회귀분석

보건빅데이터통계분석

이새봄
삼육대학교 SW융합교육원

상관분석

상관분석

- 문제의 정의
  - K속옷회사는 몸무게와 키와의 관계를 조사하고자 한다.
  - 몸무게와 키와는 어떤 관계가 있는가?

상관분석(Correlation)

	A	B
1	weight	hight
2	72	176
3	72	172
4	70	182
5	43	160
6	48	163
7	54	165
8	51	168
9	52	163
10	73	182
11	45	148
12	60	170
13	62	166
14	64	172
15	47	160
16	51	163
17	74	170
18	88	182

상관분석

- 문제의 정의
  - K의류에서는 새로운 옷을 디자인하려고 하는데, 키와 몸무게가 어떤 관계가 있는지를 보고자 한다.
  - (Ch1101.상관분석(CORR).sav)

- 가설
  - 귀무가설($H0$): 두 변수간에는 상관관계가 없다.

$$H_0 : \rho = 0$$

  - 연구가설($H1$:): 두 변수간에는 상관관계가 있다.

$$H_0 : \rho \neq 0$$

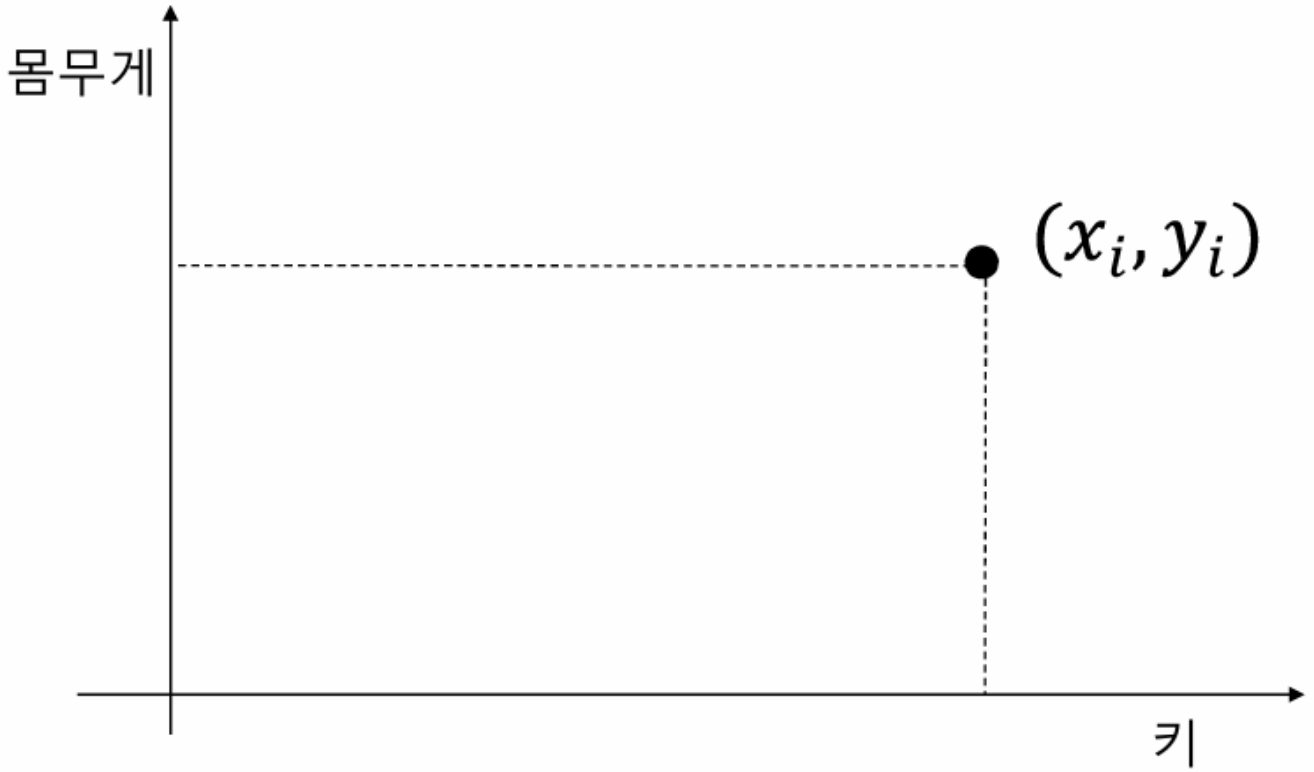
상관분석

- 두 개의 연속변수 사이의 관계성
  - 두 변수간 상호의존관계가 있을 때 이 관련성을 통계적으로 분석
  - 관련된 두 변수가 있을 때 하나의 변수에 대한 정보를 가지고 다른 변수를 예측하거나 설명할 때
  - 두 변수 사이에 강한 관련성이 있을 경우에는 한 변수에 대한 정보를 가지고 다른 변수를 예측할 수 있음

- 자료 $(x_1, y_1), (x_2, y_2), \ldots\ldots, (x_n, y_n)$
  - 예) 야구선수의 홈런수와 연봉액수
  - 라면의 선전비용과 판매량
  - 우리나라 GNP와 자동차 보유대수

상관분석

- 산점도(Scatter plot)
  - 두 변수 X,Y의 관측치 (x_i, y_i)를 좌표평면상에 점으로 나타낸 그림

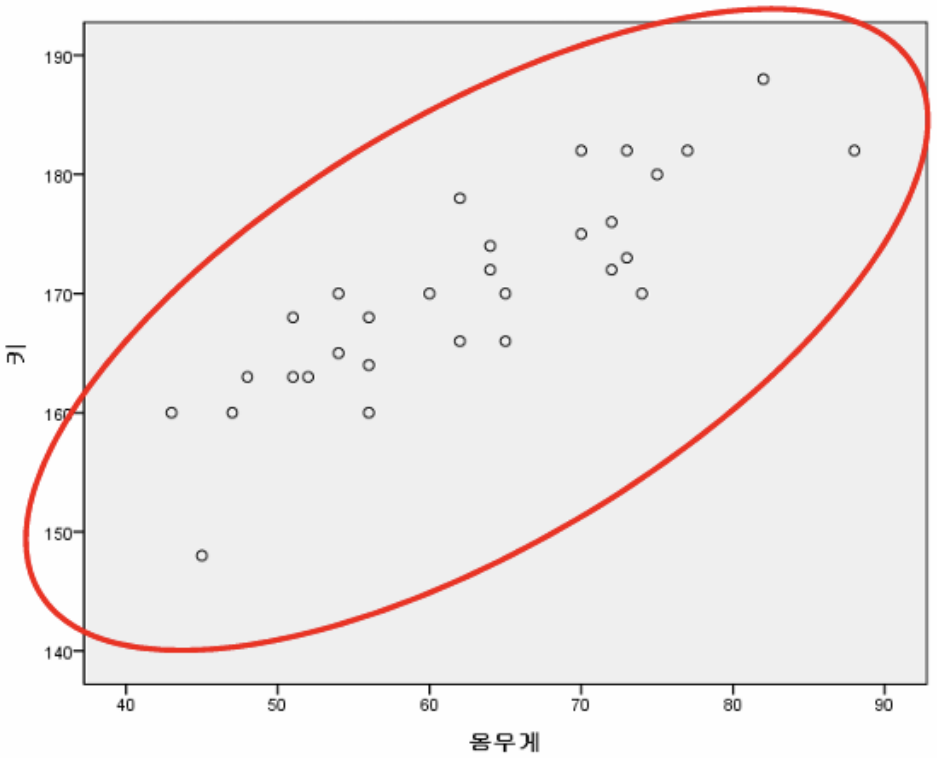


상관분석

- 산점도(Scatter plot)
  - 키와 몸무게의 관계는?
  - 상관분석

몸무게	키
72	176
72	172
70	182
43	160
48	163
54	165
51	168
52	163
73	182
45	148

$r(\text{상관계수}) = 0.857$



상관분석

- 공분산(Covariance)
  - 두 변수간의 공통분산
  - 모집단

$$\sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N}$$

  - 표본집단

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

  - 공분산은 척도단위에 따라 민감하게 반응함 → 표준화 필요

상관분석

- 상관계수(Correlation Coefficient)
  - 두 변수의 관계를 하나의 수치로 나타낸 척도
  - 모상관계수 : $\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$
  - 표본상관계수 : $r = \frac{s_{xy}}{s_x s \sigma_y}$

$$r = \frac{cov(x,y)}{\sqrt{var(x)}\sqrt{var(y)}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- 크기 $\rho(r) = \begin{cases} \text{강도: } 0 \sim 1 \\ \text{방향: } - \; or \; + \end{cases}$ $r = -1 \sim 0 \sim +1$

상관분석

- 상관계수

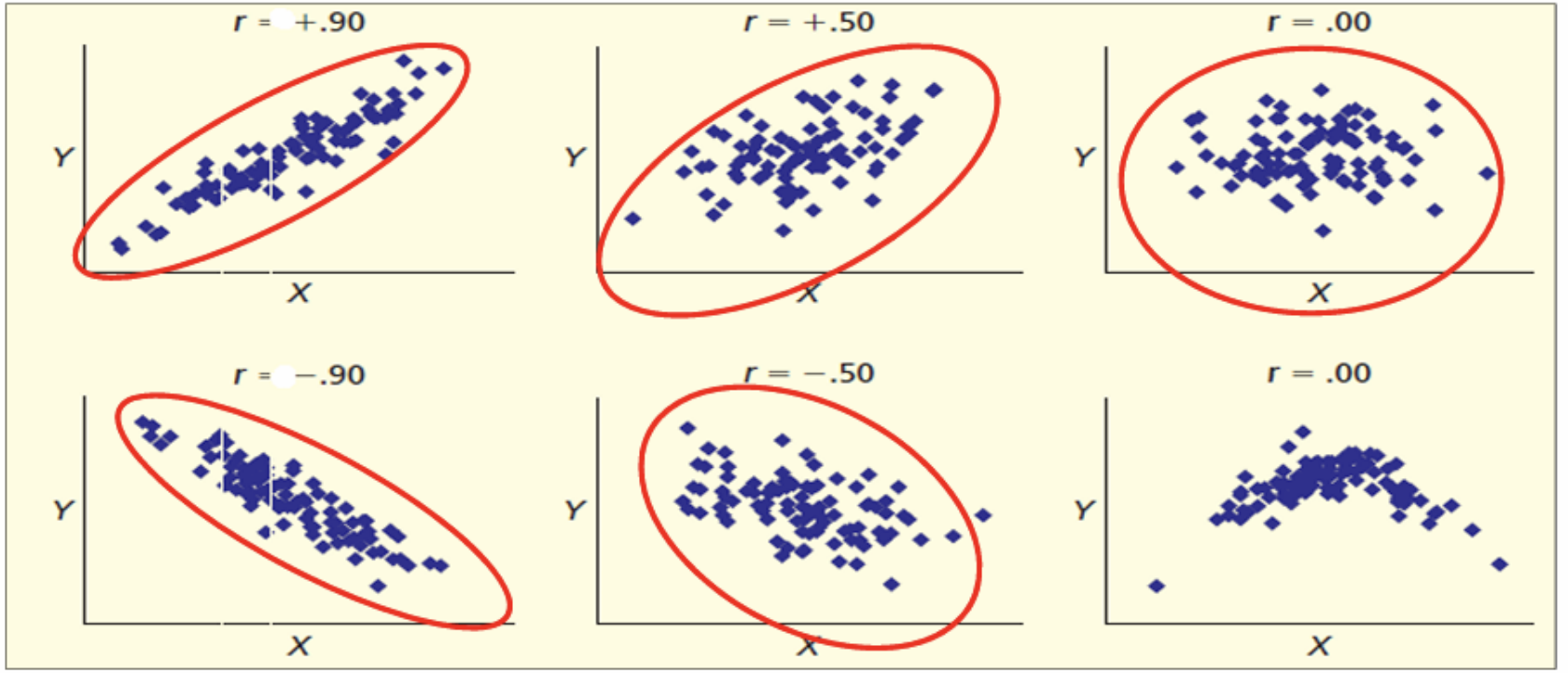
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$
$$= \frac{(72 - 62.7)(176 - 170.3) + \cdots}{\sqrt{(72 - 62.7)^2}\sqrt{(176 - 170.33)^2}}$$
$$= \frac{2{,}485}{\sqrt{3{,}848.3}\sqrt{2{,}186.7}} = 0.857$$

No	몸무게	키
1	72	176
2	72	172
3	70	182
4	43	160
5	48	163
6	54	165
7	51	168
8	52	163
…	…	…
$\bar{x}$	62.7	170

$$t = \frac{r}{\sqrt{\frac{(1 - r^2)}{n - 2}}} = \frac{r\sqrt{(n - 2)}}{\sqrt{(1 - r^2)}} = \frac{0.857\sqrt{(30 - 2)}}{\sqrt{(1 - 0.857^2)}} = 17.077$$

상관분석

- 두 개의 연속변수 사이의 관계성



상관분석

- 콜레스테롤과 중성지방 사이에는 양의 상관관계가 있는 것으로 나타났다 ($r = .86$, $p < .000$).

변수	몸무게	키
몸무게	1	
키	.86**	1

단순 선형회귀분석

회귀분석

- 회귀분석(Regression)
  - 영국의 유전학자 Francis Galton(1822~1911)에 의해 도입
  - “REGRESSION towards MEDIOCRITY in HEREDITARY STATURE”, Journal of the Anthropological Institute 15 (1886), 246–263
  - Karl Pearson(1903)에 의해 모형 정립

회귀분석

- 회귀분석(Regression)
  - 인과관계를 검정하는 분석방법
  - 하나 또는 여러 개의 원인변수(독립변수)가 다른 변수(종속변수)에 영향을 미칠 때
  - 독립변수 (Independent Variables: IV) : 종속변수에 영향을 주는 변수
  - 종속변수 (Dependent Variable: DV) : 다른 변수에 의해 영향을 받는 변수

$$X(IV) \rightarrow f(process) \rightarrow Y(DV)$$

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots$$

  - 변수들 간의 관계를 **설명**하고 **예측**하는 분석방법

회귀분석

- 회귀식을 통한 **예측**
  - 독립변수의 변화에 따라 종속변수의 값이 어떻게 변할지를 예측
  - 예) 콜레스테롤이 높으면 중성지방도 높음

$$Y(\text{중성지방}) = \beta_0 + \beta_1 x_1 = -110.819 + 1.278 \times \text{콜레스테롤}$$

- 두 변수 사이의 영향관계를 **설명**
  - 온라인게임의 몰입(즐거움)에 영향을 주는 요인

계수[a]

모형		비표준 계수		표준 계수	t	유의수준	공선성 통계	
		B	표준 오차	베타			허용 오차	VIF
1	(상수)	.029	.057		.512	.609		
	디자인	.341	.058	.352	5.860	.000	.999	1.001
	정보	.224	.057	.234	3.897	.000	1.000	1.000
	공동체	.329	.057	.344	5.741	.000	1.000	1.000
	도구	.220	.058	.227	3.788	.000	.999	1.001
	보상	.210	.057	.220	3.668	.000	1.000	1.000

a. 종속 변수: 몰입

회귀분석

- 선형회귀분석
  - 독립변수와 종속변수와의 관계가 선형(직선)일 때 이용

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \varepsilon_i$$

- 비선형회귀분석
  - 독립변수와 종속변수의 관계가 선행이 아닐 때

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2 + \cdots + \varepsilon_i$$

- 로지스틱 회귀분석
  - 종속변수의 값이 이분형(명목변수)일 때

$$Y(0,1) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \varepsilon_i$$

회귀분석

- 선형회귀분석의 종류
  - 단순 선형회귀분석
    - 독립변수가 하나인 경우

$$Y = \beta_0 + \beta_1 x_1 + \varepsilon_i$$

$Y_i = i$번째 반응치

$\beta_0 =$ 절편

$\beta_1 =$ 기울기

$X_i =$ 독립변수

$\varepsilon_i =$ 오차

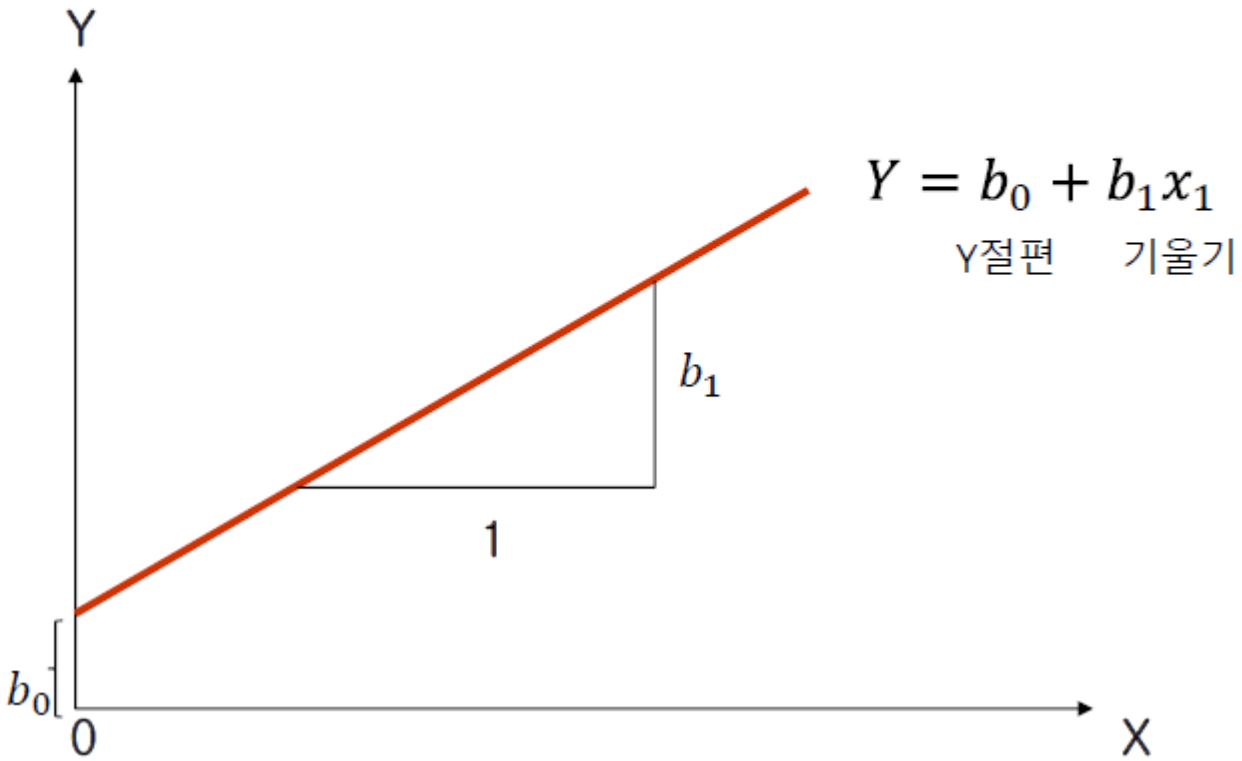
  - 다중 선형회귀분석
    - 독립변수가 여러 개인 경우

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \varepsilon_i$$

회귀분석

- 독립변수가 하나이고 두 변수가 직선관계인 회귀모형

모수(β_0, β_1) ← 추정값(b_0, b_1)

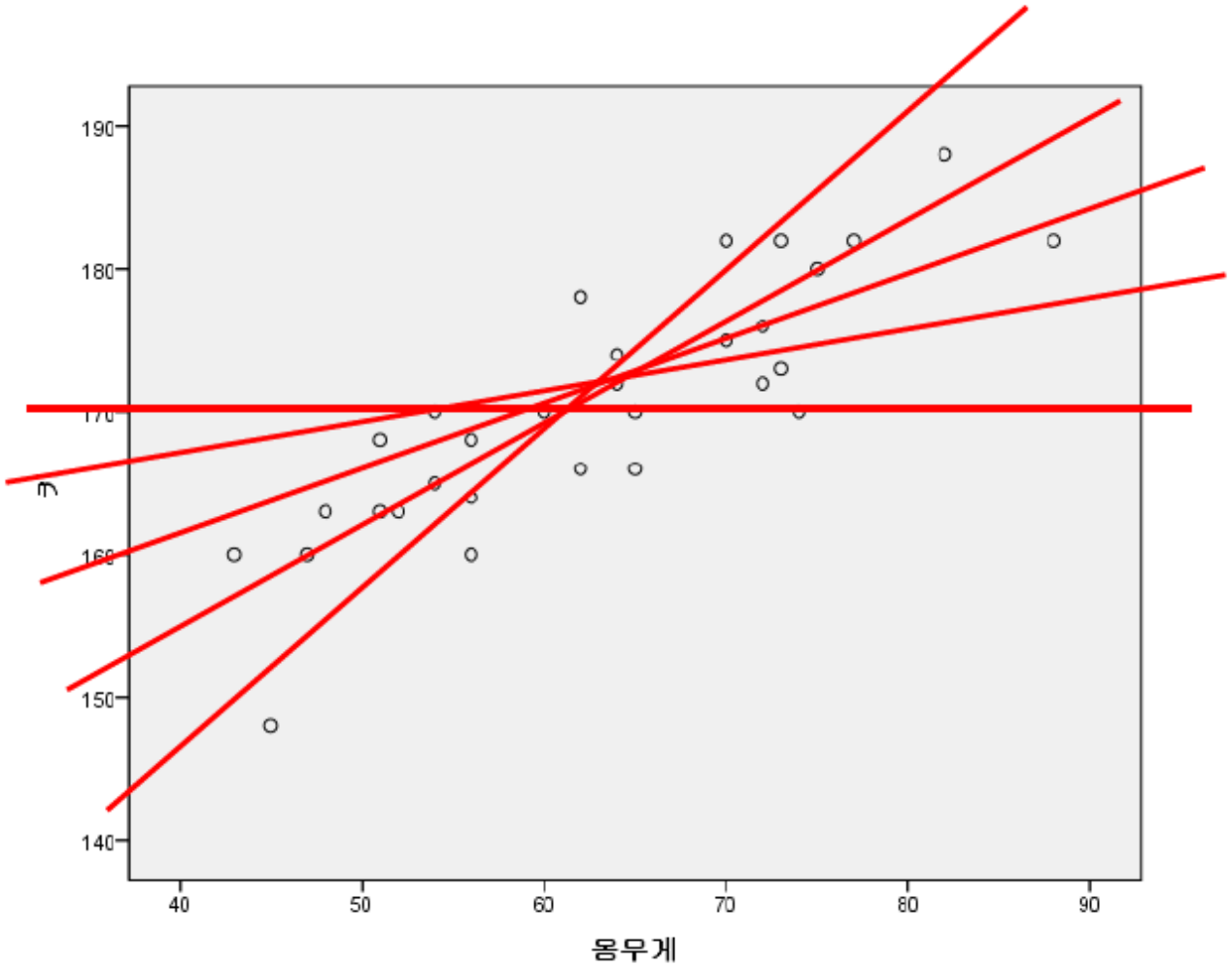


$$Y = b_0 + b_1 x_1$$

Y절편 기울기

회귀분석

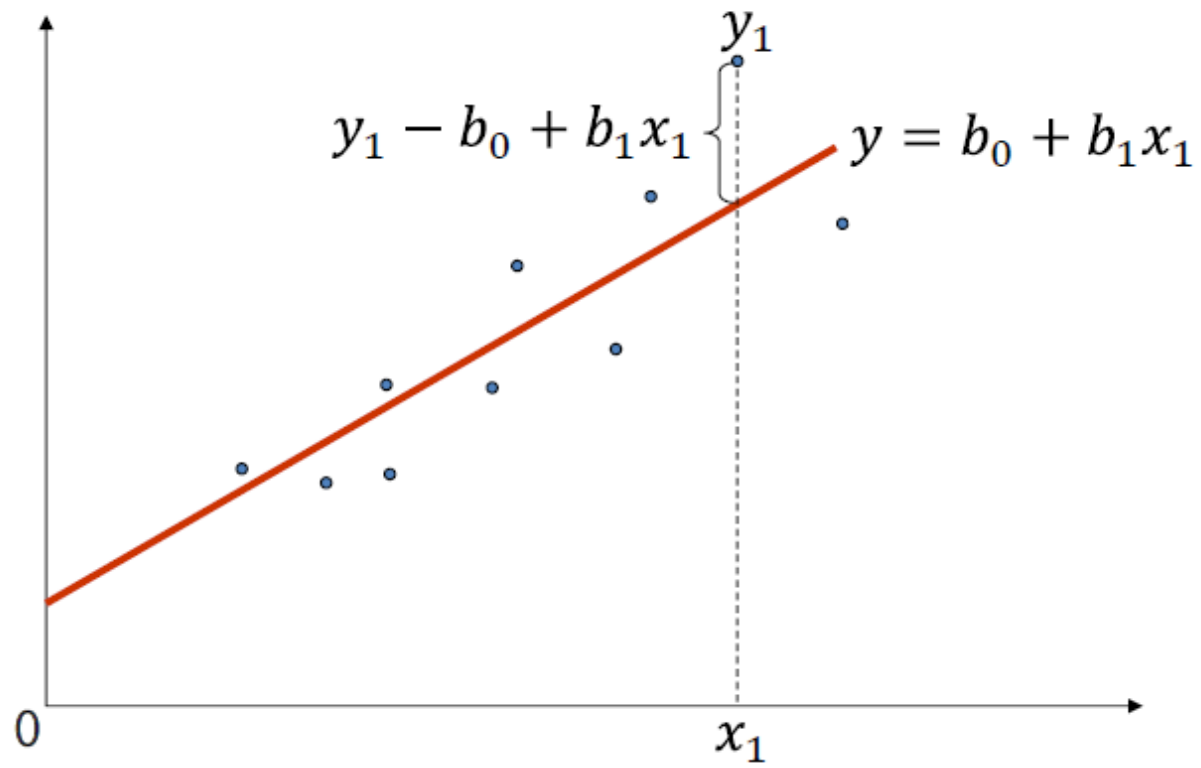
- 어떤 직선이 더 좋은가?



회귀분석

- 최소제곱법(Method of least Squares)

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 \longrightarrow \sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))^2$$



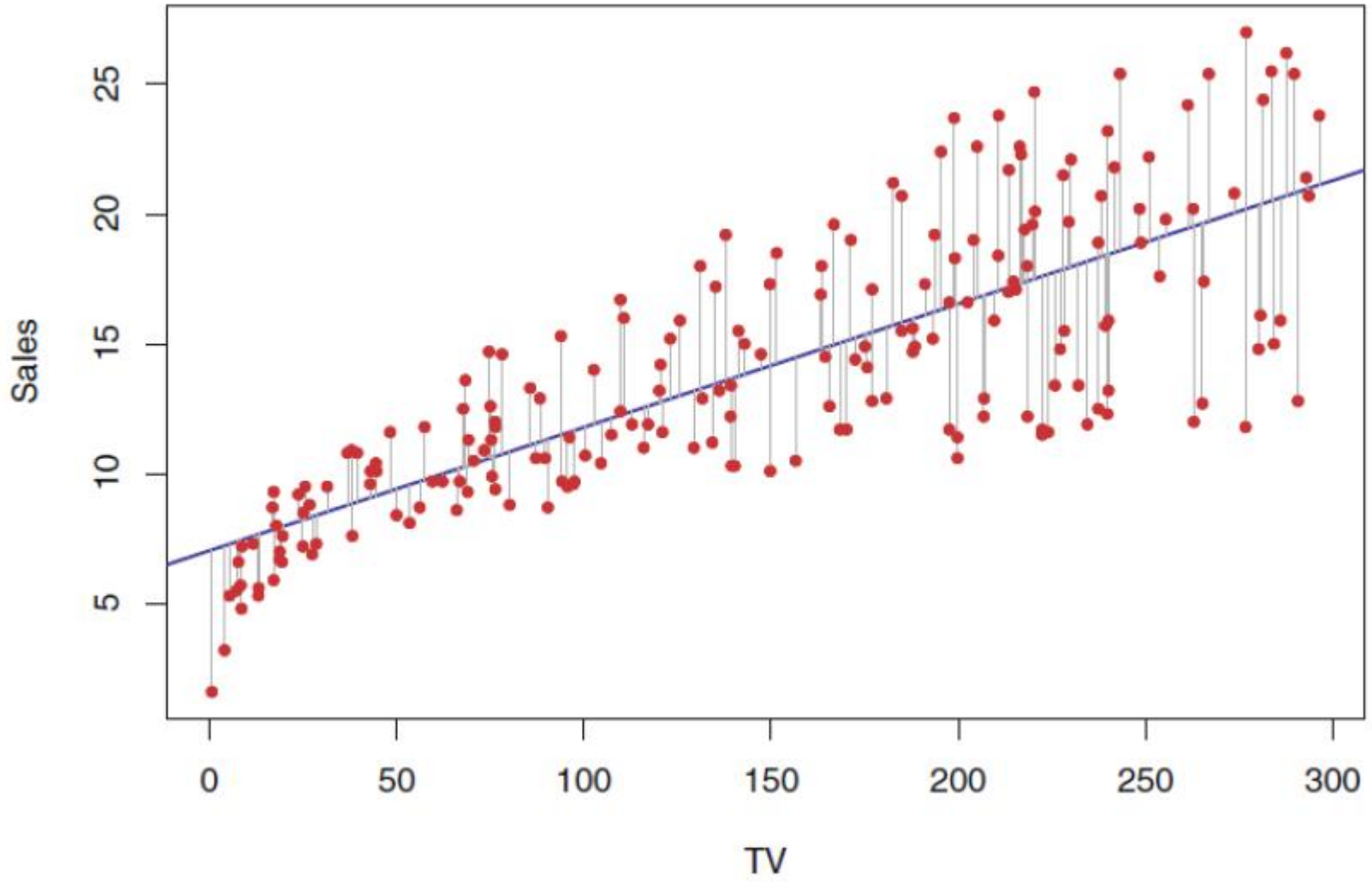
회귀분석

- 분산 모델 이용

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 \longrightarrow \sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))^2$$

- 편차(deviation)
  - 자료가 평균을 중심으로 어떻게 분포 → 분산 및 표준편차
  - 편차(개체값-평균): $(x_i - \bar{x})$
- 잔차(residuals)
  - 회귀분석에서 사용 → 모델의 적합도(회귀직선)
  - $e_i = (y_i - \hat{y}_i)$
- 오차(error)
  - 데이터마이닝 성능평가에서 사용 → 모형의 성능 (실제값 예측)
  - $e_i = (y_i - \hat{y}_i)$

회귀분석



회귀분석

- 최소제곱법(Method of least Squares)
  - 적합된 회기식에 의한 예측치 $\hat{y}_i$와 관찰치 y_i 의 차이인 잔차들의 제곱의 합이 최소가 되도록 회귀계수를 추정하는 방법(미분)

$$\sum_{i=1}^{n}(e_i)^2 = \sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))^2$$

$$\frac{\partial D}{\partial b_0} = -2\sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i)) = 0$$

$$\frac{\partial D}{\partial b_1} = -2\sum_{i=1}^{n} x_i(y_i - (b_0 + b_1 x_i)) = 0$$

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \longrightarrow \hat{\beta}_1 \qquad \beta_1 \rightarrow b_1 \rightarrow \hat{\beta}_1$$

$$b_0 = \bar{y} - b_1\bar{x} \longrightarrow \hat{\beta}_0 \qquad \beta_0 \rightarrow b_0 \rightarrow \hat{\beta}_0$$

회귀분석

- 회귀식의 분산분석

요인	제곱합 (SS)	자유도(df)	평균제곱 (MS)	F
회귀 모형	$SSR = \sum(\hat{y} - \bar{y})^2$	k	$MSR = \frac{SSR}{k}$	$\frac{MSR}{MSE}$
잔차	$SSR = \sum(y_i - \hat{y})^2$	$n - (k-1)$	$MSE = \frac{SSE}{n-(k-1)}$	
총계	$SST = \sum(y_i - \bar{y})^2$	$n - 1$	$MSR = \frac{SSR}{k}$	

- 검정통계량

$$F = \frac{MSR}{MSE} \sim F(k, n-k-1)$$

회귀분석

- 회귀식의 적합도
  - 회귀모형이 얼마나 종속변수를 잘 설명하고 있는지
  - 결정계수(R^2): 총변동 중에서 회귀모형에 의해 설명되는 비율
  - 수정 결정계수(adjusted R^2): 결정계수는 독립변수가 많아질수록 증가하기 때문에 이를 수정

$$R^2 = \frac{\text{회귀선에 의해 설명되는 변동}}{\text{전체변동}} = \frac{\sum(\hat{y} - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

회귀분석

- 회귀계수(β) 검정
  - 귀무가설(H_0) : 두 변수간에는 인과관계(영향력)가 없다.

$$H_0: \beta_1 = 0$$

  - 연구가설(H_1) : 두 변수간에는 인과관계(영향력) 가 있다.

$$H_1: \beta_1 \neq 0$$

  - 검정통계량

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MSE}{S_{xx}}}} \sim t(n-2)$$

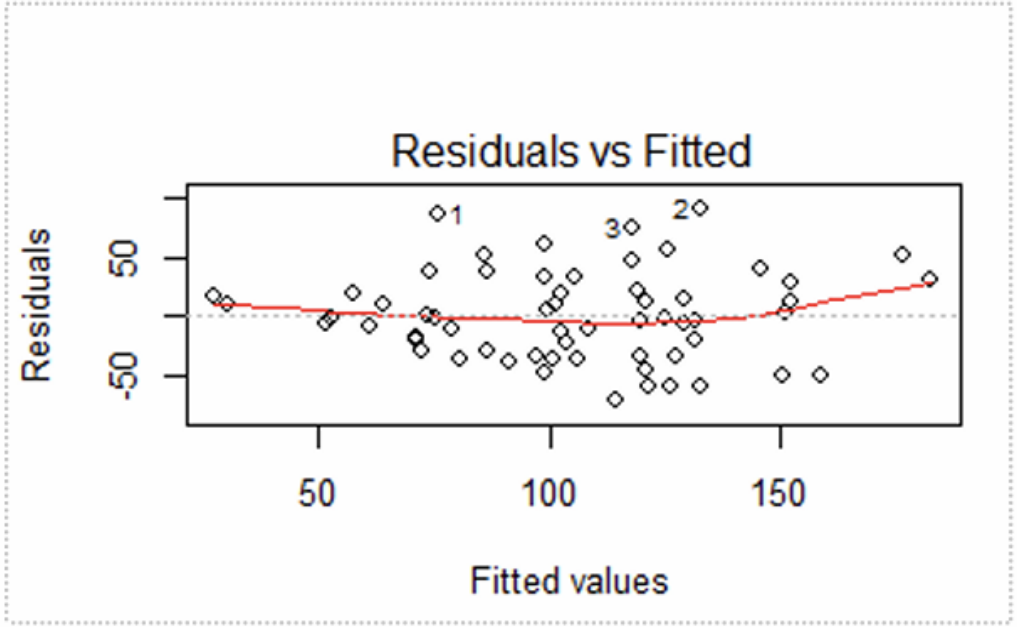
회귀분석

- 오차항의 가정
  - 회귀모형의 적절성 검정
  - 오차항 : 회귀모형의 추정치와 실제값과의 차이
  - 등분산성: 종속변수의 분산은 독립변수의 값에 관계없이 동일해야 한다.
  - 정규성: 오차는 정규분포를 이루어야 한다.
  - 독립성: 오차는 서로 독립적이어야 한다

- 잔차검정
  - 오차항을 실제로 측정 할 수 없음
  - 추정치인 잔차를 가지고 평가

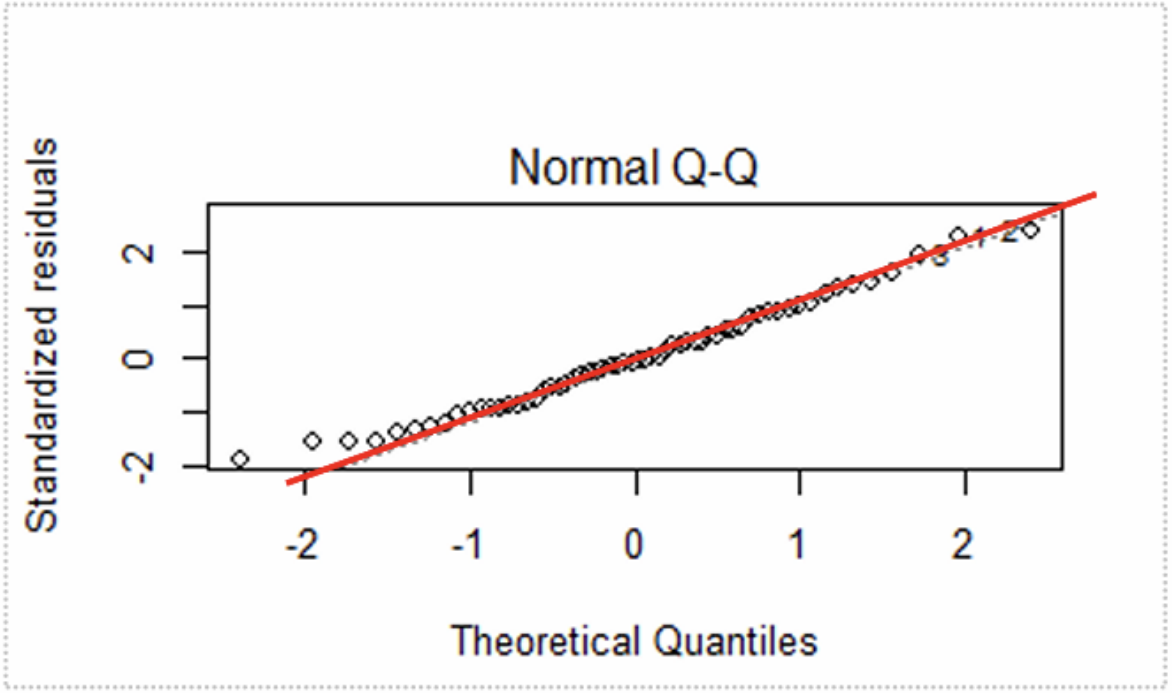
회귀분석

- 잔차의 등분산성
  - 오차 ε_i는 모든 i에 대하여 평균이 [0]이며 분산이 σ^2 인 정규분포를 따름
  - 만약 등분산성을 만족하지 않으면
  - 종속변수의 변환: $\log(y_i)$
  - 비선형 회귀분석으로 변환: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2 + \cdots + \varepsilon_i$
  - 잔차 plot으로 확인



회귀분석

- 잔차의 정규성
  - 모든 오차는 정규분포를 이루어야 한다.
  - Normal QQ pot 사용



회귀분석

- 잔차의 독립성
  - 회귀분석에서 오차 ε_i는 서로 독립적이라고 가정
  - 오차의 자기상관이란 오차들이 서로 자기상관관계가 있음을 나타냄
  - 검정방법: Durbin-Watson의 통계량
  - 주로 시계열 자료일 경우에 활용

회귀분석

- 문제의 정의
  - 일개 기업체에서 근무하고 있는 직원(100명)들의 정규적인 건강검진 결과의 일부 자료이다. 콜레스테롤이 높으면 중성지방도 높다고 말할 수 있는가? 그렇다면 콜레스테롤과 중성지방 사이의 관련성을 회귀식으로 추정하시오
  - (Ch1102.회귀분석(REG).sav)

- 가설
  - 귀무가설(H_0) : 두 변수간에는 인과관계(영향력)가 없다.

$$H_0: \beta_1 = 0$$

  - 연구가설(H_1) : 두 변수간에는 인과관계(영향력) 가 있다.

$$H_1: \beta_1 \neq 0$$

회귀분석 절차


1.데이터 불러오기
2.회귀분석
2.1 회귀식의 적합도
< 0.05
ANOVA 검정
≥ 0.2
$Adj\ R^2$
2.2 β계수 검정
< 0.05
$t-test$
3.회귀모형 가정 검정
3.1 잔차의 등분산성
잔차 plot
3.2 잔차의 정규성
≥ 0.05
Shapiro test
3.3 잔차의 독립성
≥ 1.69
Durbin-Watson
3.4 이상치 검정
< 3
cook's distance
4.결과보고서 작성

회귀분석

- 콜레스테롤과 중성지방간의 관계를 검증한 결과, 콜레스테롤은 중성지방과 관계가 있는 것으로 나타났다(F=43.5, p=0.000).
- 콜레스테롤을 통한 중성지방 예측 모델을 구하면 다음과 같다.

독립변수	B	S.E.	β	t	p	Adj R^2	F
상수	-110.819	33.178		-3.340	0.001	0.419	43.500**
콜레스테롤	1.278	0.194	0.655	6.595	0.000		

$$\text{중성지방} = -110.819 + 1.278\text{콜레스테롤}$$

Q&A