9 상관분석과 회귀분석

Topics:

- 9.1 서론
- 9.2 상관분석
- 9.3 단순선형회귀모형
- 9.4 단순선형회귀모형에서의 추론
- 9.5 잔차의 검토

9.1 서론

Topics:

- 산점도(2.3장)
- 상관분석과 회귀분석

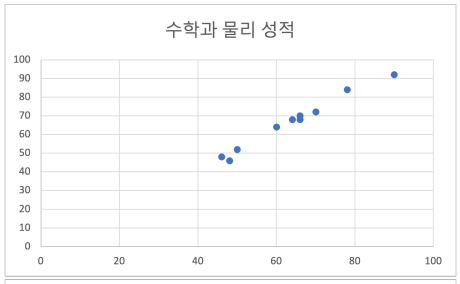
상관분석(correlation analysis)과 회귀분석(regression analysis):

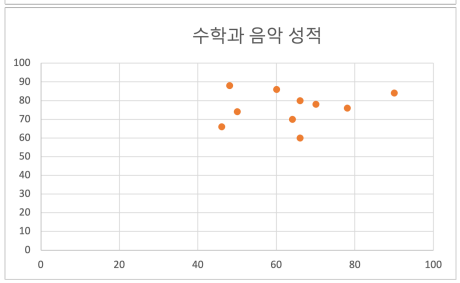
상관분석: 두 변수 사이의 _____가 있고 없음에 대한 추론
회귀분석: 두 변수 사의의 관계를 _____로 나타내어 분석하는 추론

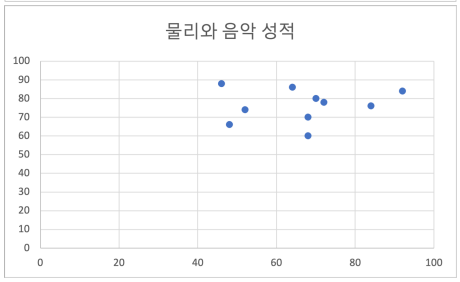
- 예: 다음은 어느 고등학교에서 랜덤하게 추출된 학생 10명의 수학, 물리, 음악성적이다.

	A	B	C	D	E	F	G	H	I	J	K
1	학생번호	1	2	3	4	5	6	7	8	9	10
2	수학	66	64	48	46	78	60	90	50	66	70
3	물리	70	68	46	48	84	64	92	52	68	72
4	음악	80	70	88	66	76	86	84	74	60	78

이 자료의 산점도를 나타내면 다음과 같다.







9.2 상관분석

Topics:

- 모상관계수와 표본상관계수
- 상관계수의 검정

표본상관계수(sample correlation coefficient):

n개의 관측 자료 $(x_1, y_1), \ldots, (x_n, y_n)$을 이용하여 구한 모집단상관계수의 추정치

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

$$= \frac{n\left(\sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\left(\sum_{i=1}^{n} y_i^2\right) - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

- 예: 표 9.1의 자료에서 수학과 물리성적, 수학과 음악성적 사이의 표본상관계수를 구하여라.

	A	B	C	D	E	F	G	H	I	J	K
1	학생번호	수학(x)	물리(y)	음악(z)	x^2	y^2	z^2	xy	xz		
2	1	66	70	80							
3	2	64	68	70							
4	3	48	46	88							
5	4	46	48	66							
6	5	78	84	76							
7	6	60	64	86							
8	7	90	92	84							
9	8	50	52	74							
10	9	66	68	60							
11	10	70	72	78							
12	합	638	664	762							
13											

상관계수의 검정:

(X, Y)가 _____를 따를 때, 표본상관계수 r에 대하여 다음이 성립한다.
귀무가설 $H_0 : \rho = 0$
검정통계량 r

기각역 (1) $H_1 : \rho > 0$ 일 때, $r \geq r_\alpha(n-2)$
(2) $H_1 : \rho < 0$ 일 때, $r \leq -r_\alpha(n-2)$
(3) $H_1 : \rho \neq 0$ 일 때, $|r| \geq r_{\alpha/2}(n-2)$

- 표11. 표본상관계수의 표본분포표 참고.

- 예: 자동차의 배기가스로 인한 공기 중의 발암성 물질에 대한 조사를 위하여, 한 도시의 12지역에서 표본을 택하고 공기 중의 일산화탄소의 농도 x(ppm)와 벤조피렌의 농도 y(ug/10^3m^3)를 측정한 결과 다음의 데이터를 얻었다.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	지역	1	2	3	4	5	6	7	8	9	10	11	12	합
2	x	5.5	5.5	5.5	5.6	5.6	6.8	9.6	10.5	11	12	12.8	13.3	103.7
3	y	1	1.3	2.2	1.1	1.5	1.9	3.9	5.5	7.3	5.7	8.1	7.8	47.3
4	x^2													
5	y^2													
6	xy													

일산화탄소와 벤조피렌의 농도 사이의 상관계수 ρ에 대하여, $H_0 : \rho = 0$, $H_1 : \rho > 0$을 유의수준 1%에서 검정하여라.

Ans.

1. 귀무가설:
   대립가설:
2. 유의수준:
3. 검정통계량:

4. 기각역:
5. 검정통계량의 관측값:

결과:즉, 일산화탄소와 벤조피렌의 농도 사이의 양의 상관계에 있다고 할 수 (있다. / 없다.)

9.3 단순선형회귀모형

Topics:

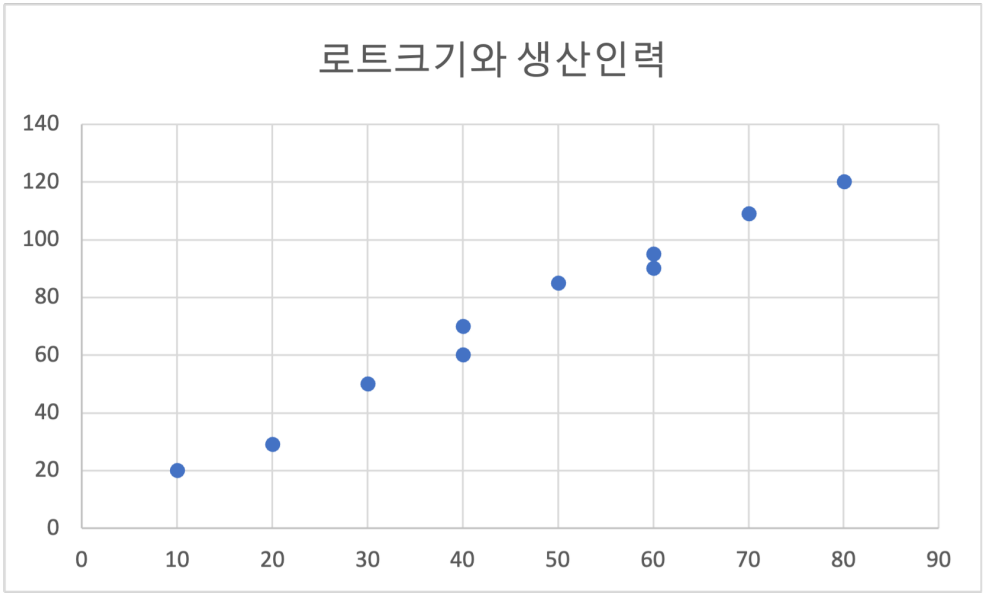
- 단순선형회귀
- 최소제곱추정량과 추정회귀직선
- 오차제곱합과 평균제곱오차(오차항의 분산에 대한 추정량)

회귀분석의 분류:

단순회귀분석(simple regression): 독립변수가 한 개 일때의 회귀분석
다중회귀분석(multiple regression): 독립변수가 두 개 이상 일때의 회귀분석
단순선형회귀분석(simple linear regression): 하나의 독립변수가 종속변수에 _____인 경우

- 예: 다음 자료는 로트크기에 따른 생산인력을 예측하고자, 한 공장에서 랜덤하게 추출된 자료이다.
(로트크기(lot size): 생산이 이루어지는 단위 수량)

	A	B	C	D	E	F	G	H	I	J	K
1	로트크기	10	20	30	40	40	50	60	60	70	80
2	생산인력	20	29	50	60	70	85	90	95	109	120



최소제곱추정량과 추정회귀직선:

$$\hat{\beta} = \frac{n\left(\sum_{n=1}^{n} x_i y_i\right) - \left(\sum_{n=1}^{n} x_i\right)\left(\sum_{n=1}^{n} y_i\right)}{n\left(\sum_{n=1}^{n} x_i^2\right) - \left(\sum_{n=1}^{n} x_i\right)^2}$$

$$\hat{\alpha} = \overline{y} - \hat{\beta}x$$

$$\hat{y} = \hat{\alpha} + \hat{\beta}x =$$

- 예: 생산인력의 자료에 대하여 α와 β의 최소제곱추정량을 구하고, 산점도에 추정회귀직선을 그려보아라.

	A	B	C	D	E	F	G	H	I	J	K	L
1	로트크기	10	20	30	40	40	50	60	60	70	80	
2	생산인력	20	29	50	60	70	85	90	95	109	120	
3												
4	x^2											
5	y^2											
6	xy											

오차제곱합과 평균제곱오차(오차항의 분산에 대한 추정량):

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

$$= \left(\sum_{n=1}^{n} y_i^2\right) - \frac{\left(\sum_{n=1}^{n} y_i\right)^2}{n} - \frac{\left(\left(\sum_{n=1}^{n} x_i y_i\right) - \frac{\left(\sum_{n=1}^{n} x_i\right)\left(\sum_{n=1}^{n} y_i\right)}{n}\right)^2}{\left(\sum_{n=1}^{n} x_i^2\right) - \frac{\left(\sum_{n=1}^{n} x_i\right)^2}{n}}$$

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-2}$$

- 예: 생산인력의 자료에 대하여 단순선형회구모형을 적용할 때, 오차항의 분산 σ^2의 추정값을 구하여라.