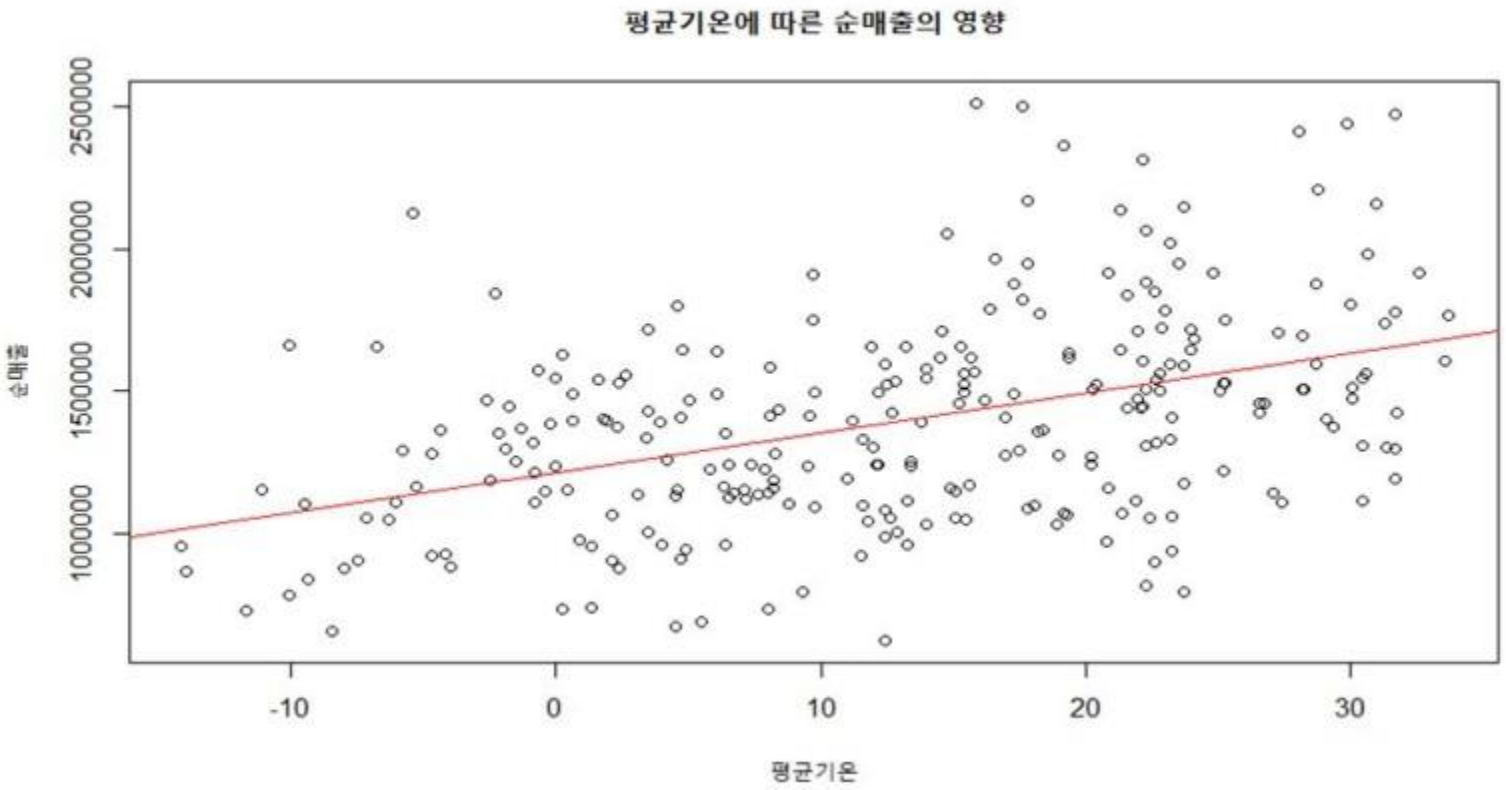
Chapter 3

상관분석과 회귀분석

Chapter 3

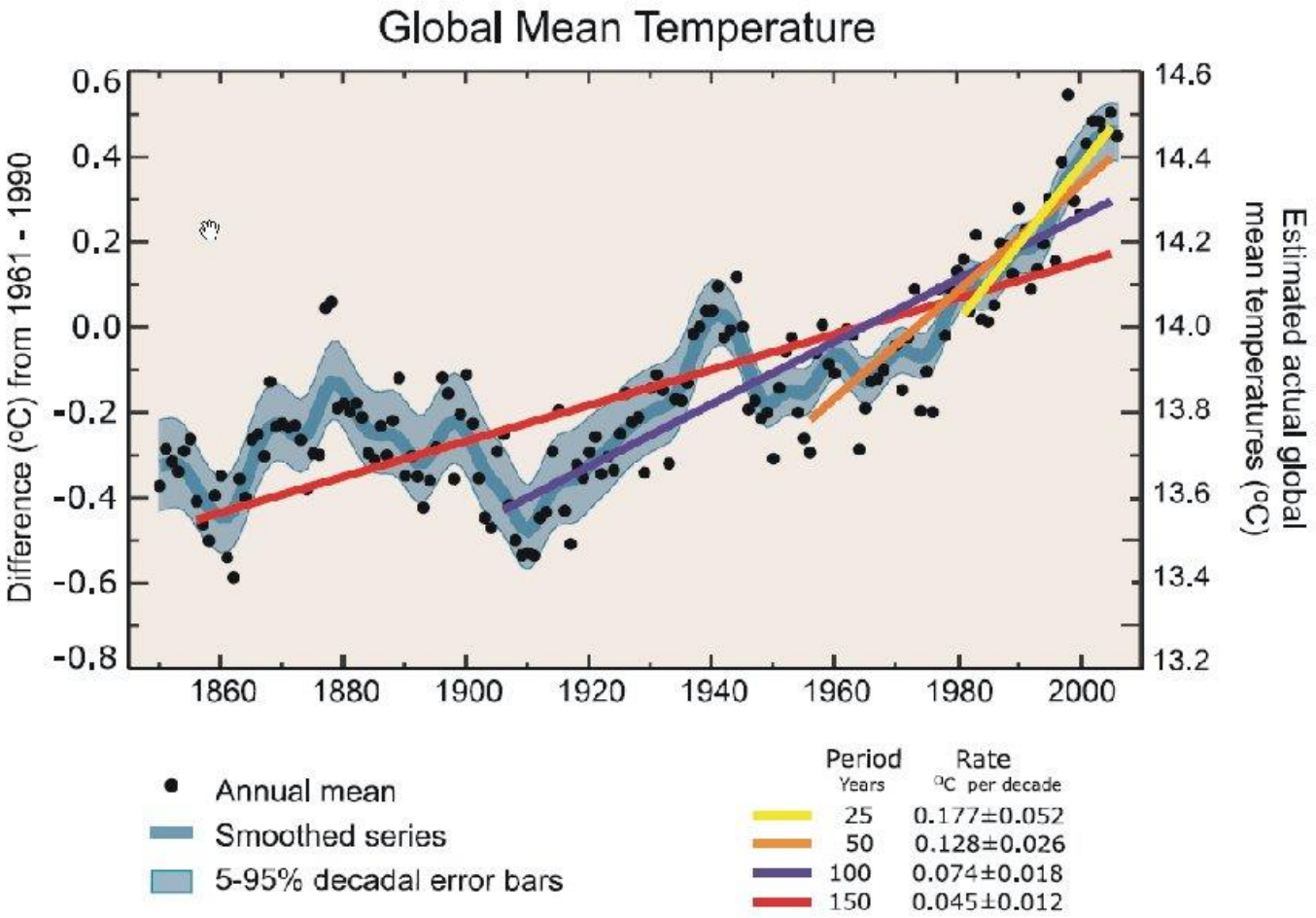
Contents

3.2 회귀분석



From https://m.blog.naver.com/kmiti/221833297598

3.2 회귀분석


Global Mean Temperature
Difference (°C) from 1961 - 1990
Estimated actual global mean temperatures (°C)
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
14.6
14.4
14.2
14.0
13.8
13.6
13.4
13.2
1860
1880
1900
1920
1940
1960
1980
2000
Annual mean
Smoothed series
5-95% decadal error bars
Period
Years
Rate
°C per decade
25
0.177±0.052
50
0.128±0.026
100
0.074±0.018
150
0.045±0.012

3.2 회귀분석

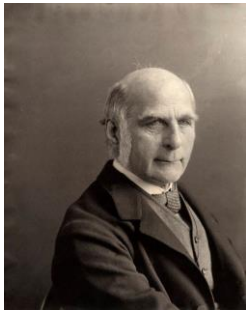
❖ 회귀분석(regression analysis)이란?

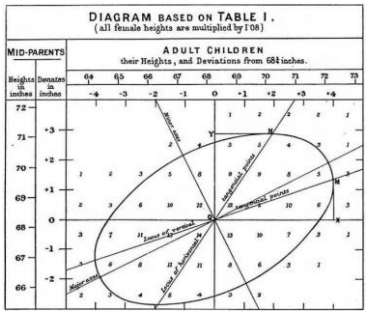
- 둘 이상의 변수 간의 관계를 보여주는 통계적 방법
- 관찰된 연속형 변수들에 대해 두 변수 사이의 모형을 구한 뒤 적합도를 측정해 내는 분석 방법 (by wiki)
- '독립변수' 또는 '예측인자(predictor)'라 부르는 변수 X와 '종속변수' 또는 '예측량(predictand)'라 부르는 변수 Y의 관계를 함수식으로 설명하는 통계적 방법 (by textbook)
- **단순(simple) 회귀분석**: 하나의 종속변수와 하나의 독립변수 사이의 관계를 규명하는 회귀분석
- **다중(multiple) 회귀분석**: 하나의 종속변수와 여러 독립변수 사이의 관계를 규명하는 회귀분석
- 선형(linear) 회귀분석: 종속변수와 독립변수 사이의 관계를 선형관계로 가정함으로써, 둘 간의 함수식을 1차 선형 방정식으로 표현함

3.2 회귀분석

❖ 회귀(regression)의 어원

- 회귀분석은 생물학자 프랜시스 골턴(Francis Galton)이 '평균으로의 회귀(regression toward the mean' 현상을 증명하기 위해 만든 것으로 알려져 있음
- 부모의 키와 아이들의 키 사이의 연관 관계를 연구하면서 부모와 자녀의 키 사이에는 선형적인 관계가 있고 키가 커지거나 작아지는 것보다는 전체 키 평균으로 돌아가려는 경향이 있다는 가설을 세움
- 회귀의 원 의미는 옛날 상태로 돌아가는 것을 의미하고, 골턴은 평균으로 돌아간다는 의미로 사용했지만 현대에 와서 회귀의 이 의미는 거의 사라짐





3.2 회귀분석

찰스 다윈에게는 그만큼이나 독특한 사상으로 무장한 사촌 동생 프랜시스 골턴 (Francis Galton, 1822~1911)이 있었다. 두 사람은 서로를 존경하며 각자의 위치에서 자신의 연구 분야를 공고히 했다.

골턴은 그의 사촌 형인 다윈에게서 많은 영향을 받았는데, 하루는 다윈의 『종의 기원』을 읽고 유전자라는 어마어마한 영향력을 가진세계를 접하게 됐다. 그리고 유전자가 중요하며 우월한 집안에서 우월한 유전자가 나온다는 결론에 도달했다.

골턴은 훌륭한 사람은 그가 처한 환경보다 유전자에 많은 영향을 받는다고 확신했다(우생학). 그래서 이러한 유전적 우월성을 구체적으로증명하기 위해 그의 사촌 형 다윈처럼 주변 사람들의 키를 전수 조사하러 다녔다. 골턴의 주장은 아버지의 키가 크면 자식도 아버지만큼 키가 크는 유전자를 물려받는다는 것이었다. 골턴이 조사해 보니 아버지가 키가큰 아이들이 또래보다 키가 크다는 사실을 알게 됐다.

그러나 뭔가 의심쩍은 부분을 발견했다. 키가 큰 아버지의 자식들이 또래보다 키가 크긴 했지만 아버지만큼 크지는 못한다는 사실이었다.

골턴은 키가 큰 사람의 자식이 부모보다 더 커지면 키 큰 유전자를 물려받는 자손은 끝도 없이 자랄 것이고, 반대로 키가 작은 집안의 자손들은 계속 작아지게 될 테니 적정 수준까지 큰다고 보았다. 그리고 사람들이 얼마까지 크는가를 고민했다.

조사 결과를 살펴보던 골턴은 놀라운 사실을 발견했다. 그가 조사한 대상 세대별 평균 키를 구하고 전체 대상의 키를 해당 평균을 기준으로 점을 찍어 분포를 확인했더니 아버지의 키가 아무리 커도 자식의 키는 평균보다는 크지만 해당 세대 평균에 가깝게 분포했다(중심극한정리). 즉, 키가 큰 아버지는 그보다 조금 작은 자식을, 키가 작은 아버지는 그보다 조금 큰 자식을 갖게 된다는 결과였다. 골턴은 이 놀라운 발견을 평균으로의 회귀(regression toward mean)라는 이름으로 공표했다.

From https://brunch.co.kr/@plusstar/139

❖ 단순 선형회귀 분석 (simple linear regression analysis)

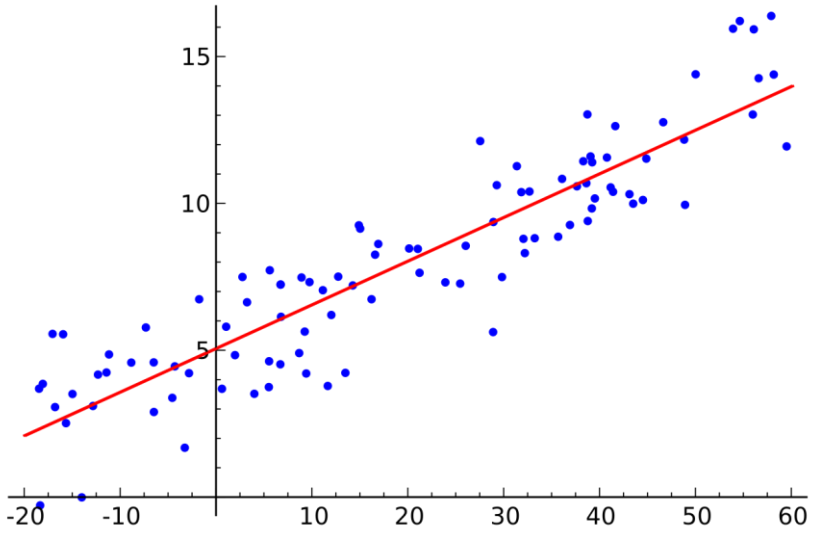
- 두 변수 X와 Y 간의 관계를 하나의 직선으로 표현하는 분석으로서 교과서 그림 6.5는 이 방법을 도식적으로 보인 것임

- 각 점(자료값)으로부터 직선까지의 거리(ε)의 제곱을 최소로 하는 직선

$$\hat{Y} = \alpha + \beta X$$

을 찾는 것이 회귀분석의 첫 과정임

- 즉, 관측값 x가 주어질 때 예측량 y에 대해 최소 오차를 만드는 과정임



3.2 회귀분석

- 일반적으로 가우스의 **최소제곱법(method of least squares)**을 사용하여 모수 α와 β를 추정하는데, 이 α와 β를 "**회귀계수(regression coefficients)**"라고 하며, 이들의 추정치를 a와 b로 표기함
- 자료값과 회귀식 간의 오차인 ε의 제곱합

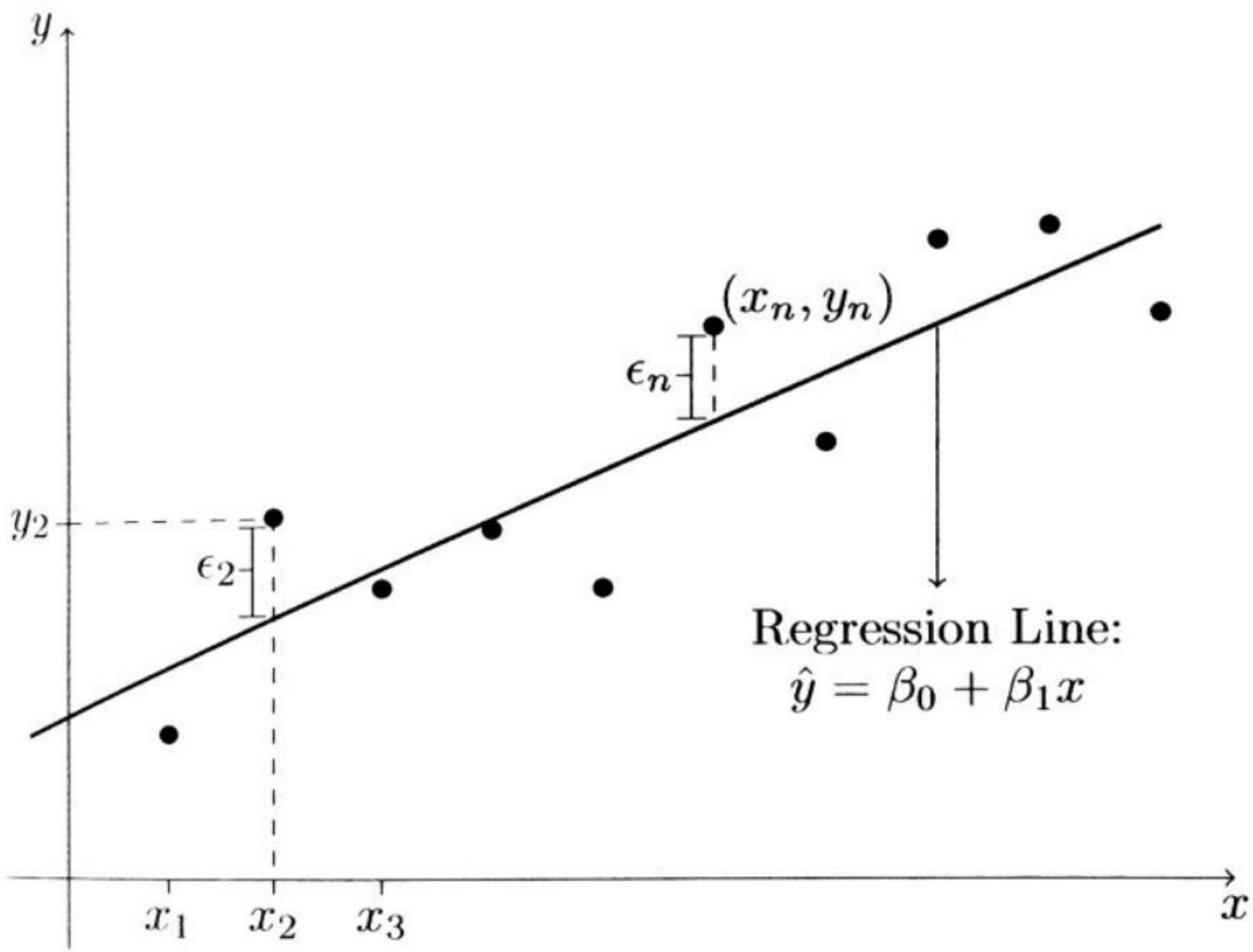
$$\sum_{i=1}^{n}(\epsilon_i)^2 = \sum_{i=1}^{n}(y_i - [\alpha + \beta x_i])^2$$

을 최소로 하는 직선 $\hat{y} = \alpha + \beta x$를 찾는 것임

- 이를 위해 위 식을 회귀계수 α, β에 대해 각각 편미분하여 0으로 놓고 연립방정식을 풀면, 최소제곱 추정치 a, b는 다음 식으로 결정

$$b = \frac{\sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})]}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad a = \bar{y} - b\bar{x}$$

3.2 회귀분석


y
(x_n, y_n)
ϵ_n
y_2
ϵ_2
Regression Line:
$\hat{y} = \beta_0 + \beta_1 x$
x_1
x_2
x_3
x

3.2 회귀분석

- $\hat{y}$는 단순 선형회귀 모형에 의한 예측값(즉 예측량)이 되는 셈임
- 각 점으로부터 직선까지의 거리 ε은 회귀모형 예측의 오차인데, "**잔차(residual)**"라 부름:

$$\epsilon_i = y_i - \hat{y}(x_i) = y_i - (a + bx_i)$$

- 또는 예측량의 참값은 회귀모형에 의한 "예측된 예측량"과 잔차의 합으로 이루어짐:

$$y_i = \hat{y}(x_i) + \epsilon_i = (a + bx_i) + \epsilon_i$$

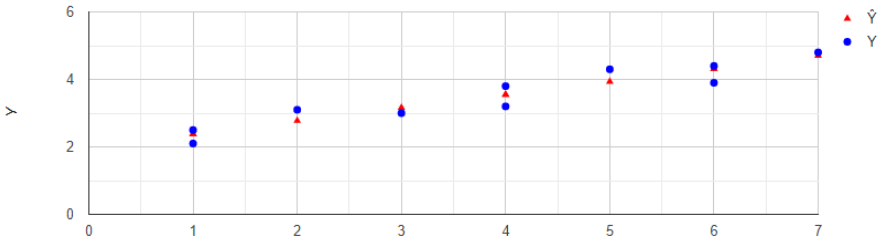
3.2 회귀분석

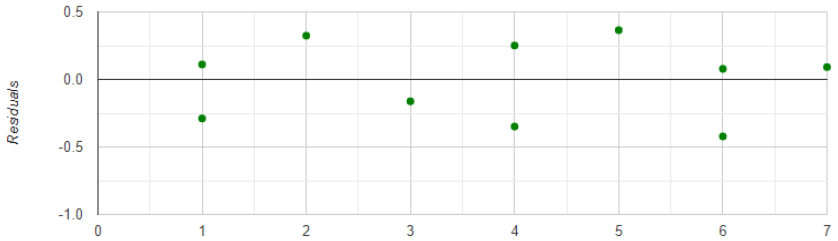
Quiz(단순 선형회귀): 어떤 화학 반응에서 생성되는 반응량(y)이 이 반응에 첨가되는 촉진제의 양(x)에 따라 어떻게 변화하는가를 10회 실험하여 다음 측정 결과를 얻게 되었다. 선형 회귀분석하여 회귀계수를 구하시오. 또한 이 회귀모형을 사용하여 촉진제 8g에 대한 반응량을 예측하시오.

촉진제 양 (g) {1, 1, 2, 3, 4, 4, 5, 6, 6, 7}

반응량 (g) {2.1, 2.5, 3.1, 3.0, 3.8, 3.2, 4.3, 3.9, 4.4, 4.8}

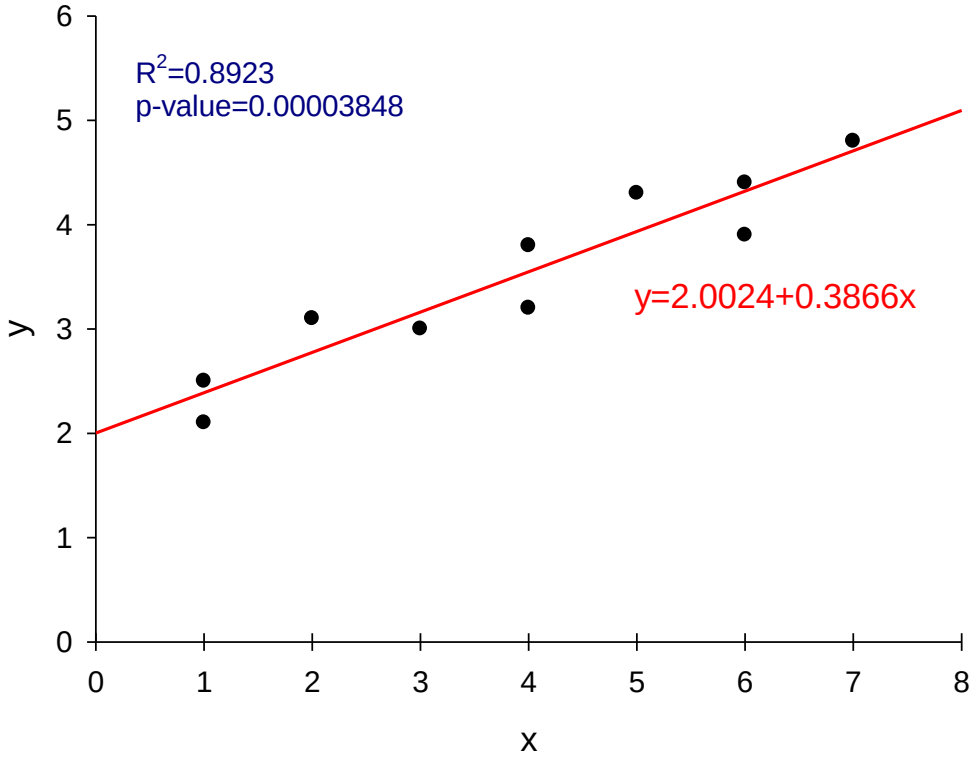
Answer: a=2.0024, b=0.3866, 예측 반응량=5.096





https://www.statskingdom.com/correlation-calculator.html

3.2 회귀분석



1	2	3-Parameters	4-Predicted	5-Residuals
1.0000	2.1000	2.0024	2.3890	-0.2890
1.0000	2.5000	0.3866	2.3890	0.1110
2.0000	3.1000		2.7756	0.3244
3.0000	3.0000		3.1621	-0.1621
4.0000	3.8000		3.5487	0.2513
4.0000	3.2000		3.5487	-0.3487
5.0000	4.3000		3.9352	0.3648
6.0000	3.9000		4.3218	-0.4218
6.0000	4.4000		4.3218	0.0782
7.0000	4.8000		4.7083	0.0917

$$b = \frac{\sum_{i=1}^{n}\left[(x_i-\bar{x})(y_i-\bar{y})\right]}{\sum_{i=1}^{n}(x_i-\bar{x})^2}, \quad a = \bar{y} - b\bar{x}$$

```
x = col(1)
y = col(2)
barx = mean(x)
bary = mean(y)

sxx = total((x-barx)^2)
sxy = total((x-barx)*(y-bary))

b = sxy/sxx
a = bary-b*barx

cell(3,1) = a
cell(3,2) = b
```

3.2 회귀분석

❖ 잔차(residual)

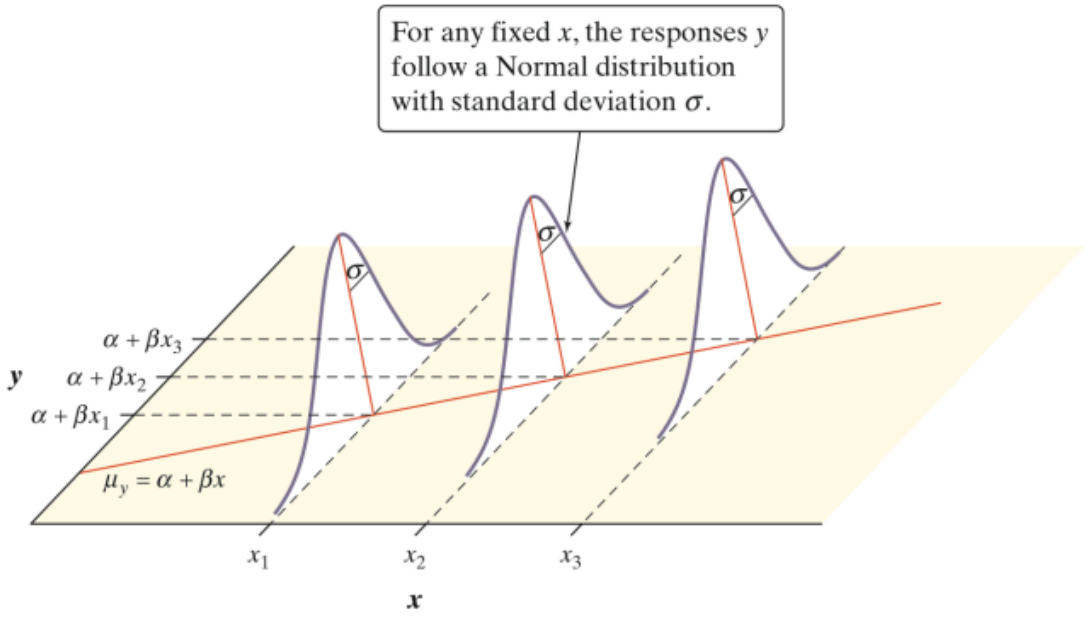
- 잔차 ε_i에 대한 다음 가정을 통해 회귀분석에 추론통계학의 개념을 도입할 수 있음:
  - ε_i는 평균이 0이고 분산이 일정한 확률변수 (실제로 평균은 0)
  - ε_i는 정규분포를 따른다: $\varepsilon_i = N(0, \sigma^2)$
- 독립변수 X의 x_i 값에서 측정된 종속변수 Y의 값 y_i가 단순 선형 회귀 모델에서 $y_i = a + bx_i + \varepsilon_i$이기 때문에, ε_i에 대한 위의 가정은 y_i가 $N(a + bx_i, \sigma^2)$의 정규분포를 따르는 모집단으로부터 얻은 한 표본임 (교과서 그림 6.6에 해당하는 다음 페이지 그림 참조)
- 잔차들의 분포는 두 변수 간에 선형관계가 있다는 조건하에서의 분포이므로(조건부 분포) 아무런 조건이 없는 종속변수의 분포(무조건부 분포) 보다 덜 퍼지게 됨 (앞 예제의 경우의 분산 비교 0.7610 vs 0.0922)
- 회귀에서 통계적 추론을 한다는 것은 잔차들의 표본으로부터 잔차의 분산을 추정하는 것임

3.2 회귀분석

- 잔차 ε_i의 평균은 0이므로 잔차의 분산은 다음과 같이 구할 수 있음:

$$s_\epsilon^2 = \frac{1}{n-2}\sum_{i=1}^{n}\epsilon_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$$

- 잔차 ε_i의 자유도가 $n-2$인 것은 두 개의 모수가 이미 추정되었기 때문임



❖ SST, SSR, SSE 간의 관계

- **SST(total sum of squares)**: 종속변수 y_i의 편차 제곱들의 합

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- **SSR(regression sum of squares)**: 예측량 $\hat{y}_i$와 종속변수의 평균 $\bar{y}$ 간 차의 제곱들의 합 → 회귀모델로 설명되는 변동성

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

- **SSE(sum of squared error)**: 잔차 제곱들의 합 → 회귀모델로 설명되지 않는 변동성

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \epsilon_i^2$$

- SST, SSR, SSE 사이에는 다음 관계가 성립함(증명):

$$SST = SSR + SSE$$

3.2 회귀분석

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 - \underbrace{\sum_{i=1}^{n} 2(\hat{y}_i - \bar{y})(y_i - \hat{y}_i)}_{0}$$

$$\hat{y}_i = a + bx_i \qquad \hat{y}_i - \bar{y} = b(x_i - \bar{x})$$

$$\bar{y} = a + b\bar{x} \qquad y_i - \hat{y}_i = (y_i - \bar{y}) - (\hat{y}_i - \bar{y})$$

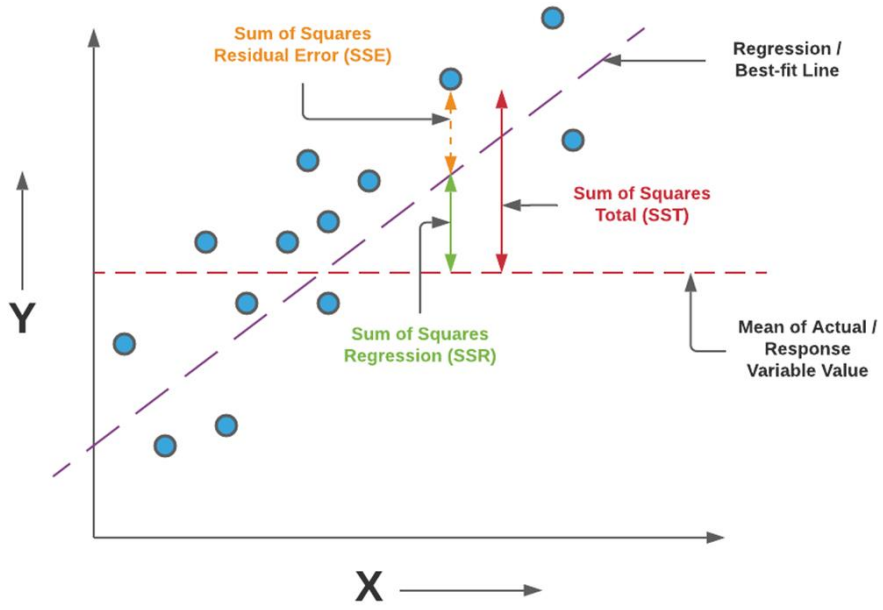
3.2 회귀분석

$$b = \frac{\sum_{i=1}^{n}\left[(x_i-\bar{x})(y_i-\bar{y})\right]}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$\begin{aligned}
\sum_{i=1}^{n} 2(\hat{y}_i-\bar{y})(y_i-\hat{y}_i) &= 2b\sum_{i=1}^{n}(x_i-\bar{x})\left[(y_i-\bar{y})-(\hat{y}_i-\bar{y})\right] \\
&= 2b\sum_{i=1}^{n}\left[(x_i-\bar{x})(y_i-\bar{y})\right]-2b\sum_{i=1}^{n}b(x_i-\bar{x})^2 \\
&= 2b\sum_{i=1}^{n}\left[(x_i-\bar{x})(y_i-\bar{y})\right]-2b\cdot b\sum_{i=1}^{n}(x_i-\bar{x})^2 \\
&= 2b\sum_{i=1}^{n}\left[(x_i-\bar{x})(y_i-\bar{y})\right]-2b\sum_{i=1}^{n}\left[(x_i-\bar{x})(y_i-\bar{y})\right] \\
&= 0
\end{aligned}$$

3.2 회귀분석

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{SST} = \underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{SSR} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{SSE}$$



❖ 회귀모형의 적합도 검정

- SAS, SPSS, R 등의 통계 패키지와 Excel, SigmaPlot 등의 스프레드시트로 회귀분석을 수행하면 아래와 같은 분산분석표(analysis of variance(ANOVA) table)를 자동으로 제공함

Source of Variation	Degrees of Freedom	Sum of Squares	Mean of Squares	F-Ratio
Regression	k	$SS_{reg} = \sum(\hat{Y}_i - \bar{Y})^2$	$MS_{reg} = SS_{reg} / k$	MS_{reg}/MS_{res}
Residual	n-k-1	$SS_{res} = \sum(Y_i - \hat{Y}_i)^2$	$MS_{res} = SS_{res} / n-k-1$	
Total	n-1	$SS_{tot} = \sum(Y_i - \bar{Y})^2$		

n ... number of observations
k ... number of independent variables

- 단순 회귀의 경우, 독립변수가 1개이므로 k=1
- 회귀모형의 적합도 측정 방법은 일반적으로 3가지: 1) MSE 또는 MSR, 2) R^2, 3) F ratio

3.2 회귀분석

- MSE(잔차의 분산): 실제 관측값에 대한 회귀모형에 의한 예측량의 평균 제곱 오차에 해당하므로 이 값이 작을수록 모형의 적합도가 높음(완전한 선형관계의 경우 0, 선형관계가 전혀 없다면 MSE=MST)

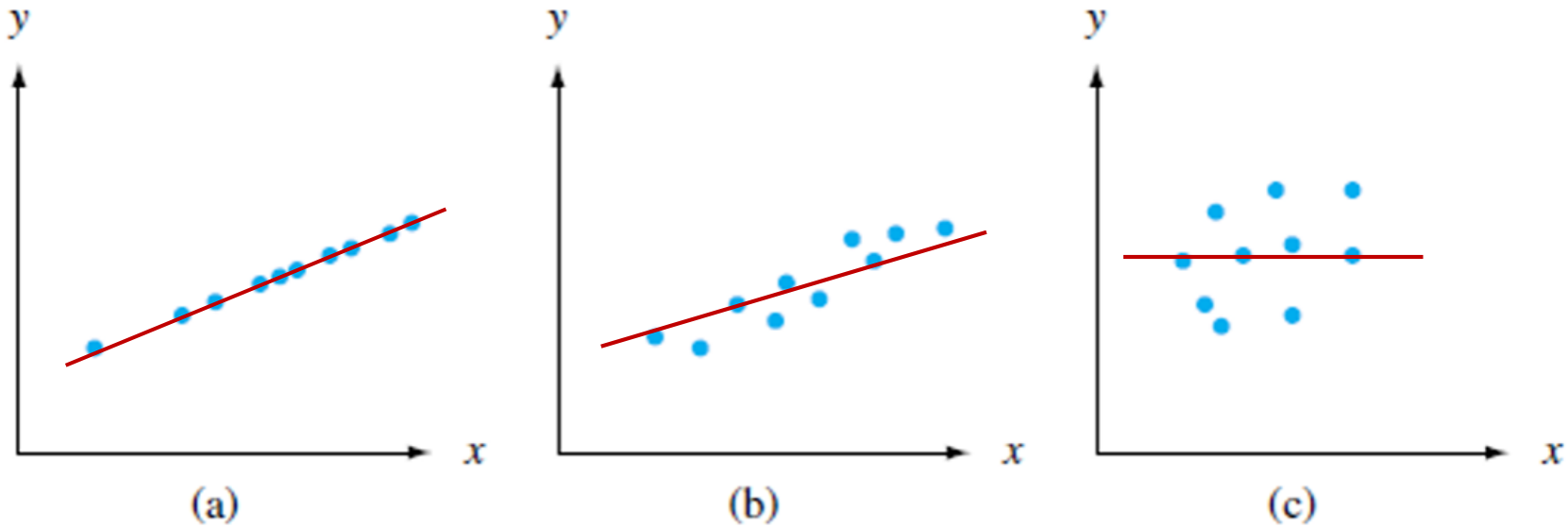
$$MSE = \frac{SSE}{n-2} = s_\epsilon^2 = \frac{1}{n-2}\sum_{i=1}^{n}\epsilon_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$$

- MSR(=SSR/1=SSR): MSE가 작으면, 즉 SSE가 작으면 SSR(즉 MSR)이 커지는데, 이것은 실제 종속변수의 변동을 회귀모형에 의한 예측량의 변동으로 많이 설명될 수 있음을 의미함. SSR이 다음과 같음을 보일 수 있으므로 회귀 직선의 기울기(b)가 클수록 SSR이 커짐을 나타냄

$$SSR = MSR = b^2\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

3.2 회귀분석

Different variability in observed *y* values:



Using the linear model to explain y variation:
(a) data for which all variation is explained;
(b) data for which most variation is explained;
(c) data for which little variation is explained

3.2 회귀분석

- R^2(결정계수): 회귀분석의 적합도로 가장 흔히 사용하는 측도로서 종속변수의 변동성에 대한 회귀모형으로 설명되는 변동성의 비를 표현하는 다음 식으로 정의됨:

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

- 결정계수가 X와 Y 간 상관계수의 제곱과 같음을 증명할 수 있음

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$= \frac{\sum_{i=1}^{n}(a + bx_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

3.2 회귀분석

- $a = \bar{y} - b\bar{x}$이므로

$$
\begin{aligned}
R^2 &= \frac{\sum_{i=1}^{n}(\bar{y} - b\bar{x} + bx_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \\
&= \frac{\sum_{i=1}^{n}\left(b(x_i - \bar{x})\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \\
&= \left(b\frac{s_x}{s_y}\right)^2 \\
&= r_{xy}^2
\end{aligned}
$$

- 결정계수의 범위는 0~1: 종속변수의 변동을 회귀모델이 완전히 다 설명할 수 있으면 1, 전혀 설명하지 못하면 0 (R^2=0.5이면 회귀모델에 의해 종속변수의 변동이 50% 설명된다고 말함)

3.2 회귀분석

- F 비$(F-ratio)$: MSE(회귀모델에 의해 설명되지 않는 변동)에 대한 MSR(회귀모델에 의해 설명되는 변동)의 비. 즉

$$F = \frac{MSR}{MSE}$$

- 두 변수의 관계가 강할수록 MSR은 크고 MSE는 작기 때문에 F 비는 커짐. F 비가 클수록 적합도가 높음

Source	Mean Square	F Statistic (df_1,df_2)
Regression (between $\hat{y}_i$ and $\bar{y}$)	6.1114	66.2838 (1,8)
Residual (between y_i and $\hat{y}_i$)	0.0922	

❖ 회귀계수의 검정

- 통계 패키지들은 회귀계수 a, b의 추정 결과와 함께 이 계수들의 표준편차를 아래의 식으로 각각 계산하여 제공함:

$$S_a = s_\epsilon \left[\sum_{i=1}^{n} x_i^2 / \left(n \sum_{i=1}^{n} (x_i - \bar{x})^2 \right) \right]$$

$$S_b = s_\epsilon \Big/ \left[\sum_{i=1}^{n} (x_i - \bar{x})^2 \right]^{1/2}$$

- 각 계수들의 t 값은 S_a, S_b에 대한 a, b의 비, 즉 아래의 식으로 계산함:

$$t_a = a/S_a, \quad t_b = b/S_b$$

- 이 t 값이 각각 H_0: $\alpha = 0$과 H_0: $\beta = 0$ 이라는 회귀계수에 대해 가설검정하는 통계량으로 사용됨

3.2 회귀분석

- 이 t 값들로부터 p-value를 구함으로써 귀무가설 H_0의 기각 여부를 판정함 (상관계수 가설검정 참조)
- t_b는 회귀직선의 적합성을 직접적으로 판단하게 해주는 통계량인데, 주어진 유의수준에서 귀무가설을 기각하지 못하면 회귀직선은 통계적으로 의미가 없음

3.2 회귀분석

Quiz(피어슨 상관계수): 15명 학생의 수학 성적과 영어 성적이 다음과 같이 나타났을 때, 수학 성적을 독립변수, 영어 성적을 종속변수로 하는 단순 선형 회귀모형을 제시하고, 결정계수를 통해 이 모형의 적합도를 판정하시오. 또한, 유의수준 5%에서 양측 검정을 통한 회귀계수의 통계적 유의성을 검정하시오.

수학 성적 {64, 75, 68, 87, 76, 70, 90, 72, 60, 70, 80, 76, 84, 81, 57}

영어 성적 {83, 80, 67, 91, 73, 92, 85, 76, 72, 81, 96, 67, 69, 87, 64}

Answer: $y = 43.7225 + 0.4749x,\ R^2 = 0.2013,\ t_a = 2.2346, t_b = 1.8101,$
$(p-value)_a = 0.0436, (p-value)_b = 0.0934$

Scatter Diagram & Linear Regression: Math vs Eng