# 오차항의 분산 (1/2)

- 오차항은 관측될 수 없으므로, 잔차의 표준편차에 의해서 추정됨
- 잔차(residual):

$$e_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$$

- 잔차제곱합(error sum of squares; SSE):

$$SSE = \sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2$$

- 잔차평균제곱(mean square error; MSE):

$$MSE = \hat{\sigma}^2 = s^2 = \frac{SSE}{n-2} = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

  - MSE 는 $\sigma^2$ 의 불편 추정량이다. (증명해 볼 것)

# 회귀계수에 대한 검정 (1/3)

- **두 회귀계수의 추정량 $\hat{\beta}_1$와 $\hat{\beta}_0$ 도 확률변수이며 정규 분포를 따름**

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$$

$$\hat{\beta}_0 \sim N(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right)), \quad \hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{\sum(X_i - \bar{X})^2})$$

- $E[\hat{\beta}_1] = \beta_1$
- $E[\hat{\beta}_0] = \beta_0$

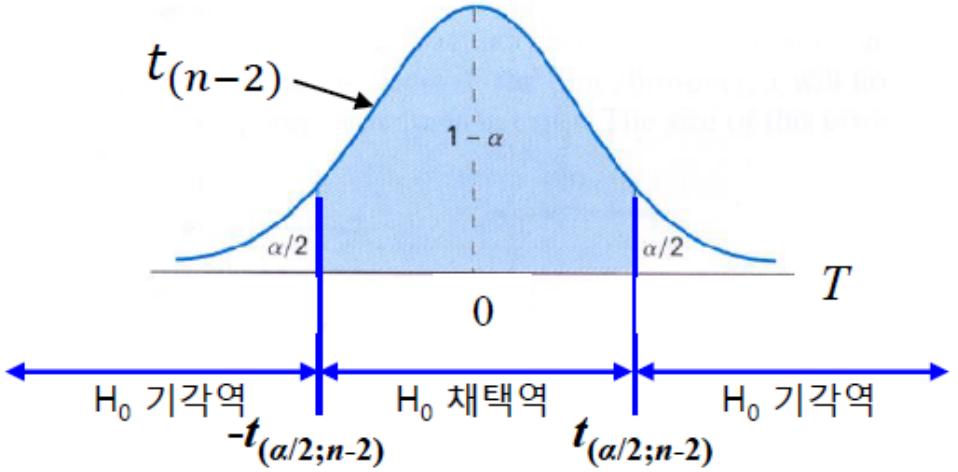
⇒ 두 회귀계수의 추정량 $\hat{\beta}_1$와 $\hat{\beta}_0$ 은 각각 $\beta_1$ 와 $\beta_0$의 불편추정량임

# 회귀계수에 대한 검정 (2/3)

- $\beta_1$에 대한 t-검정
  - 가설: $H_0 : \beta_1 = 0$
    $H_1 : \beta_1 \neq 0$
  - 검정통계량

$$T_1 = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \qquad se(\hat{\beta}_1) = \sqrt{\frac{MSE}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

- $T_1$은 자유도 n-2 인 t 분포를 따름



# 데이터

1 to 25 of 17000 entries Filter

| index | longitude | latitude | housing_median_age | total_rooms | total_bedrooms | population | households | median_income | median_house_value |
|---|---|---|---|---|---|---|---|---|---|
| 0 | -114.31 | 34.19 | 15.0 | 5612.0 | 1283.0 | 1015.0 | 472.0 | 1.4936 | 66900.0 |
| 1 | -114.47 | 34.4 | 19.0 | 7650.0 | 1901.0 | 1129.0 | 463.0 | 1.82 | 80100.0 |
| 2 | -114.56 | 33.69 | 17.0 | 720.0 | 174.0 | 333.0 | 117.0 | 1.6509 | 85700.0 |
| 3 | -114.57 | 33.64 | 14.0 | 1501.0 | 337.0 | 515.0 | 226.0 | 3.1917 | 73400.0 |
| 4 | -114.57 | 33.57 | 20.0 | 1454.0 | 326.0 | 624.0 | 262.0 | 1.925 | 65500.0 |
| 5 | -114.58 | 33.63 | 29.0 | 1387.0 | 236.0 | 671.0 | 239.0 | 3.3438 | 74000.0 |
| 6 | -114.58 | 33.61 | 25.0 | 2907.0 | 680.0 | 1841.0 | 633.0 | 2.6768 | 82400.0 |
| 7 | -114.59 | 34.83 | 41.0 | 812.0 | 168.0 | 375.0 | 158.0 | 1.7083 | 48500.0 |
| 8 | -114.59 | 33.61 | 34.0 | 4789.0 | 1175.0 | 3134.0 | 1056.0 | 2.1782 | 58400.0 |
| 9 | -114.6 | 34.83 | 46.0 | 1497.0 | 309.0 | 787.0 | 271.0 | 2.1908 | 48100.0 |
| 10 | -114.6 | 33.62 | 16.0 | 3741.0 | 801.0 | 2434.0 | 824.0 | 2.6797 | 86500.0 |
| 11 | -114.6 | 33.6 | 21.0 | 1988.0 | 483.0 | 1182.0 | 437.0 | 1.625 | 62000.0 |
| 12 | -114.61 | 34.84 | 48.0 | 1291.0 | 248.0 | 580.0 | 211.0 | 2.1571 | 48600.0 |
| 13 | -114.61 | 34.83 | 31.0 | 2478.0 | 464.0 | 1346.0 | 479.0 | 3.212 | 70400.0 |

# 상관계수

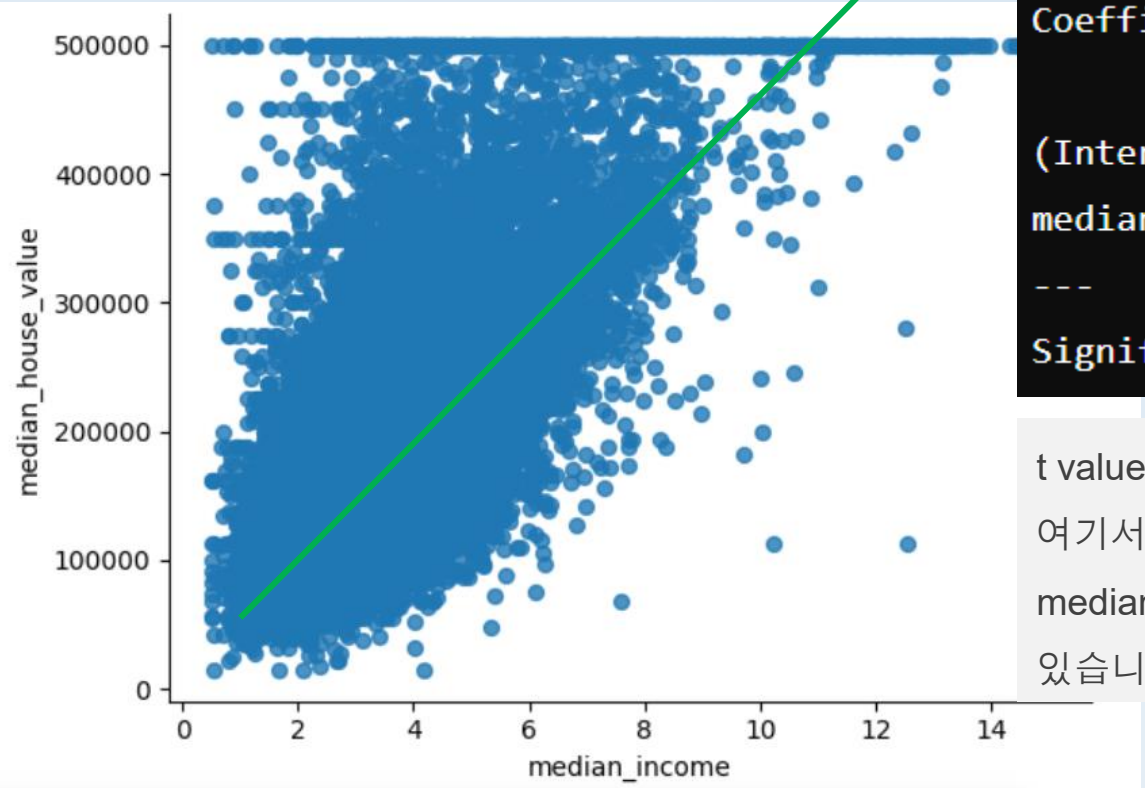
1 to 9 of 9 entries Filter

| index | longitude | latitude | housing_median_age | total_rooms | total_bedrooms | population | households | median_income | median_house_value |
|---|---|---|---|---|---|---|---|---|---|
| longitude | 1.0 | -0.93 | -0.11 | 0.05 | 0.07 | 0.1 | 0.06 | -0.02 | -0.04 |
| latitude | -0.93 | 1.0 | 0.02 | -0.04 | -0.07 | -0.11 | -0.07 | -0.08 | -0.14 |
| housing_median_age | -0.11 | 0.02 | 1.0 | -0.36 | -0.32 | -0.3 | -0.3 | -0.12 | 0.11 |
| total_rooms | 0.05 | -0.04 | -0.36 | 1.0 | 0.93 | 0.86 | 0.92 | 0.2 | 0.13 |
| total_bedrooms | 0.07 | -0.07 | -0.32 | 0.93 | 1.0 | 0.88 | 0.98 | -0.01 | 0.05 |
| population | 0.1 | -0.11 | -0.3 | 0.86 | 0.88 | 1.0 | 0.91 | -0.0 | -0.03 |
| households | 0.06 | -0.07 | -0.3 | 0.92 | 0.98 | 0.91 | 1.0 | 0.01 | 0.06 |
| median_income | -0.02 | -0.08 | -0.12 | 0.2 | -0.01 | -0.0 | 0.01 | 1.0 | 0.69 |
| median_house_value | -0.04 | -0.14 | 0.11 | 0.13 | 0.05 | -0.03 | 0.06 | 0.69 | 1.0 |

# 결과 표 해석하기 (1/2)

```
Call:
lm(formula = median_house_value ~ median_income, data = data_df)
```

$$y = \alpha + \beta x,$$



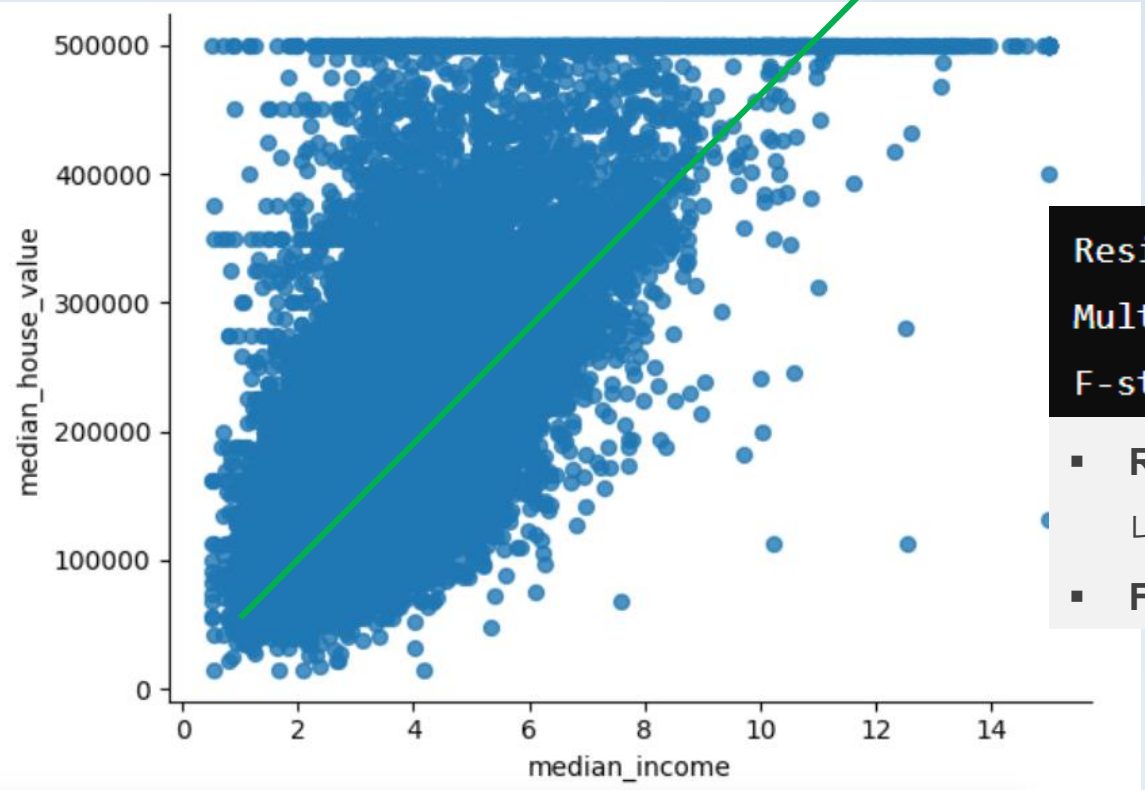

median_house_value = 294610.2 + 20,1* median_income

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    294610.2    29948.9   9.839  <2e-16 ***
median_income     20.1        11.7   1.718  0.0887 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

t value와 Pr(>|t|)는 각 계수의 유의성을 테스트하는 **t-통계량과 p-값**입니다. 여기서 median_income의 p-값이 0.0887로 0.05보다 크므로, 이 변수가 median_house_value에 미치는 영향이 **통계적으로 유의하지 않다고** 할 수 있습니다.

# 결과 표 해석하기 (2/2)

median_house_value = 294610.2 + 20,1* median_income




```
Residual standard error: 87590 on 98 degrees of freedom
Multiple R-squared:  0.03007,   Adjusted R-squared:  0.02017
F-statistic: 2.952 on 1 and 98 DF,  p-value: 0.08874
```

- **R-squared**: 결정 계수는 모델이 데이터의 변동성을 얼마나 잘 설명하는지 나타내며, 이 경우에는 약 3.01%입니다.
- **F-statistic**과 그 p-값은 모델 전체가 유의한지를 나타냅니다.

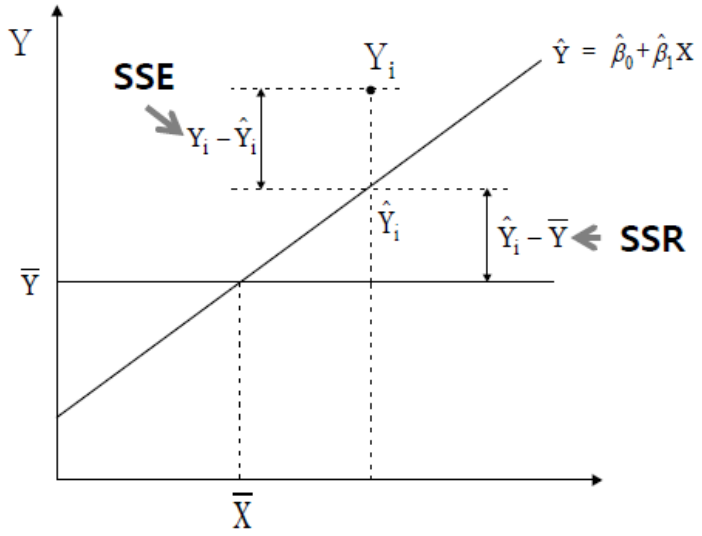
```
Call:
lm(formula = median_house_value ~ median_income, data = data_df)

Residuals:
     Min       1Q   Median       3Q      Max
-150196.3 -71794.5  -2145.9  67274.4 172644.8

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    294610.2    29948.9   9.839  <2e-16 ***
median_income     20.1        11.7   1.718  0.0887 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 87590 on 98 degrees of freedom
Multiple R-squared:  0.03007,   Adjusted R-squared:  0.02017
F-statistic: 2.952 on 1 and 98 DF,  p-value: 0.08874
```

# 모형의 적합도 검정 (1/4)

- **모형의 적합도와 결정계수 ($R^2$)**
  - 정의 : SST에 대한 SSR의 비율, 즉 모형으로 설명할 수 있는 부분의 비율

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

  - $0 \le R^2 \le 1$




**전체제곱합의 분할: SST=SSR+SSE**

**전체제곱합(SST):** $SST = \sum (Y_i - \bar{Y})^2$

**회귀제곱합(SSR):** $SSR = \sum (\hat{Y}_i - \bar{Y})^2$

**잔차제곱합(SSE):** $SSE = \sum (Y_i - \hat{Y}_i)^2$

# 모형의 적합도 검정 (3/4)

- **회귀모형에 대한 유의성 검정(F-검정)**
  - 단순선형 회귀모형에서는 t-검정과 동일 (Part I page 10)

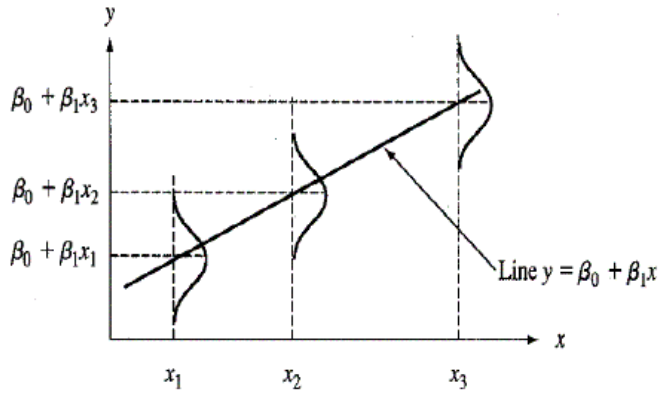
- **검정과정**
  - 가설: $H_0 : \beta_1 = 0$
    $H_1 : \beta_1 \neq 0$
  - 검정통계량: $F_0 = \dfrac{SSR/1}{SSE/(n-2)} = \dfrac{MSR}{MSE}$

(가설 $H_0$가 옳을 때 위의 $F_0$는 자유도 (1, n-2)를 갖는 F-분포를 따름.)

- **기각결정**
  - $F_0 \geq F_{(\alpha;\, 1,\, n-2)}$이면 $H_0$ 를 기각한다.

# 잔차 분석 (1/4)

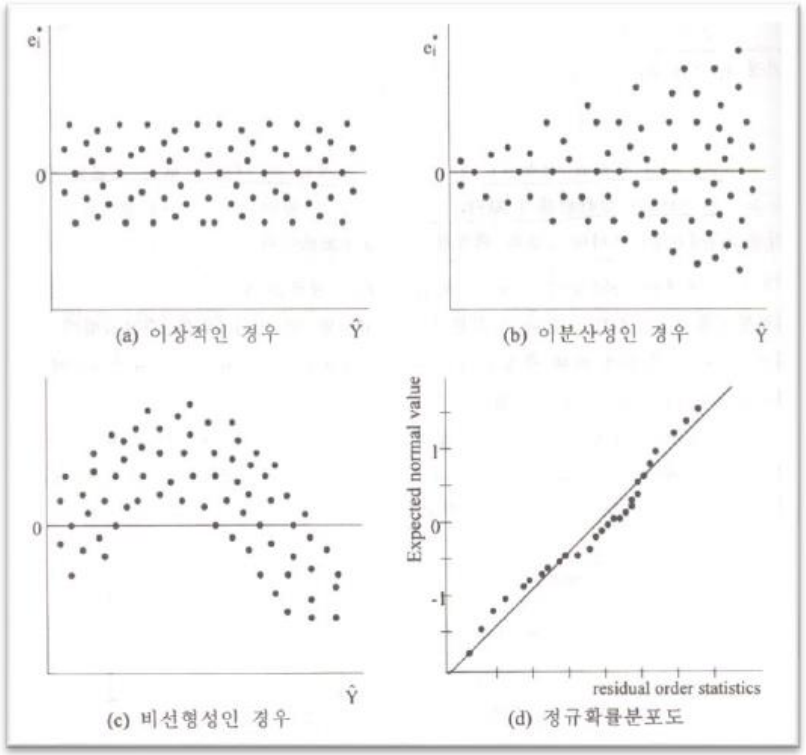
- **단순회귀분석에서 모수의 추정은 오차항에 대한 가정을 바탕으로 함**

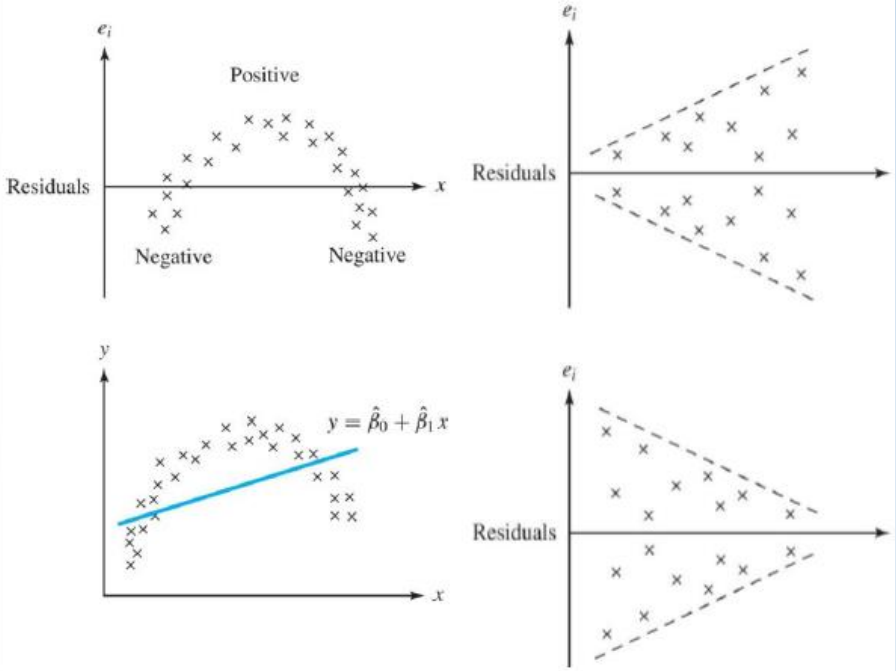



- 오차는 정규분포를 따른다
- $E[\varepsilon_i] = 0,\ i=1,2,\ldots,n$
  → 모든 입력변수 값에 대해 평균이 0이다
- 등분산성: $Var[\varepsilon_i] = \sigma^2\ ,\ i=1,2,\ldots,n$
  → 모든 입력변수 값에 대해 산포가 동일하다
- $Cov[\varepsilon_i, \varepsilon_i] = 0,\ i \neq j,\ i,j=1,2,\ldots,n$
  →어떤 Y값에 대한 오차는 다른 Y값의 오차와 독립이다

# 잔차 분석 (2/4)

- **잔차 산점도를 통해 오차항의 가정을 판단함**




오차항의 가정에 위배되는 경우, 변수의 변환 혹은 다른 회귀모형을 사용해야함


OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | MedHouseVal | R-squared: | 0.479 |
| Model: | OLS | Adj. R-squared: | 0.479 |
| Method: | Least Squares | F-statistic: | 9502. |
| Date: | Mon, 04 Apr 2022 | Prob (F-statistic): | 0.00 |
| Time: | 04:27:24 | Log-Likelihood: | -25505. |
| No. Observations: | 20640 | AIC: | 5.102e+04 |
| Df Residuals: | 20637 | BIC: | 5.104e+04 |
| Df Model: | 2 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 0.5950 | 0.016 | 36.836 | 0.000 | 0.563 | 0.627 |
| MedInc | 0.4342 | 0.003 | 134.497 | 0.000 | 0.428 | 0.440 |
| AveRooms | -0.0381 | 0.002 | -15.375 | 0.000 | -0.043 | -0.033 |

| | | | |
|---|---|---|---|
| Omnibus: | 4804.179 | Durbin-Watson: | 0.692 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 12852.863 |
| Skew: | 1.250 | Prob(JB): | 0.00 |
| Kurtosis: | 5.949 | Cond. No. | 20.3 |

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

$$y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \epsilon$$