1.1 Simple linear regression

如果你是一位吳柏毅的數據分析師
你可能會有以下任務：

- 該如何訂定下個月的成本控制目標？
- 有哪些因素會影響每天的成本？

一個結果是由很多原因造成的

- 每月成本

	基本成本	外送員分潤	店家分潤	人事成本	總支出
1月	10	4	6.6	4.4	25
2月	10	6.5	3	5	24.5
3月	10	8	13	7	38

(百萬)

Correlation vs. Regression

1. A scatter plot cane be used to show the relationship between two variables
2. Correlation analysis is used to measure the strength of the association between two variables
   - Correlation is only concerned with strength of the relationship
   - No causal effect is implied with correlation

Introduction to Regression Analysis

Regression analysis is used to：

① Predict the value of a Y based on the value of at least one X.

② Explain the impact of changes in an X on the Y

- ***Y :***

  the variable we wish to predict or explain

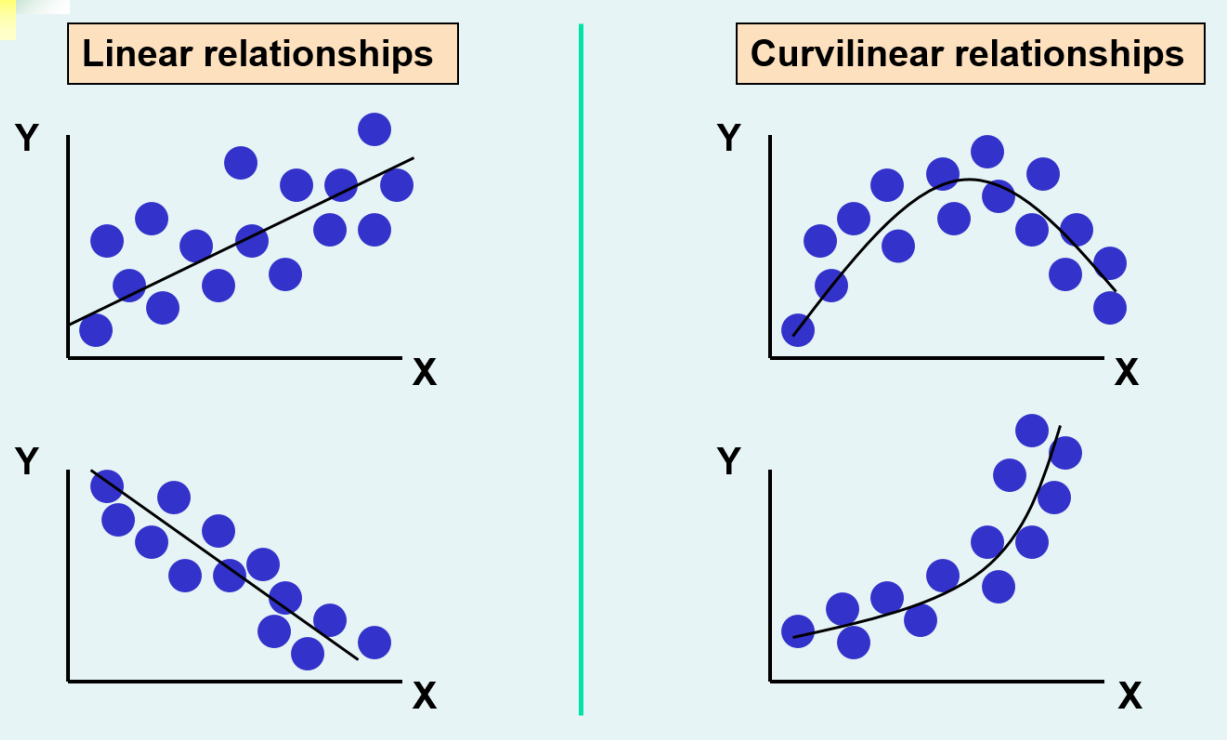
- ***X :***

  the variable used to predict or explain the dependent variable

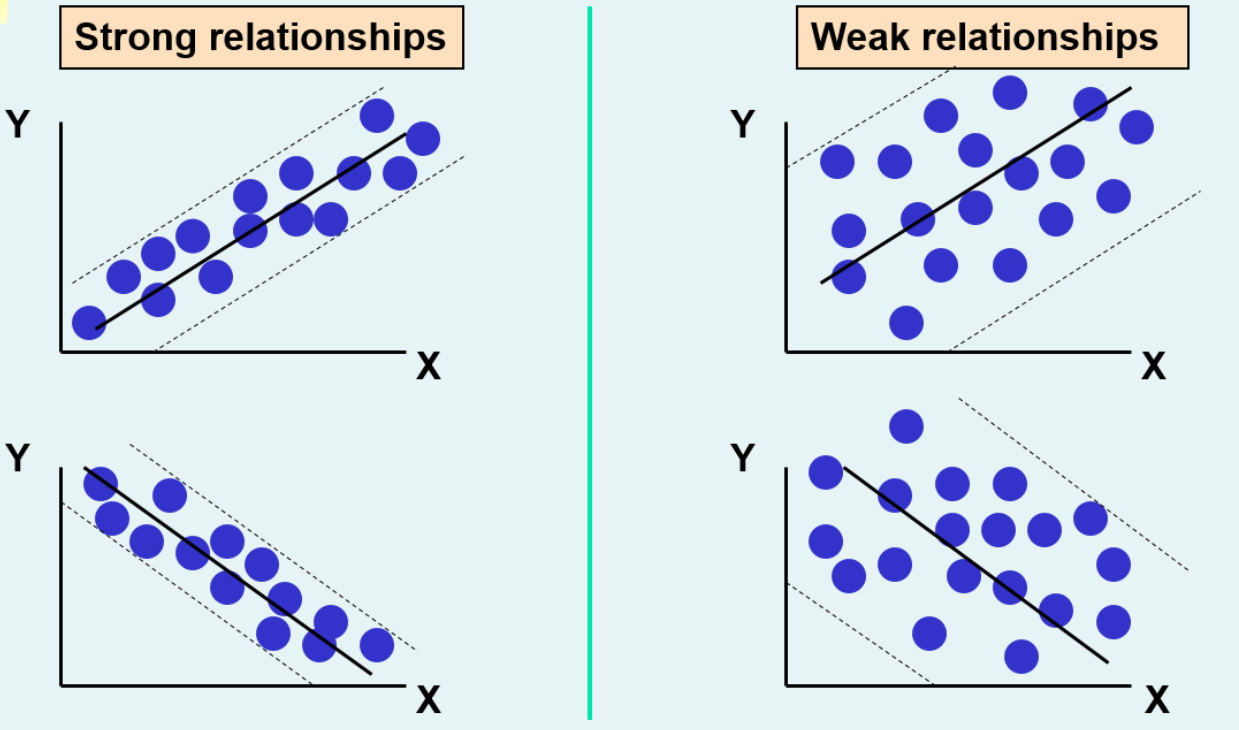
Simple Linear Regression Model

- Only ***one*** X
- Relationship between X and Y is described by a linear function
- Changes in Y are assumed to be related to changes in X

Type of Relationships

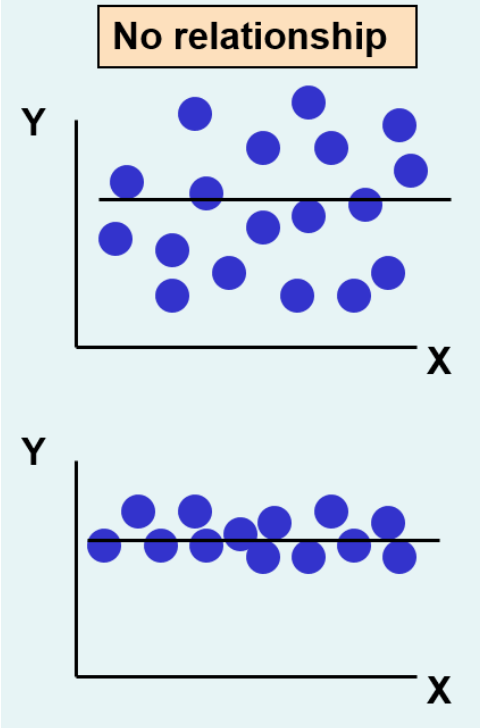


Type of Relationships (continued)

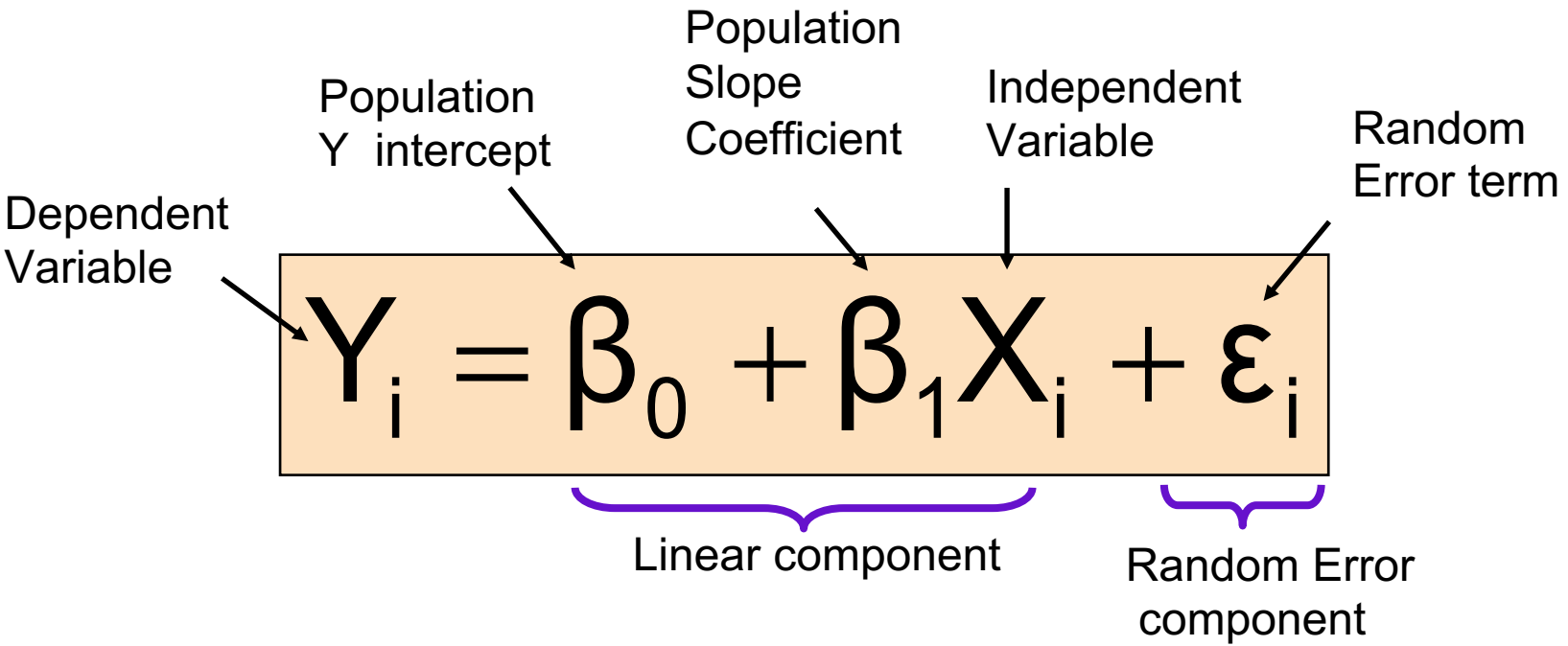


Type of Relationships

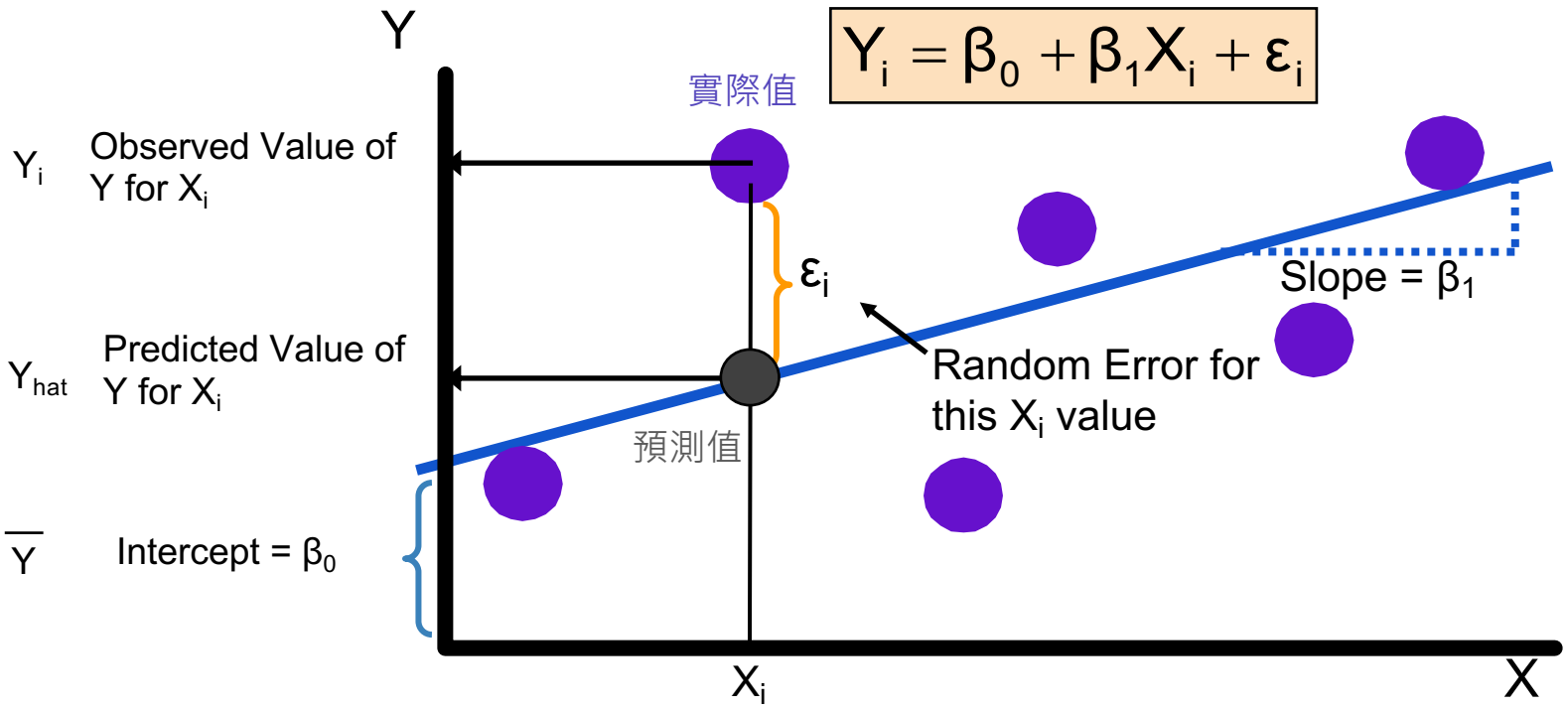
(continued)



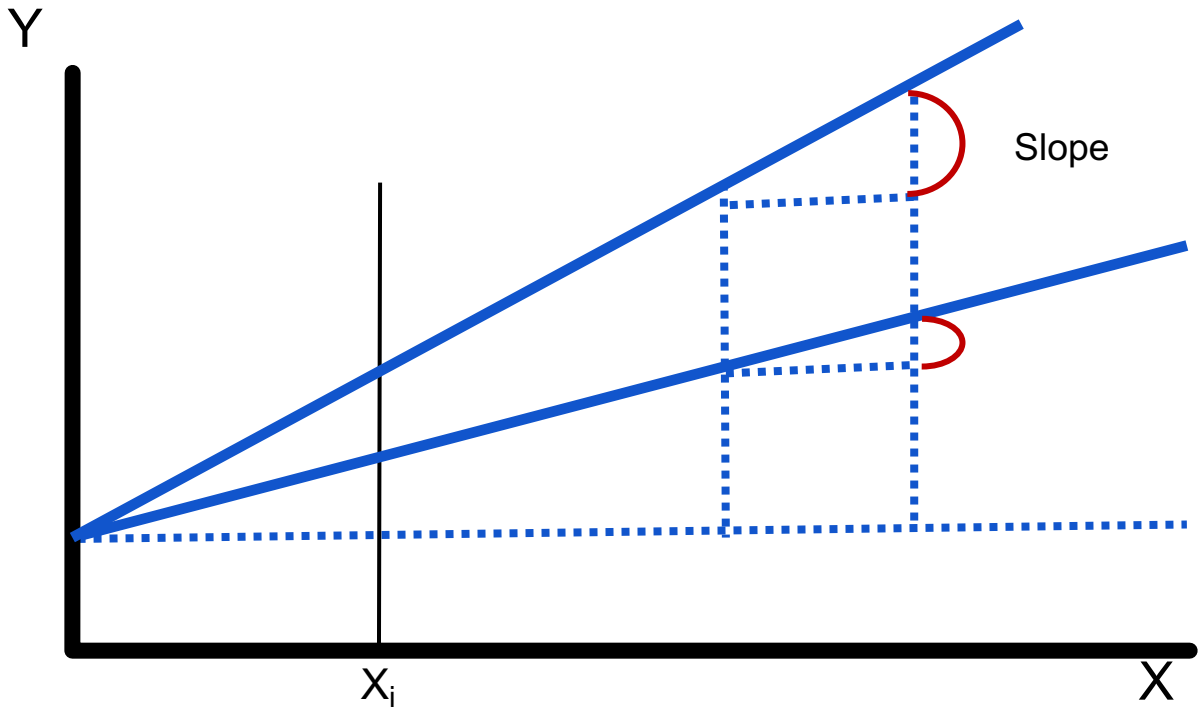
Simple Linear Regression Model



Simple Linear Regression Model



Simple Linear Regression Model



The Least Squares Method

- b_0 and b_1 are obtained by finding the values of that **minimize the sum of the squared differences** between Y and :

$$\min \sum (Y_i - \hat{Y}_i)^2 = \min \sum (Y_i - (b_0 + b_1 X_i))^2$$

Interpretation of the Slope and the Intercept

- b_0 is the estimated average value of Y when the value of X is zero
- b_1 is the estimated change in the average value of Y as a result of a one-unit change in X

Measures of Variation

Total variation is made up of two parts:

$$SST = SSR + SSE$$

Total Sum of Squares	Regression Sum of Squares	Error Sum of Squares
$SST = \sum (Y_i - \overline{Y})^2$	$SSR = \sum (\hat{Y}_i - \overline{Y})^2$	$SSE = \sum (Y_i - \hat{Y}_i)^2$

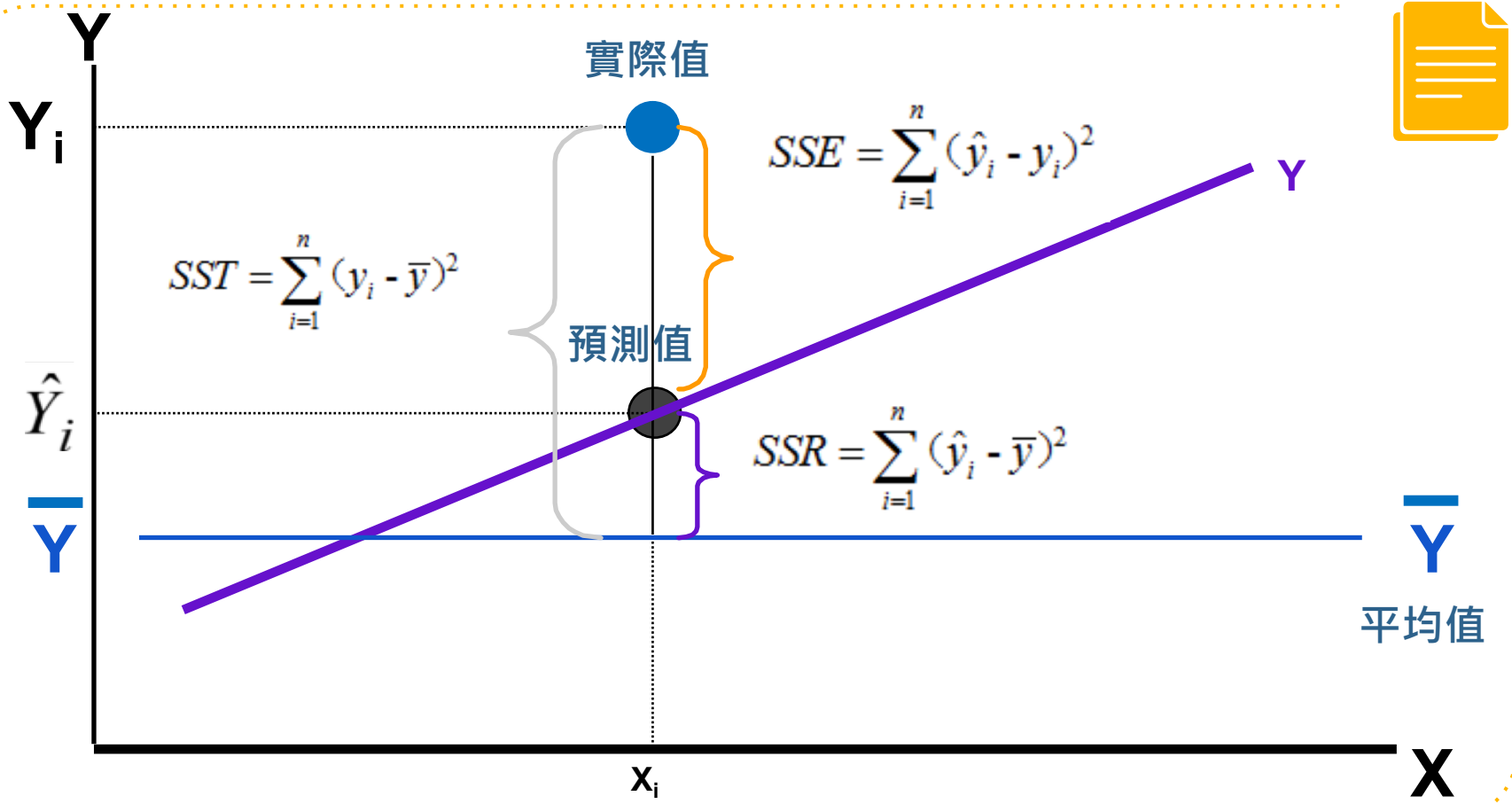
where:

$\overline{Y}$ = Mean value of the dependent variable

Y_i = Observed value of the dependent variable

$\hat{Y}_i$ = Predicted value of Y for the given X_i value

- **SST = total sum of squares** (Total Variation)
  - Measures the variation of the Y_i values around their mean Y
- **SSR = regression sum of squares** (Explained Variation)
  - Variation attributable to the relationship between X and Y
- **SSE = error sum of squares** (Unexplained Variation)
  - Variation in Y attributable to factors other than X


Y
Y_i
實際值
$SSE = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$
Y
$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$
預測值
$\hat{Y}_i$
$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$
$\bar{Y}$
$\bar{Y}$
平均值
x_i
X

Coefficient of Determination

R^2

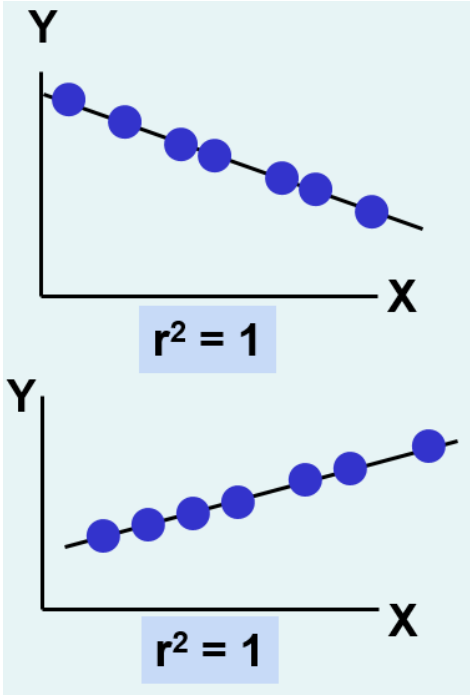
① The **coefficient of determination** is the portion of the total variation in the dependent variable that is explained by variation in the independent variable

② The coefficient of determination is also called **r-squared** and is denoted as **r^2**

$$r^2 = \frac{SSR}{SST} = \frac{\text{regression } sum \text{ of squares}}{total \text{ sum of squares}}$$

note: $0 \le r^2 \le 1$

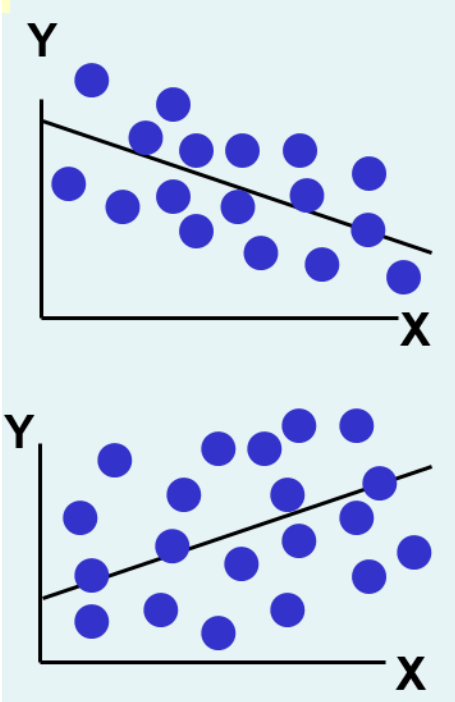
Examples of Approximate r^2 Values



$r^2 = 1$

- **Perfect linear relationship between X and Y :**
- **100% of the variation in Y is explained by variation in X**

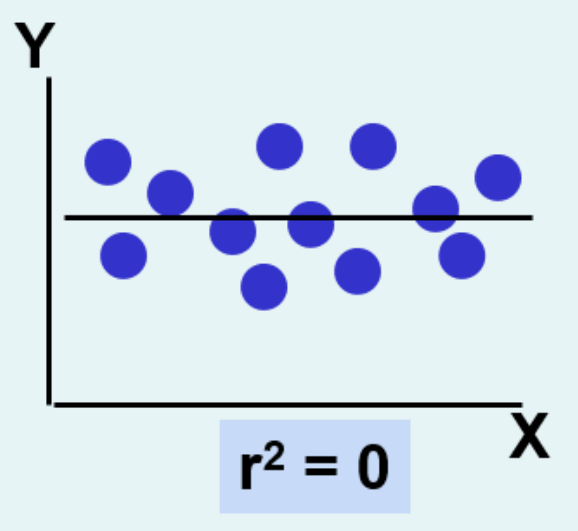
Examples of Approximate r^2 Values



$$0 < r^2 < 1$$

- **Weaker linear relationships between X and Y :**
- **Some but not all of the variation in Y is explained by variation in X**

Examples of Approximate r^2 Values



$r^2 = 0$

- No linear relationship between X and Y :
- The value of Y does not depend on X. (None of the variation in Y is explained by variation in X)

Assumptions of Regression L.I.N.E

① Linearity
- The relationship between X and Y is linear

② Independence of Errors
- Error values are statistically independent

③ Normality of Error
- Error values are normally distributed for any given value of X

④ Equal Variance (also called homoscedasticity)
- The probability distribution of the errors has constant variance

Thanks!