Finance Analytics

Chapter 2. 자료분석 시작하기 PART 2

권태연

한국외대 국제금융학과

어낼리틱스 수업에서는..

- 통계학적 분석도구를 사용하여 경영/경제/금융과 관련된 자료를 분석, 이론연결
- 이론? – 변수들 (현상들)간의 체계적으로 설명될 수 있는 연관성

ex. 수요 분석

sonata의 판매량은 sonat의 가격과 관련이 있다.

➡ sonata의 가격과 sonata의 판매량 간의 관계식을 만들 수 있다.

sonata의 판매량= f(sonata 의 가격) + 가격이 설명하지 못하는 부분

➡ 함수 f의 형태는? 간단한 형태부터 출발

➡ 가격만 고려해야 할까?

sonata의 판매량= f(sonata 의 가격, 경쟁사 자동차의 가격,…) + f() 즉 체계적 부분이 설명하지 못하는 부분

모형 (statistical model, econometric model)

모형의 체계적 부분

모형의 비체계적 부분
=오차 (error)

Model

- sonata의 판매량= f(sonata 의 가격, 경쟁사 자동차의 가격,…) + f() 즉 체계적 부분이 설명하지 못하는 부분

모형 (statistical model, econometric model)

모형의 체계적 부분

모형의 비체계적 부분
=오차 (error)

목적 변수 = 함수 (다른(원인) 변수들) + 설명되지 않는 부분

1. 목적 변수와 연관성(관련성)이 있는 원인 변수들을 찾아

2. 그들간의 관계를 수량화/수치화 → 모형화

→ 변수들 간의 관계를 살펴보자.

2. 두 변수의 관계를 그래프로 살펴보기

Scatter plot으로 살펴보기

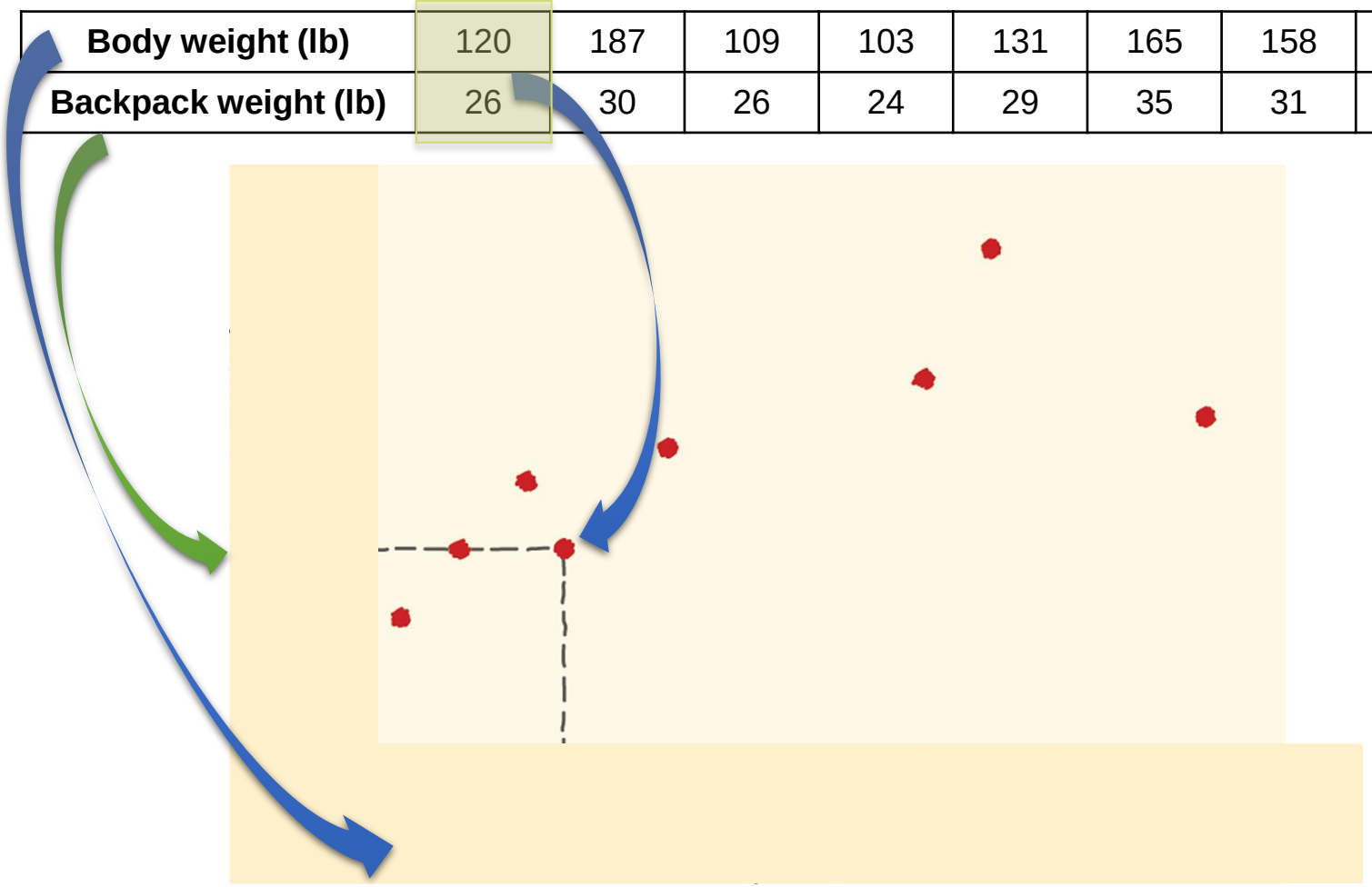
산포도, 산점도 (scatterplot)

- 동일한 개체에 대해 측정한 두 개 정량변수 사이의 관계를 보여주는 그래프
- 한 변수의 값은 수평축에 나타내고 다른 변수의 값은 수직축에 나타낸다.
- 자료 상의 각 개체들은 해당 개체에 대한 두 개 변수의 값으로 결정된 도표 상의 점으로 나타낸다.

Scatter plot으로 살펴보기

Example: Make a scatterplot of the relationship between body weight and pack weight for a group of hikers.

Body weight (lb)	120	187	109	103	131	165	158	116
Backpack weight (lb)	26	30	26	24	29	35	31	28



Interpreting Scatterplots

How to Examine a Scatterplot

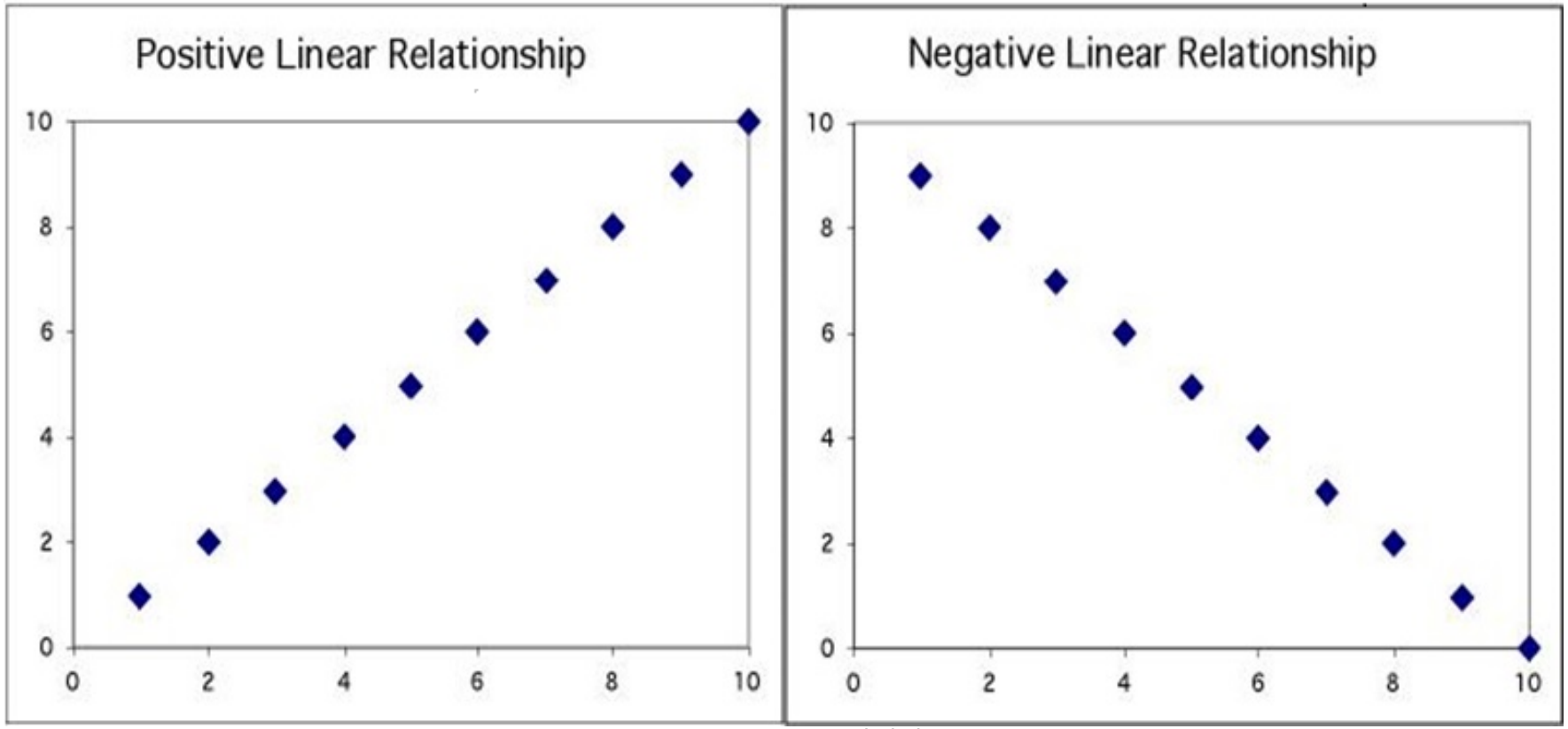
As in any graph of data, look for the *overall pattern* and for striking *departures* from that pattern.

•두 변수 간 관계의 방향(direction), 형태(form), 강도(strength)를 통해 산포도의 전반적인 패턴을 설명

•이탈현상 중 중요한 것은 이탈값(outlier)으로 전반적인 패턴 밖에 위치하는 개체의 값.

Scatter plot을 통해 알 수 있는 두 변수의 관계 :

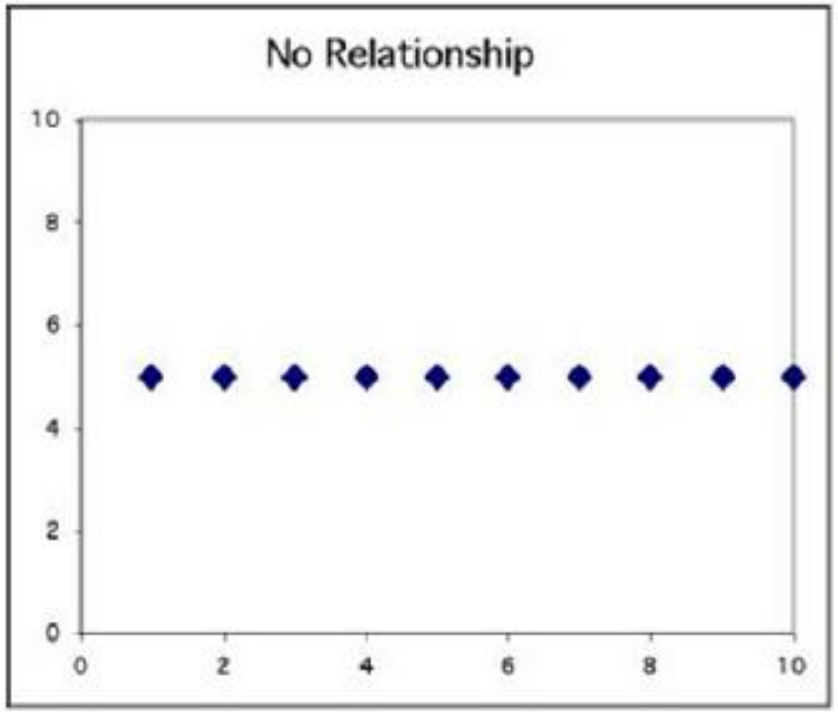
1. 관계의 방향성: 양의 관계 및 음의 관계

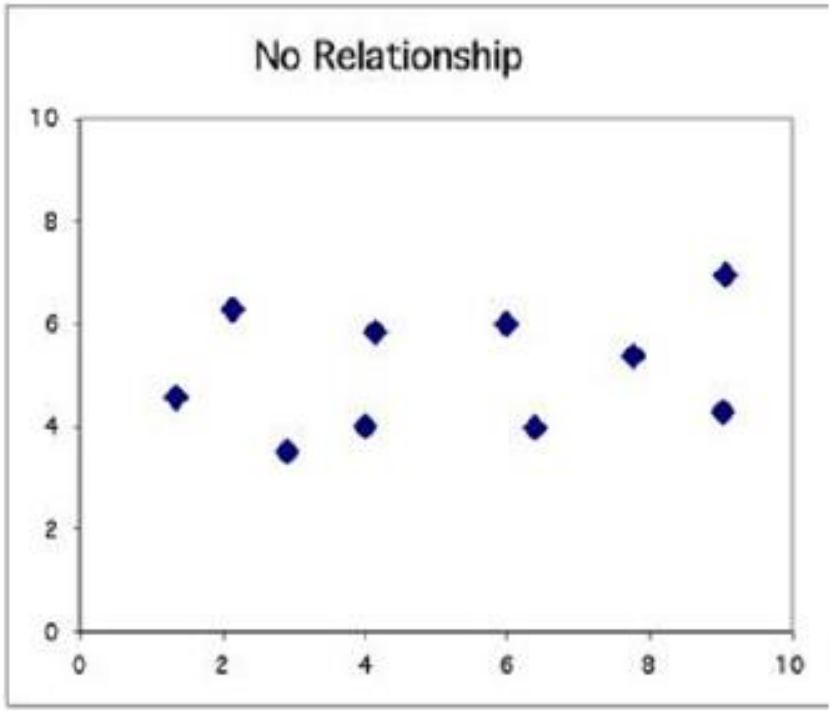


Scatter plot을 통해 알 수 있는 두 변수의 관계 :

1. 관계의 방향성: No relationship

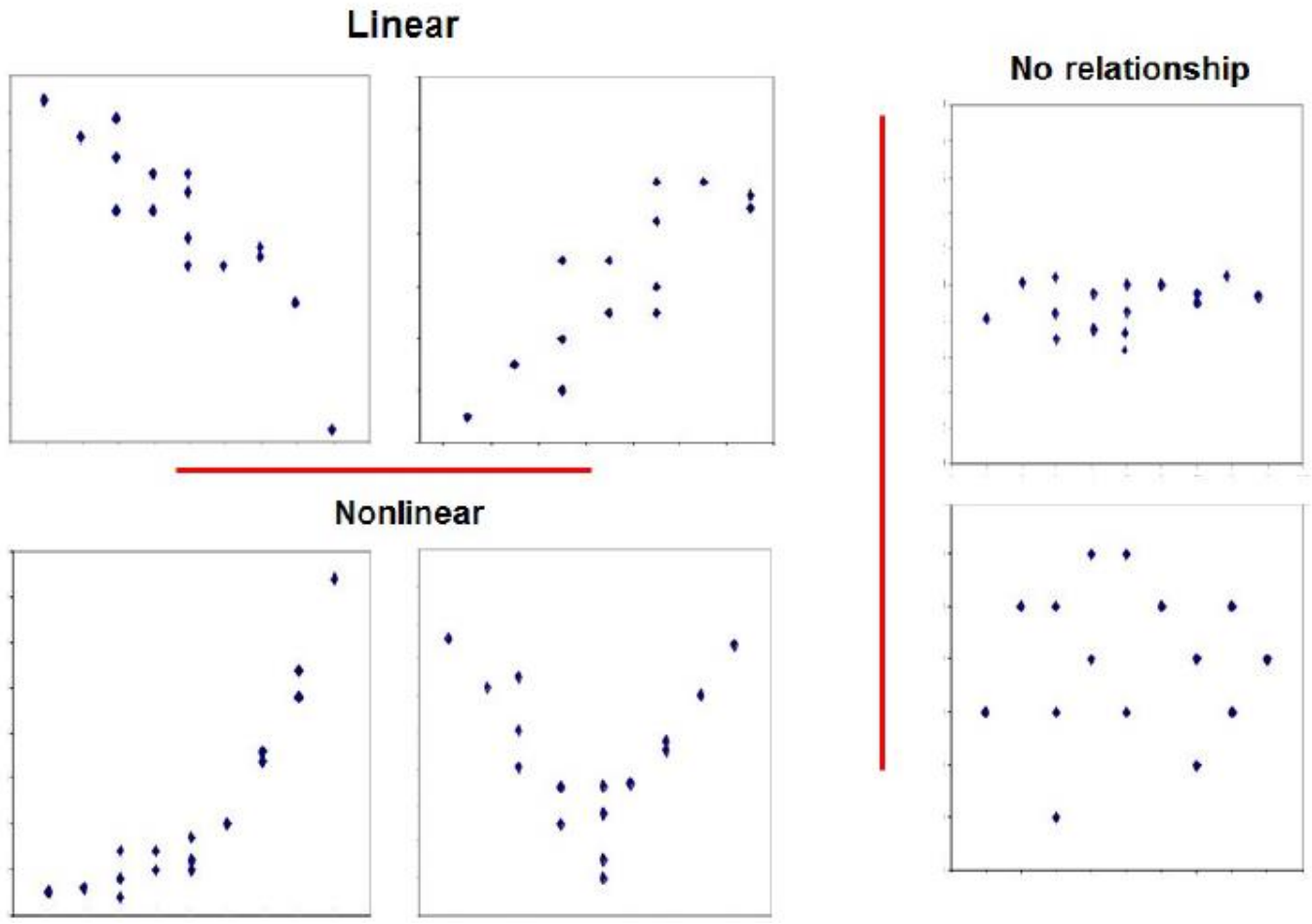
- No relationship: X and Y vary independently. Knowing X tells you nothing about Y.





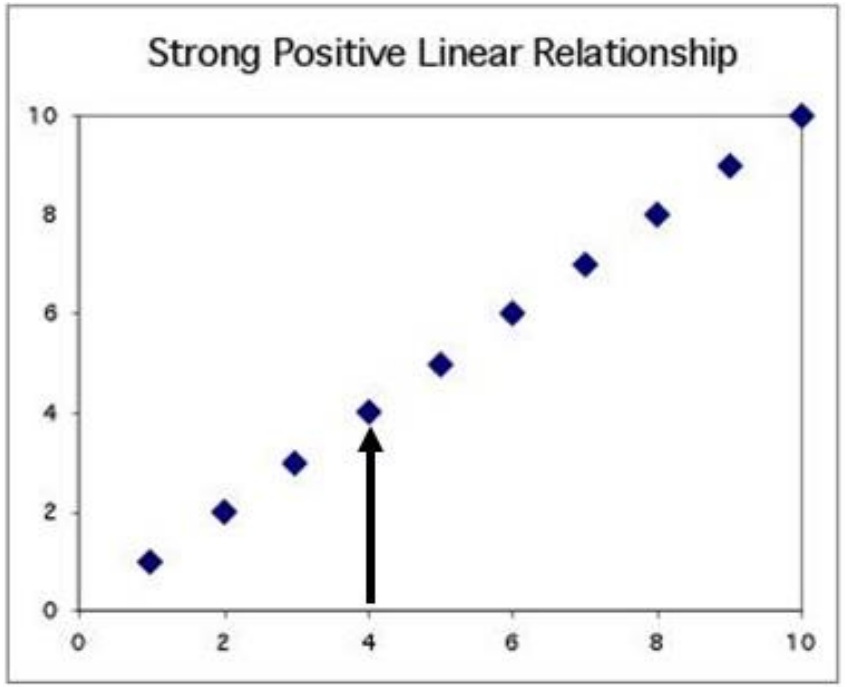
Scatter plot을 통해 알 수 있는 두 변수의 관계 :

2. Form of Relationship

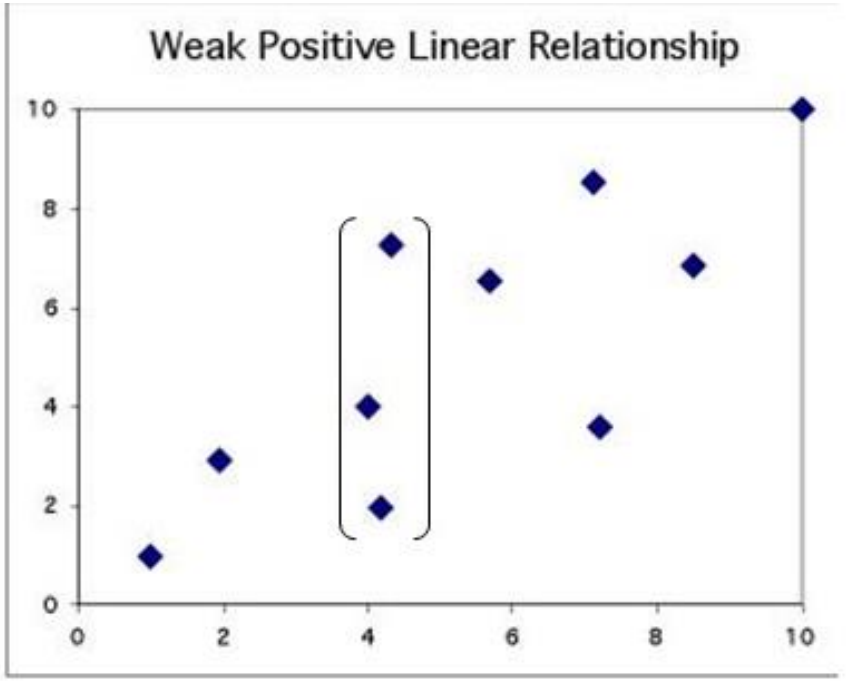


Scatter plot을 통해 알 수 있는 두 변수의 관계 :

3. Strength of Relationship



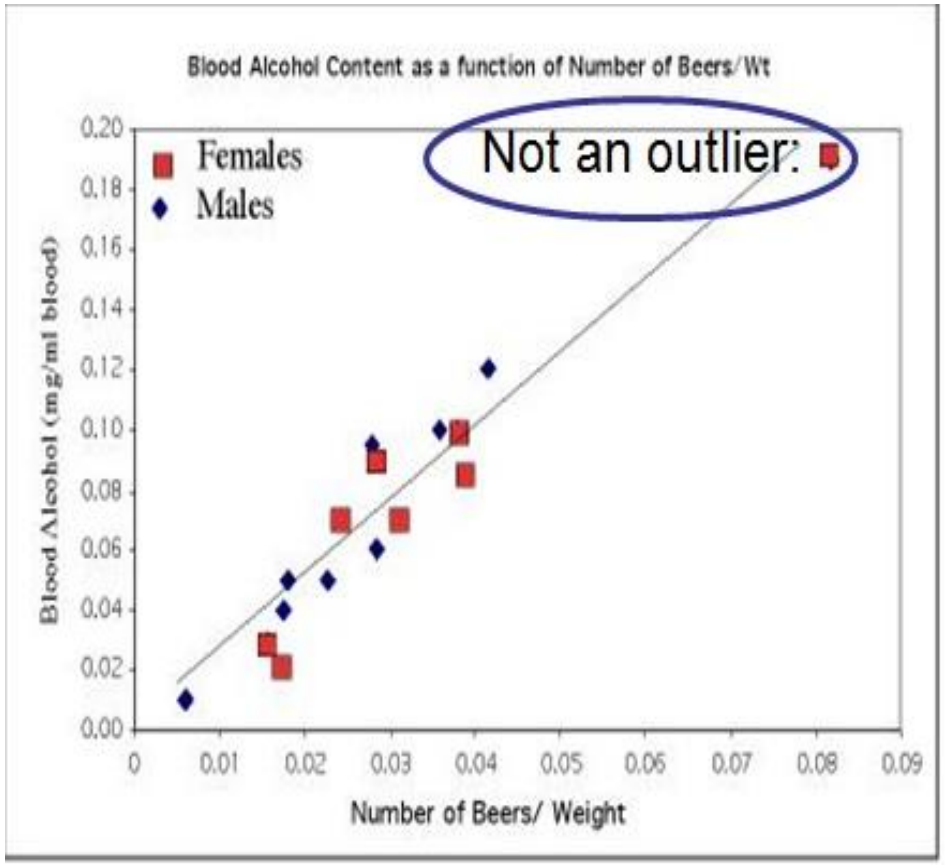
With a strong relationship, you can get a pretty good estimate of *y* if you know *x*.

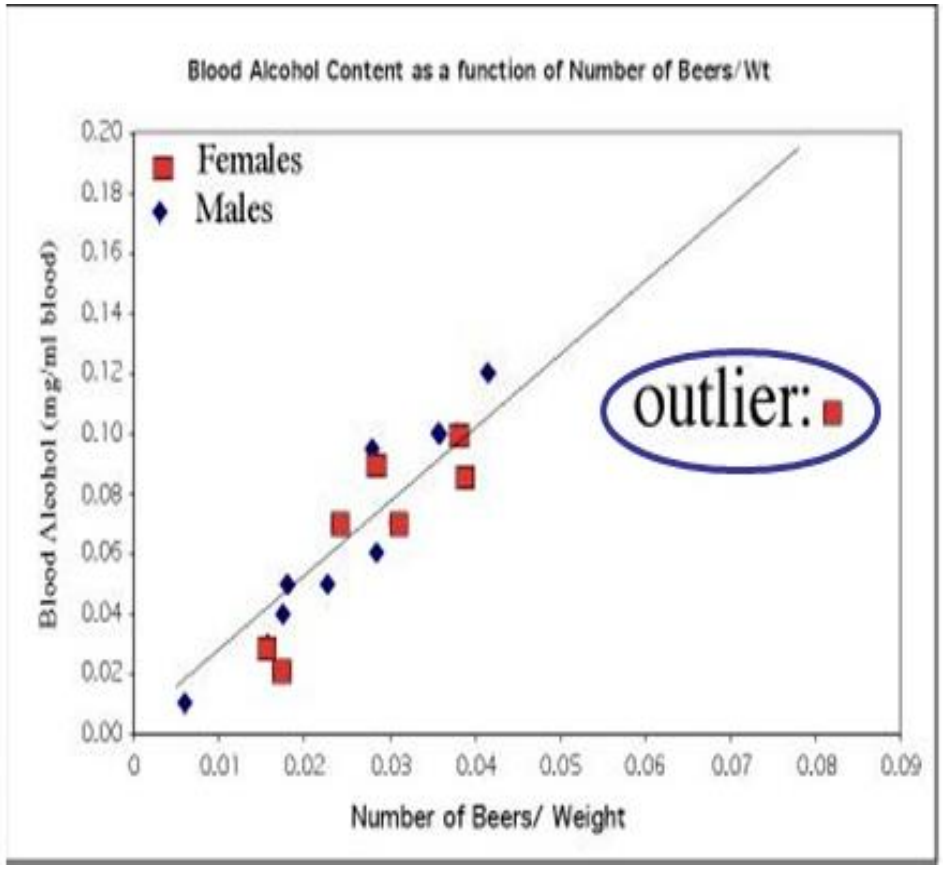


With a weak relationship, for any *x* you might get a wide range of *y* values.

Scatter plot을 통해 알 수 있는 두 변수의 관계 :

4. *departures* from that pattern: outlier





3. 두변수의 관계 수량화 하기

: 상관계수

두 변수의 관계 수량화 하기:
1. 상관계수

- A scatterplot displays the strength, direction, and form of the relationship between two quantitative variables.

- 상관계수 (correlation coefficient) 두 개의 정량변수 사이에 존재하는 "선형관계"의 방향(direction) 및 강도(strength)를 측정한다.

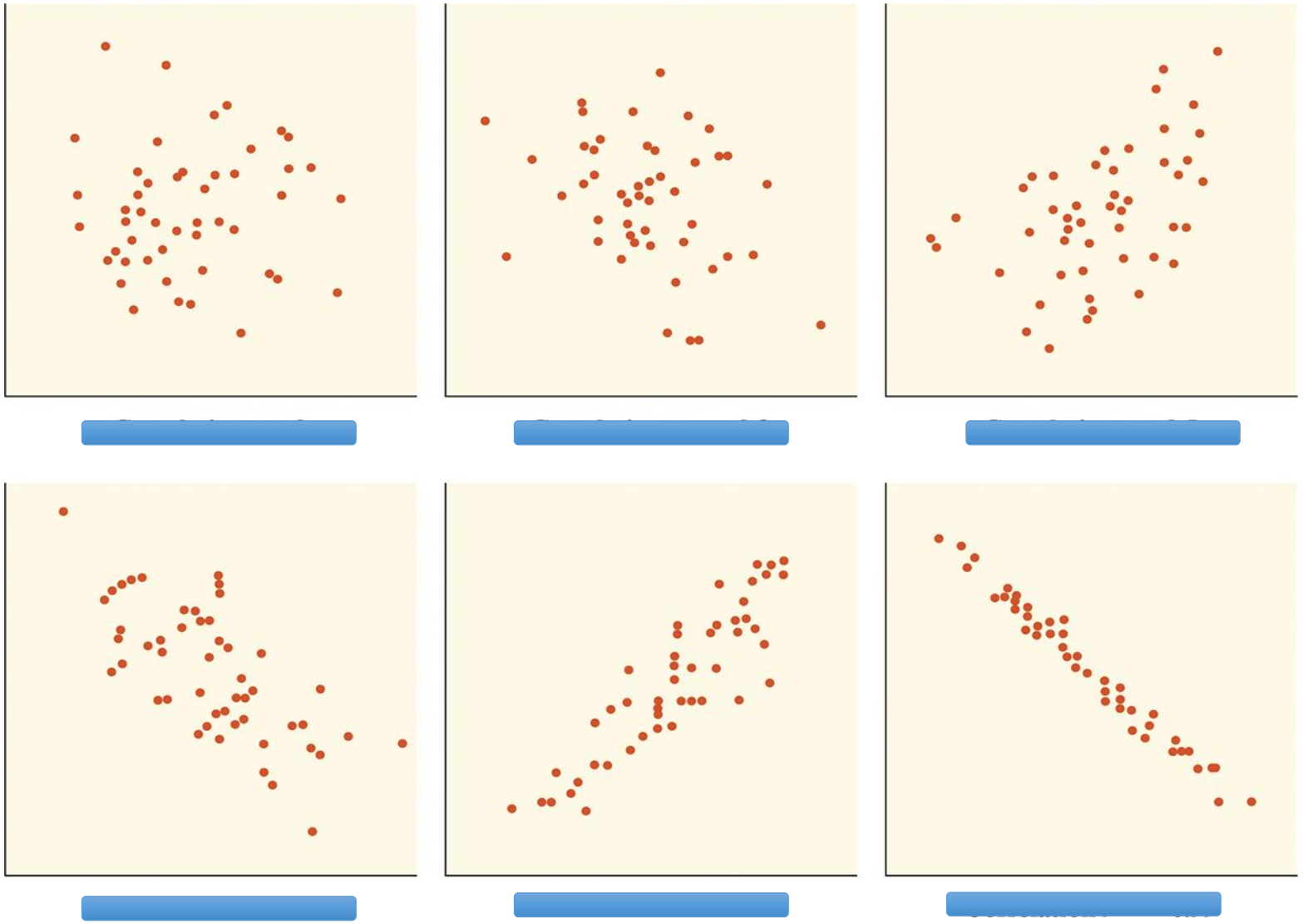
상관계수 (correlation, correlation coefficient)

The **correlation *r*** measures the strength of the linear relationship between two quantitative variables.

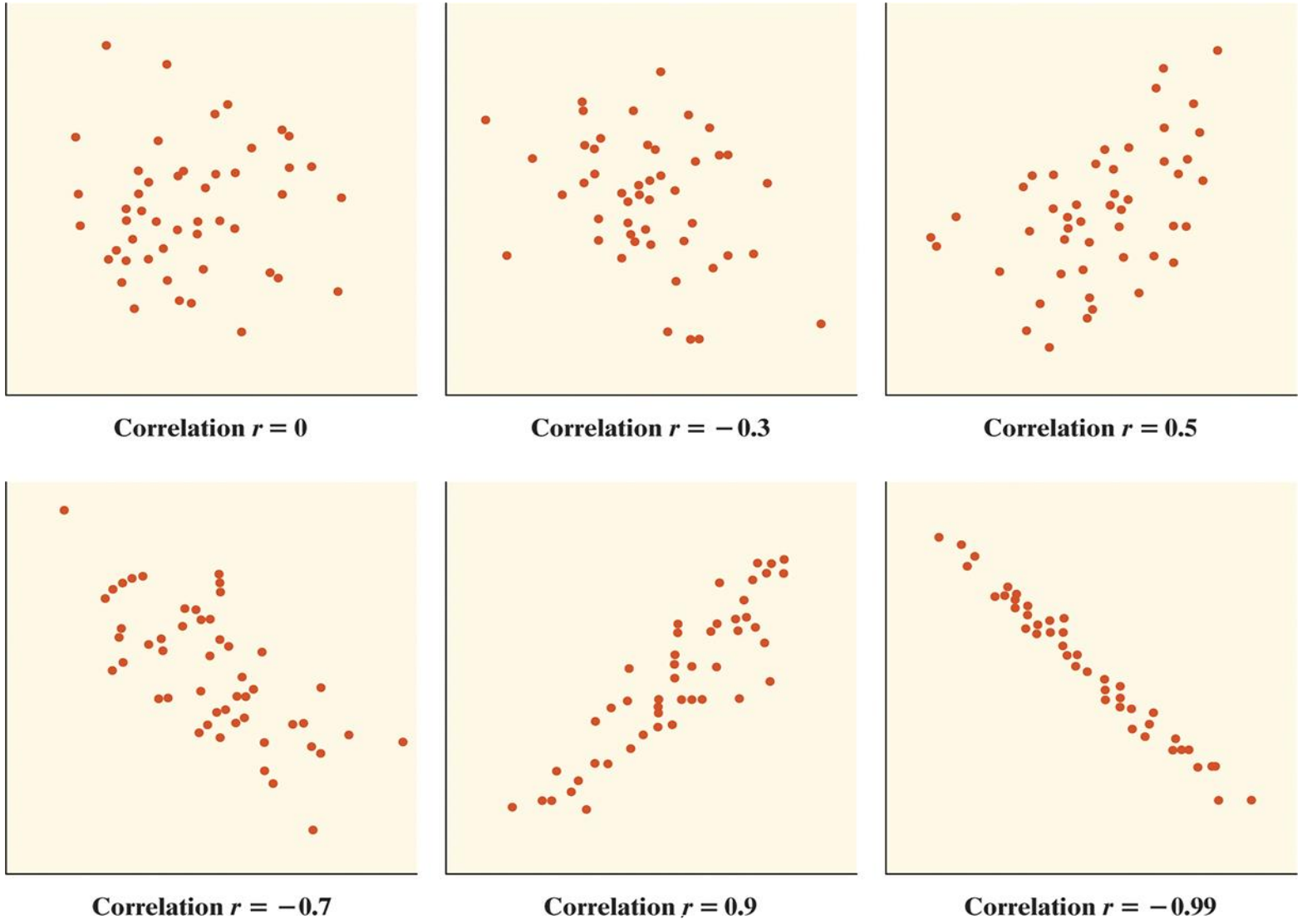
$$r = \frac{1}{n-1}\sum \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

- *r* is always a number between -1 and 1.
- $r > 0$ indicates a positive association.
- $r < 0$ indicates a negative association.
- Values of *r* near 0 indicate a very weak linear relationship.
- The strength of the linear relationship increases as *r* moves away from 0 toward -1 or 1.
- The extreme values $r = -1$ and $r = 1$ occur only in the case of a perfect linear relationship.

Scatter plot vs Correlation coefficient?
-0.3, -0.7, -0.99, 0, 0.5, 0.9



Scatterplot and Correlation r



Facts About Correlation

1. Correlation makes no distinction between explanatory and response variables.
2. *r* has no units and does not change when we change the units of measurement of *x*, *y*, or both.
3. Positive *r* indicates positive association between the variables, and negative *r* indicates negative association.
4. The correlation *r* is always a number between -1 and 1.

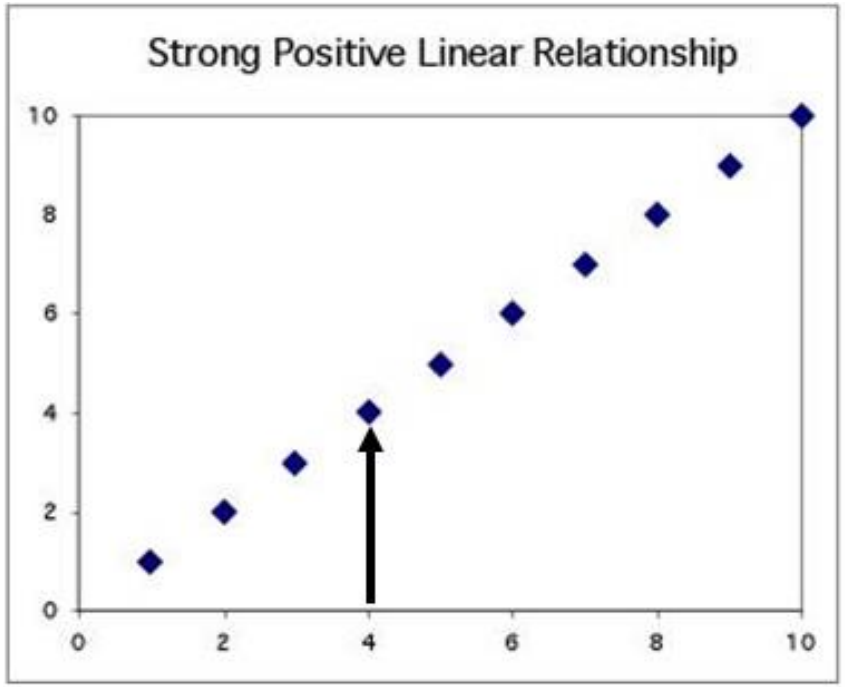
Cautions:

- Correlation requires that both variables be quantitative.
- Correlation does not describe curved relationships between variables, no matter how strong the relationship is.
- Correlation is not resistant. *r* is strongly affected by a few outlying observations.
- Correlation is not a complete summary of two-variable data.

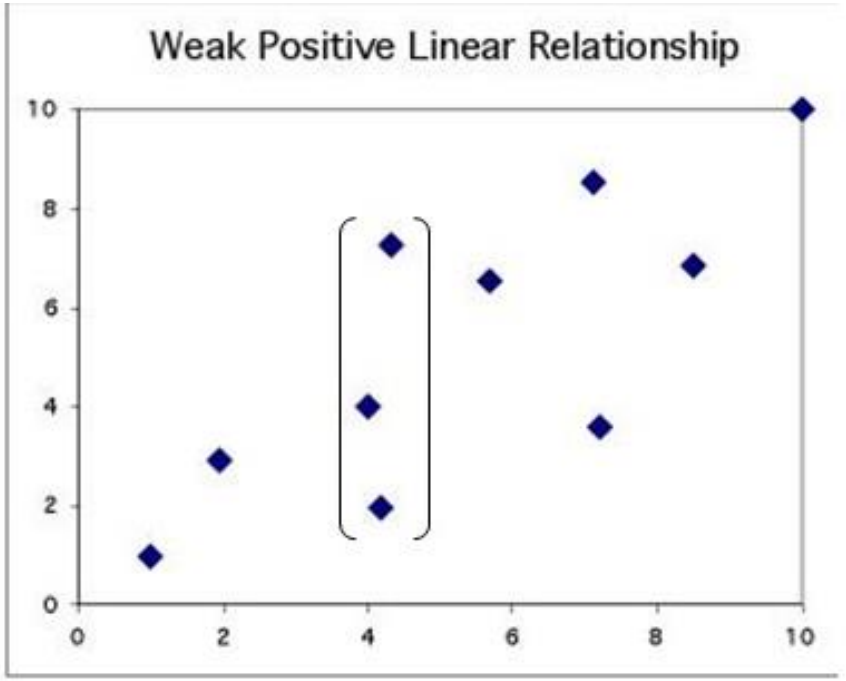
4. 두변수의 관계 모형화 하기

: 선형 회귀모형
(Linear Regression Model, LRM)

Recall: Scatter plot을 통해 알 수 있는 두 변수의 관계 : Strength of Relationship



With a strong relationship, you can get a pretty good estimate of *y* if you know *x*.

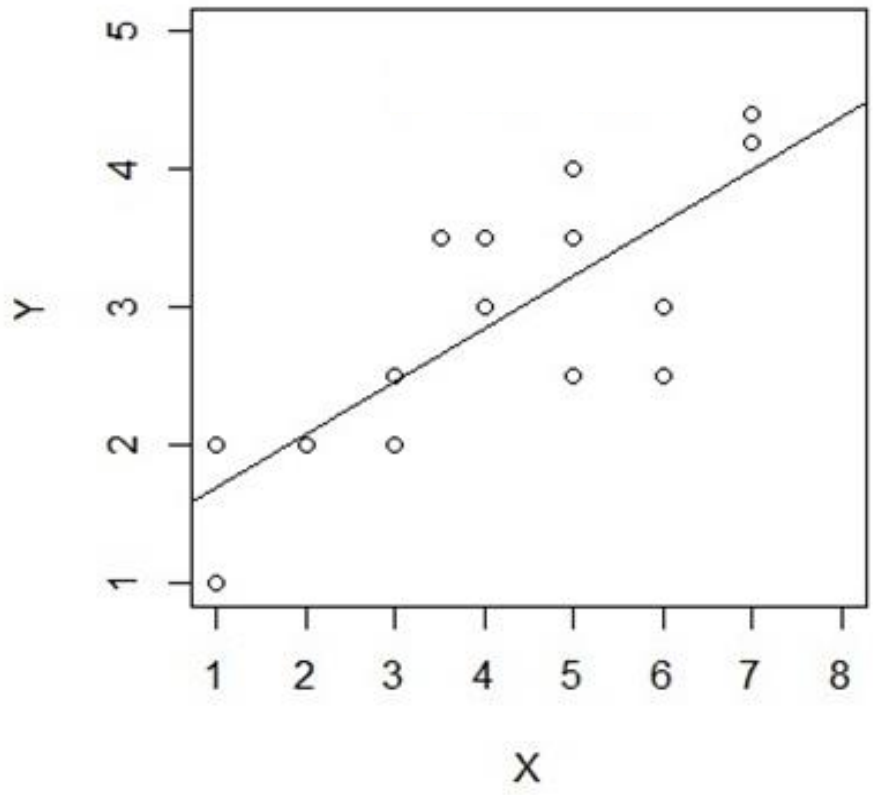


With a weak relationship, for any *x* you might get a wide range of *y* values.

선형 회귀선

- 선형 회귀선(linear **regression line)은** 설명변수 x가 변화함에 따라 반응변수 y가 "평균적으로" 어떻게, 얼마만큼 변화하는지를 보여주는 직선

- 주어진 x값에 대해 y값을 예측하는데 회귀선(regression line)을 이용한다.



y = a + bx

좀더 통계적으로 명확한 표현

$$\hat{y} = a + bx$$

or

$$y = a + bx + \varepsilon$$

주의: 선형회귀모형에서의 종속/독립변수

반응변수 or 종속변수 or Y	설명변수 or 독립변수 or X
정량 (& 연속형)	정량
정량 (& 연속형)	범주
범주	정량
범주	범주