단순 선형 회귀
(Simple Linear Regression)

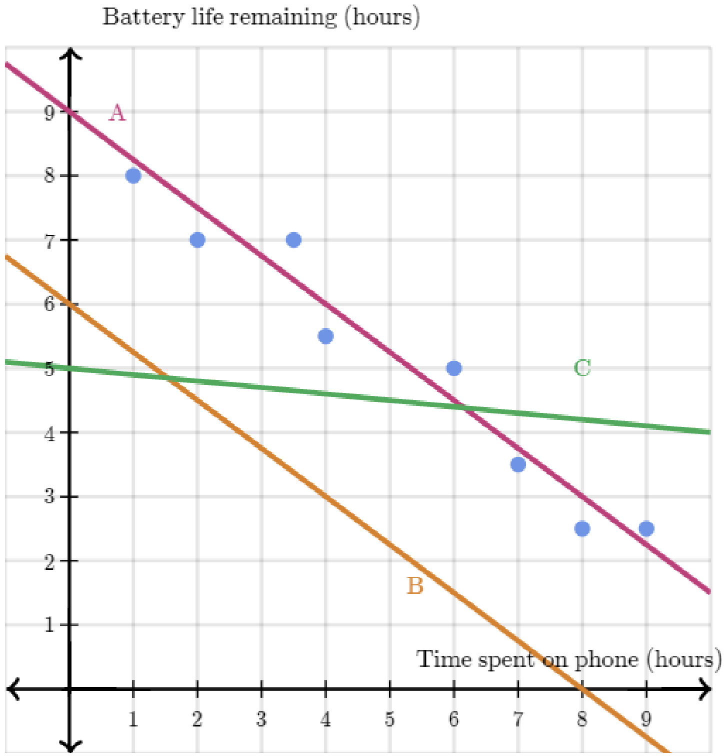
Simple Linear Regression

부산대학교 정보·의생명공학대학
정보컴퓨터공학부

Equation of trend lines in Scatter plot

❖ Example) Phone data

Time spent on phone (hours)	1	2	3.5	4	6	7	8	9
Battery life remaining (hours)	8	7	7	5.5	5	3.5	2.5	2.5



1. Which line fits the data graphed above ?
2. Which equation describes the best trend line above ?

① $y = -0.5x + 9$

② $y = -0.75x + 9$

③ $y = -x + 9$

④ $y = -2x + 9$

3. Use the equation of the trend line to predict the battery life remaining **after 3.6 hours** of phone use.
4. Use the trend line to predict the battery life remaining **after 20 hours** of phone use.
5. Does the prediction from problem 4 seem reasonable in the context of the problem?

① Yes. Plugging in 20 gave us a prediction of -6 hours

② No. It doesn't make sense for battery life to be negative in this context.

6. What is the best interpretation of the slope of this trend line?

① For each additional 1 hour of time spent on the phone, the predicted battery life remaining increases by 0. 75 hours.

② For each additional 1 hour of time spent on the phone, the predicted battery life remaining decreases by 0. 75 hours.

③ For each additional 0.75 hours of time spent on the phone, the predicted battery life remaining decreases by 1 hours.

7. What is the best interpretation of the y-intercept of this trend line?

① When the time spent on her phone is 9 hours, the predicted battery life remaining is 0 hours.

② When the time spent on her phone is 0 hours, the predicted battery life remaining is 9 hours.

8. Yuna wants to turn her phone off when the battery has 15 minutes remaining, just in case she has an emergency and needs her phone later. According to the trend line, how long can she spend on her phone before she needs to turn it off?

Simple Linear Regression – Least Squares Regression Equations

- $y = \beta_0 + \beta_1 x$
  - Simple Linear Regression Model
  - β_0 : intercept
  - β_1: slope

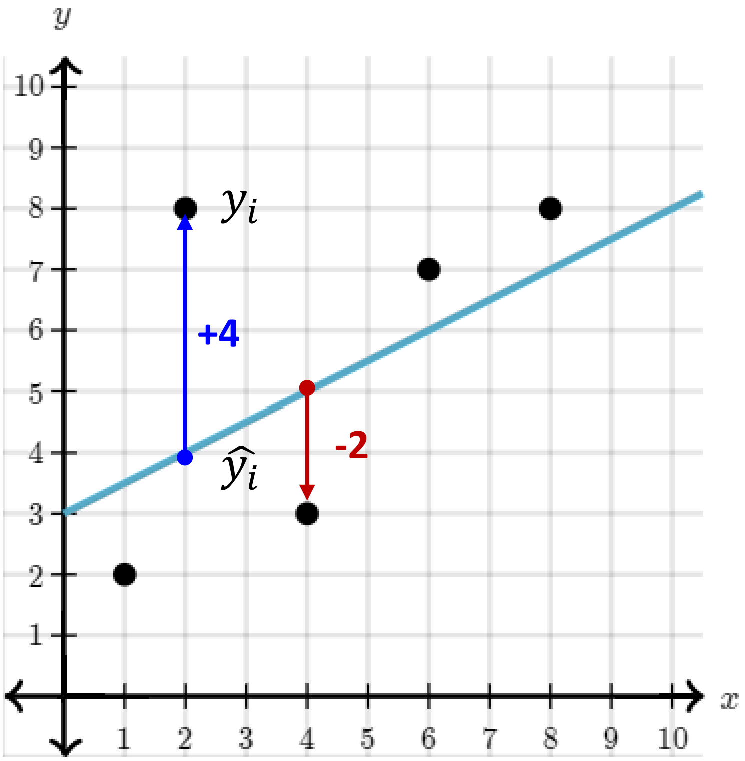
- $y_i \sim \beta_0 + \beta_1 x_i + \varepsilon_i$
  - ε_i : Error term , $\varepsilon_i \sim_{iid} N(0, \sigma^2)$
  - $\hat{y}_i = \beta_0 + \beta_1 x_i$: Fitted/Predicted Value

- Least (Residual Sum of) Squares Methods
  - 최소제곱법 / 최소 자승법

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{\beta_0,\beta_1} \sum_i^n [y_i - (\beta_0 + \beta_1 x_i)]^2$$

$$= \arg\min_{\beta_0,\beta_1} \sum_i^n (y_i - \hat{y}_i)^2$$

- Residual (잔차)



Least Squares Regression Equations

$$\boldsymbol{\hat{\beta}_1 = r_{xy} \cdot \frac{s_y}{s_x}}$$

- r_{xy} : correlation coefficient
- s_x : standard deviation of x
- s_x : standard deviation of y

$$\boldsymbol{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}}$$

- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \, , \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, the sample mean

❖ Example

- A personal trainer wants to look at the relationship between number of hours of exercise per week and resting heart rate of her clients. The data show a linear pattern with the summary statistics shown below

	mean	Standard deviation
$\boldsymbol{x}$ = hours of exercise per week	$\bar{x} = 8.9$	$s_x = 4.8$
y = resting heart rate (beats per minute)	$\bar{y} = 74.3$	$s_y = 7.2$
	$r_{xy} = -0.88$	

- Find the equation of the least-squares regression line for predicting resting heart rate from the hours of exercise per week.
- $\hat{y} = \beta_0 + \beta_1 x, \ \beta_0 = ?\, , \beta_1 = ?$
- $\beta_1 = -0.88 \cdot \frac{7.2}{4.8} = -1.32, \ \beta_0 = 74.3 - \beta_1 \cdot 8.9 = 86.048$

Calculating Slope and Intercept

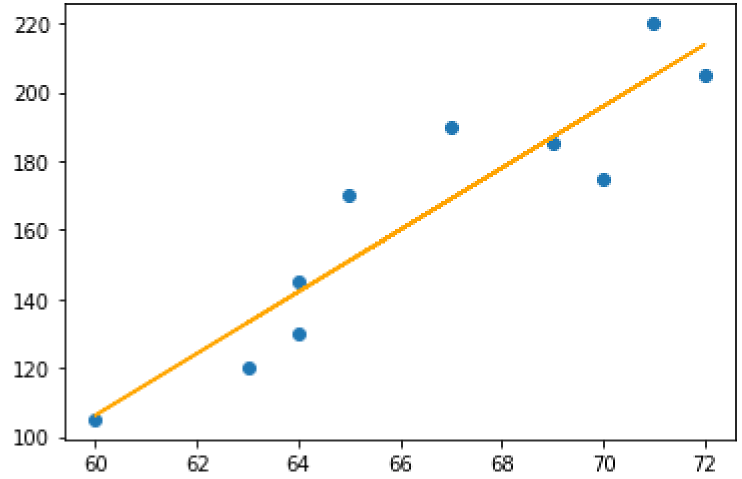
- $\hat{\beta}_1 = r_{xy} \cdot \frac{s_y}{s_x}$, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

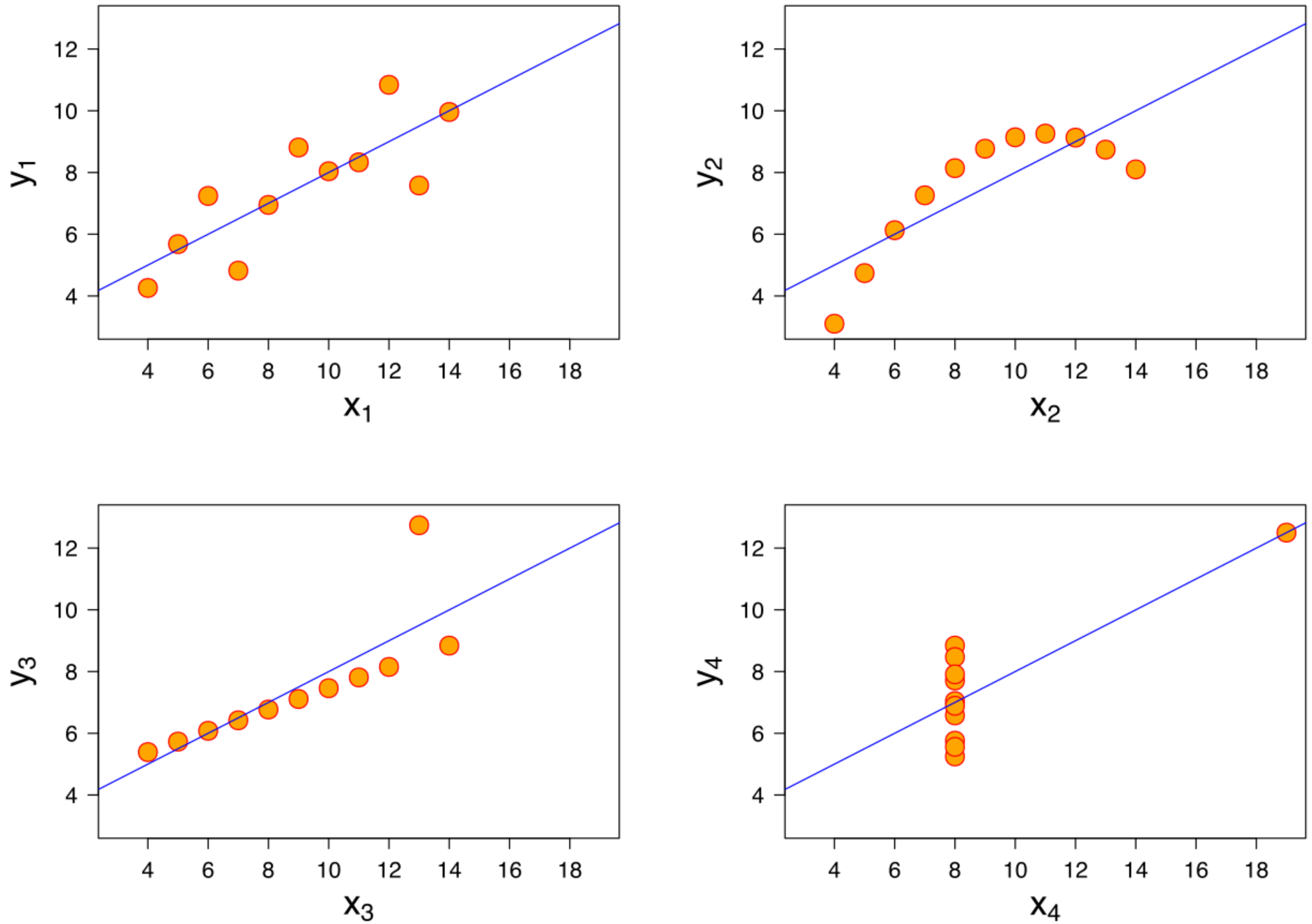
friends x	minutes y		
70	175	$\bar{x}$	66.5
65	170	$\bar{y}$	164.5
72	205	s_x	3.922867432
63	120	s_y	38.03872296
71	220	r_{xy}	0.9251759349
64	130		
60	105		
64	145		
67	190		
69	185		

- Spreadsheet
  - =SLOPE(D2:D11, C2:C11) ; =SLOPE(Y_Values, X_Values)
  - =INTERCEPT(D2:D11, C2:C11);
    =INTERCEPT(Y_Values, X_Values)
  - =FORECAST(X, Y_Vales, X_Values)

Scikit-Learn Linear Regression Model

```
import pandas as pd
datum = pd.read_csv('https://raw.githubusercontent.com/inetguru/IDS-CB35533/main/datum.csv', index_col='id')
datum.head()

from sklearn.linear_model import LinearRegression

reg = LinearRegression().fit(datum[['friends']],datum[['minutes']])

print('Slope = ', reg.coef_, ' Intercept =', reg.intercept_)
print('Predicted Value when friends = 70, y_hat =',reg.predict([[70]]))

import matplotlib.pyplot as plt
plt.scatter(datum['friends'], datum['minutes'])
plt.plot(X,reg.predict(X),color='orange')
plt.show()
```





Source : https://en.wikipedia.org/wiki/Linear_regression#/media/File:Anscombe's_quartet_3.svg

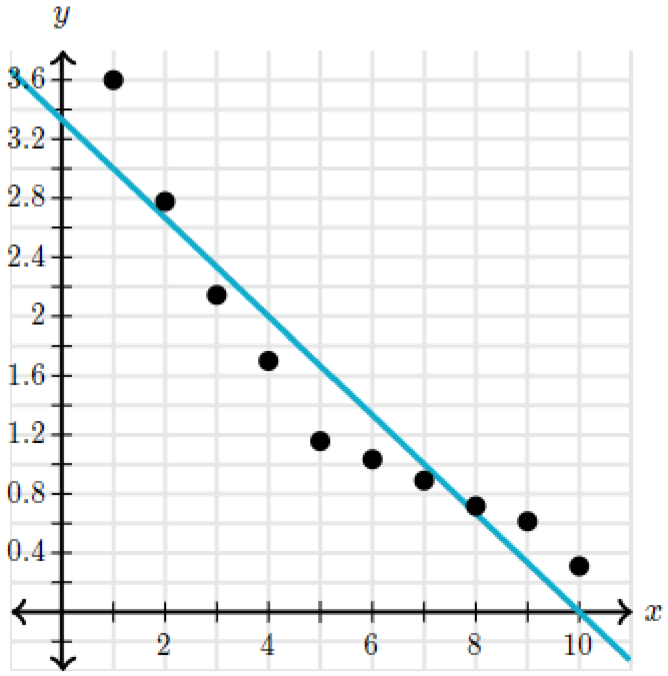
The data sets in the Anscombe's quartet are designed to have approximately the same linear regression line (as well as nearly identical means, standard deviations, and correlations) but are graphically very different. This illustrates the pitfalls of relying solely on a fitted model to understand the relationship between variables.

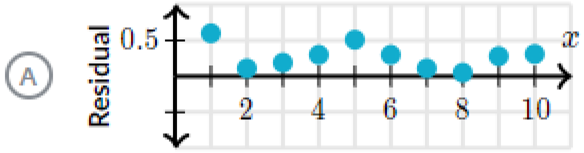
Residual Plot

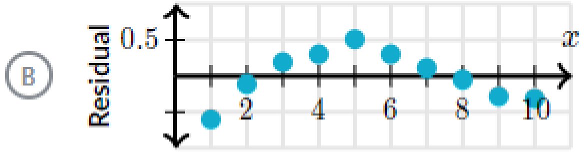
❖ Residual Plot

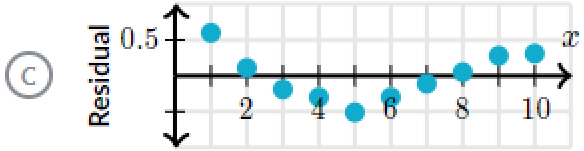
- A **residual plot** is a **graph** that shows the **residuals** on the vertical axis and the independent variable on the horizontal axis.

❖ Example







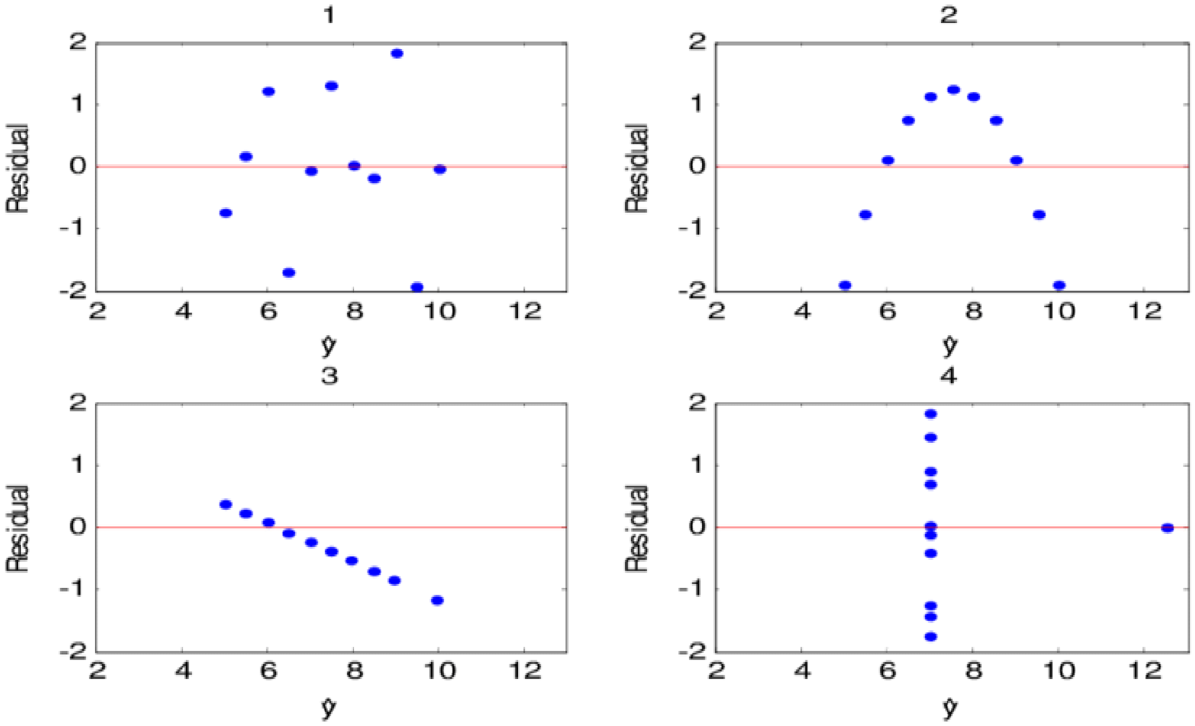


Random/Non-random Residual plot

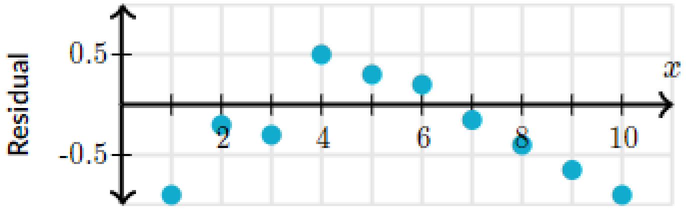
- ❖ If the points in a residual plot are randomly dispersed around the horizontal axis,

  a linear regression model is appropriate for the data;

  otherwise, a nonlinear model is more appropriate.

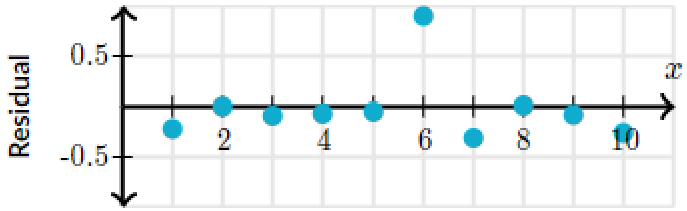


- ❖ Example – What can you conclude from the residual plot ?



① There appears to be a linear relationship between x and y

② There does not appear to be a linear relationship between x and y.



① When x = 6, the least squares regression equation overestimates y.

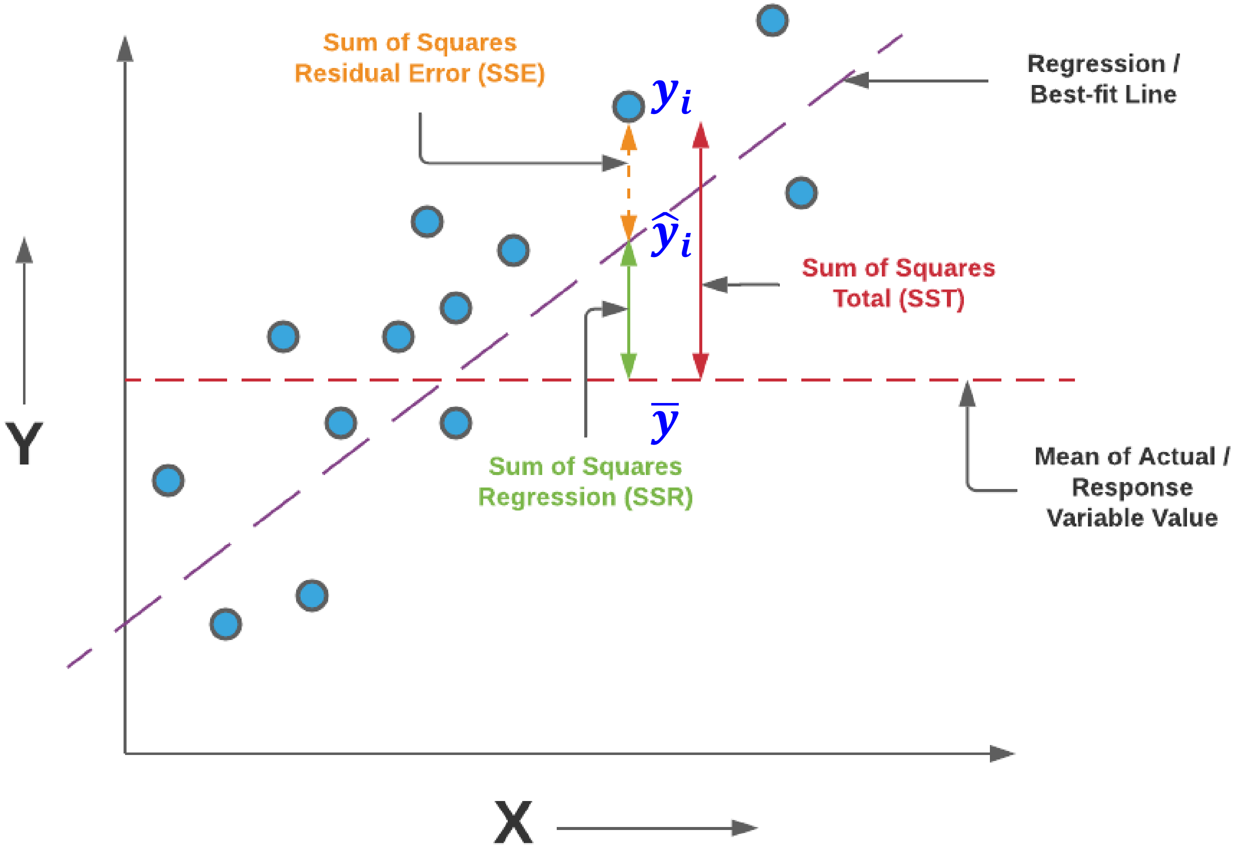
② The slope of the least-squares regression line is 0.

③ The least squares regression equation overestimates y more often than it underestimates x.

Coefficient of Determination : r^2

- The Goodness of fit of a model
  - describes how well it fits a set of observations
  - 모형 적합도
- Coefficient of Determination : $\boldsymbol{r^2}$
  - 결정계수(決定係數)
  - Measures **how much prediction error is eliminated** when we use least-squares regression.
  - $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$: Total sum of squares
  - $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$: Regression sum of squares
  - $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$: Error sum of squares
  - $SST = SSR + SSE$

$$r^2 = \frac{SSR}{SST}$$

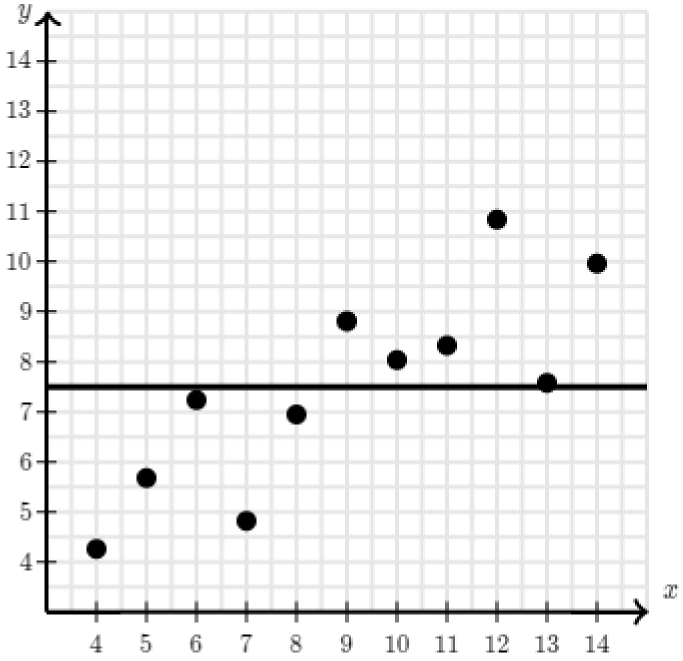


Sum of Squares Residual Error (SSE)
y_i
Regression / Best-fit Line
$\hat{y}_i$
Sum of Squares Total (SST)
$\bar{y}$
Sum of Squares Regression (SSR)
Mean of Actual / Response Variable Value
Y
X

Intuition for r^2

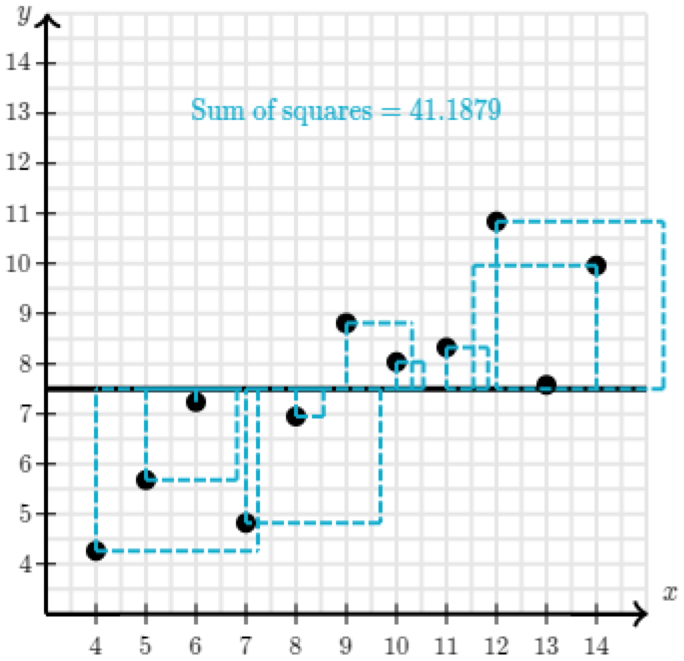
❖ Predicting without regression vs. with regression

- We use linear regression to predict y given some value of x.
- But suppose that we had to predict a y value without a corresponding x value.
- Without using regression on the x variable, our most reasonable estimate would be to simply predict **the average** of the y values.



❖ Without regression

- Notice that this line doesn't seem to fit the data very well. One way to measure the fit of the line is to calculate the sum of the squared residuals—this gives us an overall sense of how much prediction error a given model has.
- **So without least-squares regression, our sum of squares is 41.1879 (SST)**



Intuition for r^2

❖ With Regression

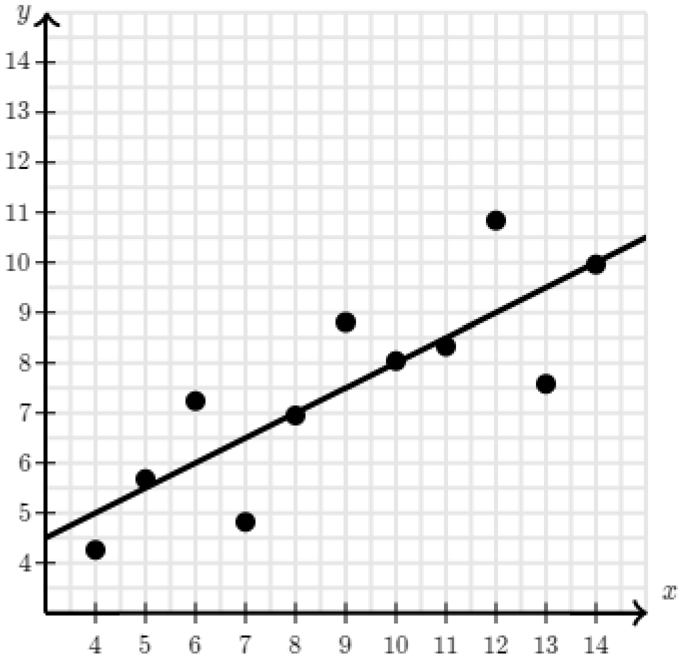
Equation	r	r^2
$\hat{y} = 0.5x + 1.5$	0.816	0.6659

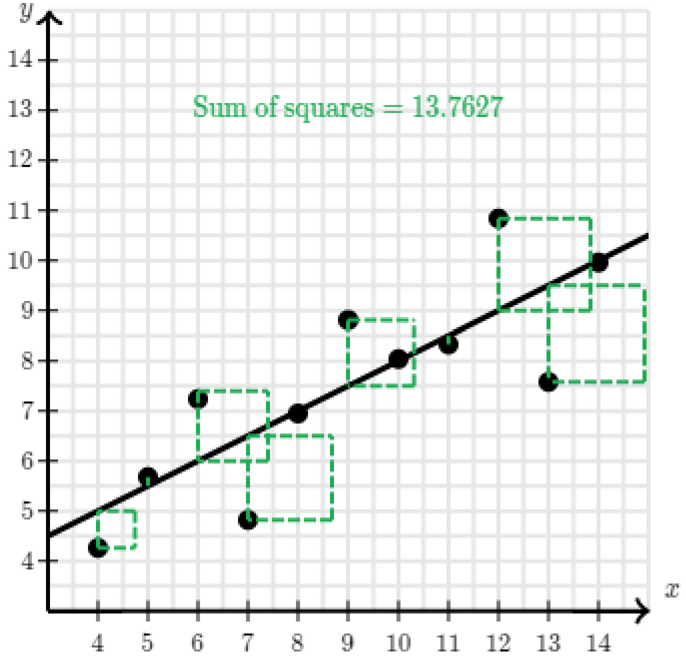
- how much better it fits ?
  - the sum of the squared residuals : 13.7627 (SSE)

- Using least-squares regression reduced the sum of the squared residuals **from 41.1879 to 13.7627**

- So using least-squares regression eliminated a considerable amount of prediction error. How much?

$$\frac{41.1879 - 13.7267}{41.7879} \approx 66.59\% = r^2$$





Summary

❖ Statistical Relationship between a pair of variables

- Explanatory(Independent) Variable X & Response (Dependent) Variable Y

❖ Tools

- Scatter Plot
- Correlation Coefficient

$$\boldsymbol{r_{xy}} = \frac{\sum_{i=1}^{n}(\boldsymbol{x_i} - \overline{\boldsymbol{x}})(\boldsymbol{y_i} - \overline{\boldsymbol{y}})}{\sqrt{\sum_{i=1}^{n}(\boldsymbol{x_i} - \overline{\boldsymbol{x}})^2} \cdot \sqrt{\sum_{i=1}^{n}(\boldsymbol{y_i} - \overline{\boldsymbol{y}})^2}}$$

- Linear Regression - 2 Bivariate Numerical Data

❖ Simple Linear Regression

- $\hat{y} = \beta_0 + \beta_1 x$
  - Slope, intercept
- Residual
- Least (Residual Sum of) Squares Methods

$$\widehat{\boldsymbol{\beta}}_1 = \boldsymbol{r_{xy}} \cdot \frac{\boldsymbol{s_y}}{\boldsymbol{s_x}}, \quad \widehat{\boldsymbol{\beta}}_0 = \overline{\boldsymbol{y}} - \widehat{\boldsymbol{\beta}}_1 \overline{\boldsymbol{x}}$$

❖ The Goodness of fit of a model

- Coefficient of determination

$$\boldsymbol{r^2} = \frac{\sum_{i=1}^{n}(\widehat{\boldsymbol{y}}_i - \overline{\boldsymbol{y}})^2}{\sum_{i=1}^{n}(\boldsymbol{y_i} - \overline{\boldsymbol{y}})^2}$$