

선형관계의 분석

변수명	코딩명	변수 설명	척도
성별 구분	Gender	남자=0, 여자=1	명목척도
연령	Age	단위: 세, 만 나이	비율척도
신장	Hight	단위: Cm	비율척도
체중	Weight	단위: kg	비율척도
체지방량	BFM	단위: kg	비율척도
제지방량	FFM	단위: kg, 체중 - 체지방량	비율척도
근육량	SLM	단위: kg	비율척도
뼈가 된 무기질	MIN	단위: kg, 제지방량 - 근육량,	비율척도
골격근량	SMM	단위: kg	비율척도
내장지방	VFA	단위: cm ²	비율척도
복부지방율	WHR	단위: cm/cm	비율척도
요추골밀도	LBMD	단위: g/cm ² , 요추(L2~4) T-값	비율척도
대퇴골밀도	FBMD	단위: g/cm ² , 대퇴골경부 T-값	비율척도
√ 연령 구분	G_Age	-39, 40-44, 45-49, 50-54, 55-60, 60+	서열척도
체질량지수	BMI	단위: kg/m ² , 체중 ÷ 신장 ²	비율척도
√ 비만 판정	D-OBE	저체중=1, 정상=2, 과체중=3, 비만=4	서열척도
체지방지수	I_BFM	단위: kg/cm ² , 체지방량 ÷ 신장 ²	비율척도
체지방율	R_BFM	단위: kg/kg, 체지방량 ÷ 체중 ²	비율척도
√ 요추골밀도 판정	D-LBMD	정상=1, 골감소증=2, 골다공증=3	서열척도
√ 대퇴골밀도 판정	D-FBMD	정상=1, 골감소증=2, 골다공증=3	서열척도
요추골밀도 이상 여부	N/A-LBMD	정상=0, 비정상=1	명목척도
대퇴골밀도 이상 여부	N/A-FBMD	정상=0, 비정상=1	명목척도
요추 골다공증 여부	N/A-L	정상=0, 골다공증=1	명목척도
대퇴 골다공증 여부	N/A-F	정상=0, 골다공증=1	명목척도

연구문제

- ◆ 키가 큰 사람은 연인으로 키가 큰 사람을 선호?
- ◆ 키가 큰 사람은 연인으로 키가 작은 사람을 선호?

연구 문제 10-1

6가구의 음주량(단위: 소주 병수/일주일) 가족들 사이의 관계는?

Sample ID	Father	Mother	Son	Daughter
1	0	4	0	0
2	1	3	2	2
3	2	3	2	4
4	3	1	3	5
5	4	1	3	1
6	5	0	5	0

1. 아버지와 어머니의 주량의 관계는?
2. 아버지와 아들의 주량의 관계는?
3. 아버지와 딸의 주량의 관계는 어떠한가? 또 어머니와 딸의 주량의 관계는?
4. 딸의 주량은 아버지와 어머니 중 누구와 더 관련이 있는가?

연구문제

연구 문제 10-2

체성분과 골밀도 변수간의 관련성

- ◆ 연령구분과 비만판정 등은 서열척도임에 주의
- ◆ 서열척도는 선형관계를 어떻게 분석해야 하나?

연구 문제 10-3

시험성적(S)과 몸무게(W)

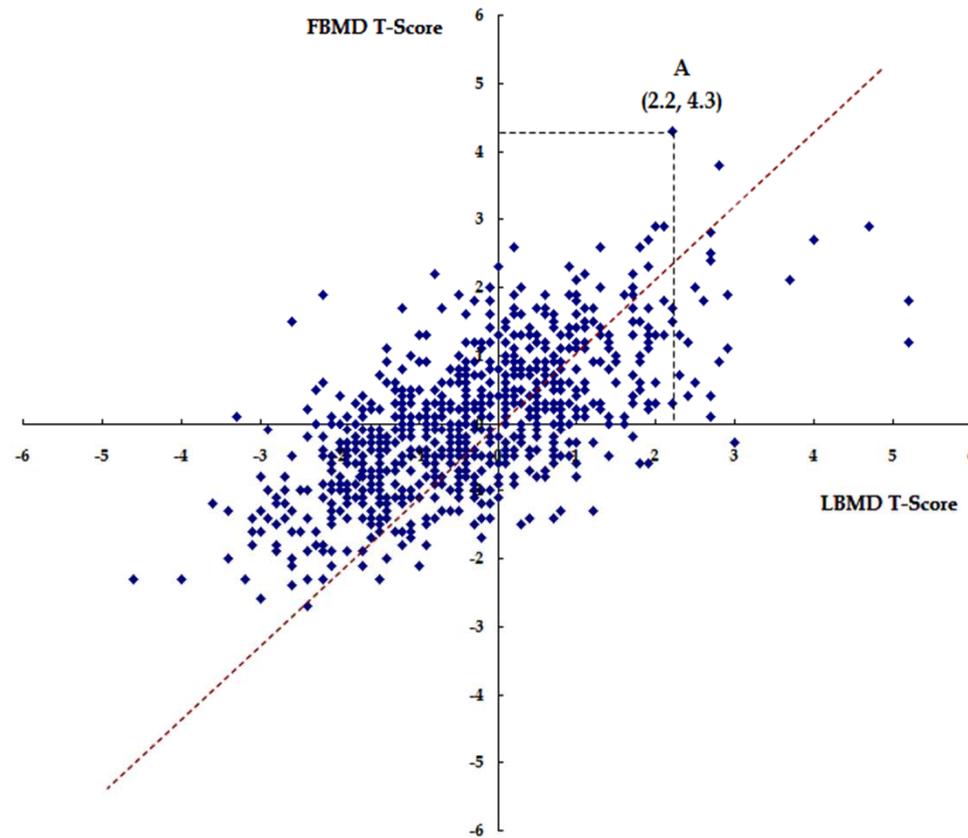
- ◆ 초등학교 학생 21명을 대상으로 ‘학습능력’에 영향을 미치는 요인을 조사하였다. 학습능력은 시험의 성적으로 측정되었다. 여러 가지 조사변수들 중 ‘몸무게(weight, W)’와 ‘시험성적(score, S)’ 사이의 관계는 다음과 같았다. (Excel을 이용하여 Scatter plot을 작성하시오)

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
W	21	20	19	22	23	24	27	26	25	28	29	30	33	32	31	34	35	36	39	38	37
S	40	42	44	50	52	54	60	62	64	70	72	74	80	82	84	88	90	92	94	96	98

1. ‘몸무게(Weight)’와 ‘시험성적(Score)’ 사이의 관계는 어떠한가?
2. 두 변수의 관계에 대해 직관적으로 동의할 수 있는가? 자료가 잘못된 것이 아니라면, 무엇이 문제인가?

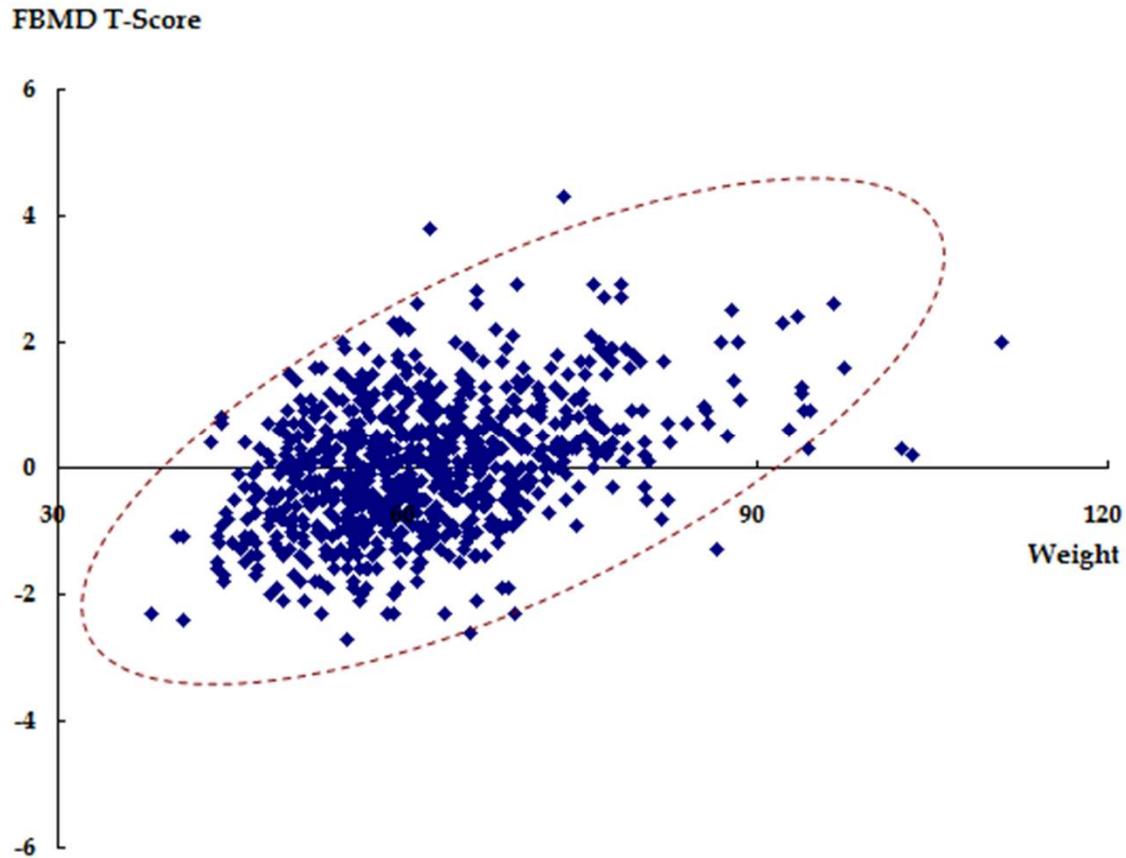
두 수치형 변수간 선형관계의 이해

- ◆ 요추골밀도(LBMD T-Socre)와 대퇴골밀도(FBMD T-Score)의 관계



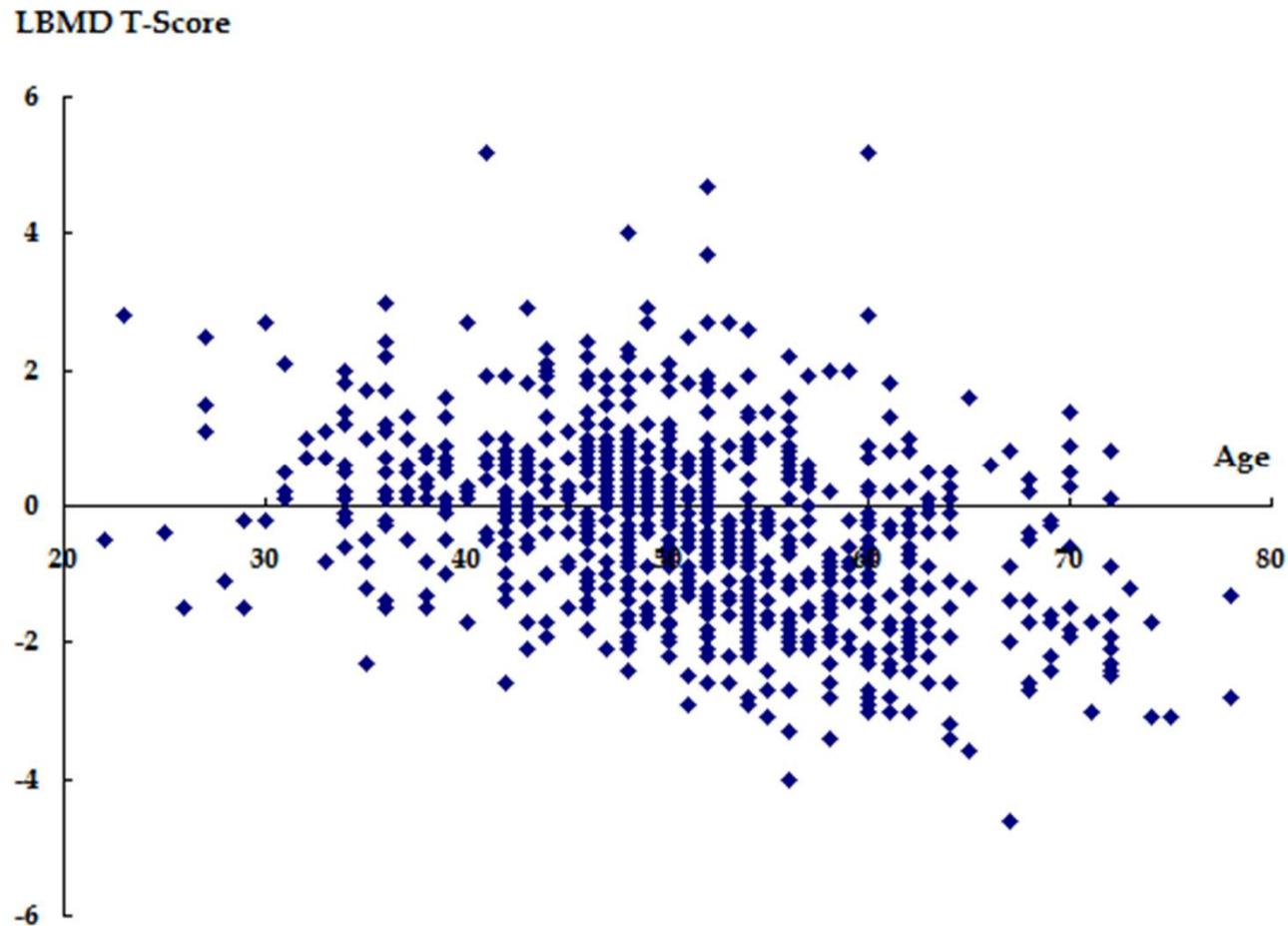
두 수치형 변수간 선형관계의 이해

◆ 체중과 대퇴골밀도의 관계



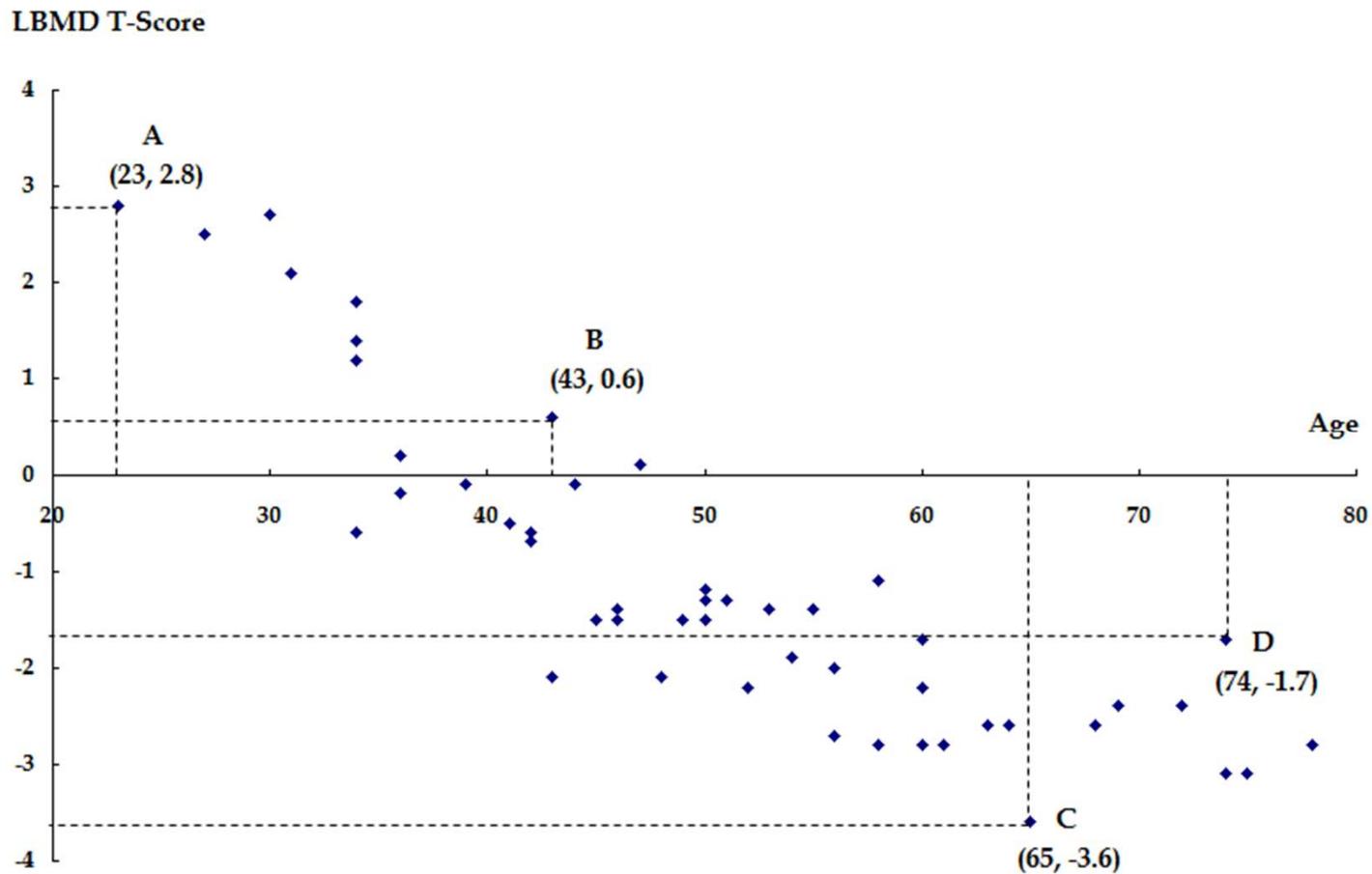
두 수치형 변수간 선형관계의 이해

◆ 연령과 요추골밀도의 관계



두 수치형 변수간 선형관계의 이해

◆ 50개 표본: 연령과 요추골밀도의 관계



선형관계의 측정: 공분산과 상관계수

◆ 공분산의 이해(음주량 사례)

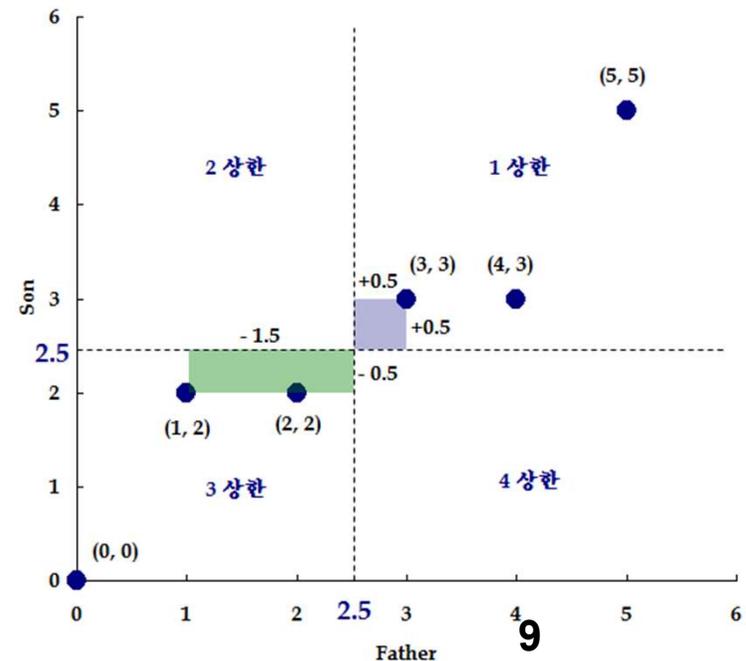
- $(x_1, y_1) (x_2, y_2) (x_3, y_3) (x_4, y_4) (x_5, y_5) (x_6, y_6) =$
- $(0, 0) (1, 2) (2, 2) (3, 3) (4, 3) (5, 5)$

◆ 공분산
$$S_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}, \quad (\bar{x} = \frac{\sum x_i}{n}, \quad \bar{y} = \frac{\sum y_i}{n})$$

Father	Son			
x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
0	0	-2.5	-2.5	6.25
1	2	-1.5	-0.5	0.75
2	2	-0.5	-0.5	0.25
3	3	0.5	0.5	0.25
4	3	1.5	0.5	0.75
5	5	2.5	2.5	6.25
15	15			14.50

$\bar{x} = 2.5 \quad \bar{y} = 2.5$

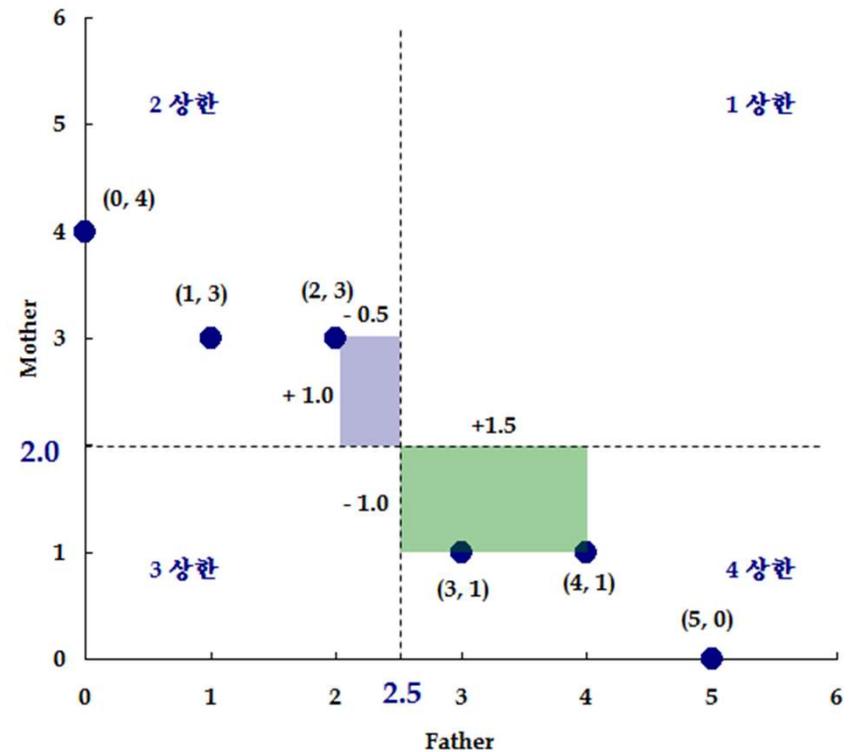
$s_{xy} = 14.5/5 = 2.9$



선형관계의 측정: 공분산과 상관계수

◆ 아버지와 엄마 주량의 공분산

Father		Mother		
x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
0	4	-2.5	2.0	-5.0
1	3	-1.5	1.0	-1.5
2	3	-0.5	1.0	-0.5
3	1	0.5	-1.0	-0.5
4	1	1.5	-1.0	-1.5
5	0	2.5	-2.0	-5.0
15	12			-14.0
$\bar{x} = 2.5$		$\bar{y} = 2.0$		$s_{xy} = -14/5 = -2.8$



선형관계의 측정: 공분산과 상관계수

◆ 아버지와 독립인 누군가와와의 공분산

Father	Alcoholic			
x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
0	20	-2.5	0.0	0.00
1	20	-1.5	0.0	0.00
2	20	-0.5	0.0	0.00
3	20	0.5	0.0	0.00
4	20	1.5	0.0	0.00
5	20	2.5	0.0	0.00
15	120			0.00

$$\bar{x} = 2.5 \quad \bar{y} = 20.0$$

$$s_{xy} = 0/5 = 0$$

공분산

모공분산(population covariance):
$$\sigma_{xy} = \frac{\sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y)}{N}, \quad (\mu_x = \frac{\sum x_i}{N}, \quad \mu_y = \frac{\sum y_i}{N})$$

표본공분산(sample covariance):
$$S_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{N}, \quad (\bar{x} = \frac{\sum x_i}{n}, \quad \bar{y} = \frac{\sum y_i}{n})$$

선형관계의 측정: 공분산과 상관계수

◆ 공분산을 통해 알 수 있는 사실

- 공분산의 부호: 선형관계의 방향성
- 공분산의 크기: 선형관계의 강도 (관계의 강도를 객관적으로 비교하기는 불가능 → 상관계수로 보완)

◆ 상관계수 (correlation coefficient)

상관계수	
모상관계수(population correlation coefficient):	$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}, \quad -1 \leq \rho_{xy} \leq +1$
표본상관계수(sample correlation coefficient):	$r_{xy} = \frac{S_{xy}}{S_x S_y}, \quad -1 \leq r_{xy} \leq +1$

선형관계의 측정: 공분산과 상관계수

C2		fx =CORREL(B6:B11,C6:C11)				
	A	B	C	D	E	F
1						
2			0.94337			
3						
4		Father	Son			
5		x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
6		0	0	-2.5	-2.5	6.25
7		1	2	-1.5	-0.5	0.75
8		2	2	-0.5	-0.5	0.25
9		3	3	0.5	0.5	0.25
10		4	3	1.5	0.5	0.75
11		5	5	2.5	2.5	6.25
12		15	15			14.50
13						
14		$\bar{x} = 2.5$	$\bar{y} = 2.5$			
15		$s_x = 1.87$	$s_y = 1.64$			
16						

C		=CORREL(B6:B11,C6:C11)			
	C	D	E	F	
1					
2		0.94337			
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					

Son					
	x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
8	3	2	-0.5	-0.5	0.25
9	4	3	0.5	0.5	0.25
10	5	3	1.5	0.5	0.75
11	6	5	2.5	2.5	6.25
12	21	15			14.50
13					
14	$\bar{x} = 3.5$	$\bar{y} = 2.5$			
15	$s_x = 1.87$	$s_y = 1.64$			

Son					
	x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
8	3	2	-0.5	-0.5	0.25
9	4	3	0.5	0.5	0.25
10	5	3	1.5	0.5	0.75
11	6	5	2.5	2.5	6.25
12	21	15			14.50
13					
14	$\bar{x} = 3.5$	$\bar{y} = 2.5$			
15	$s_x = 1.87$	$s_y = 1.64$			

x=>x+1 (아버지 주량이 1병 증가)

$\bar{x} = 2.5$ $\bar{y} = 2.5$
 $s_x = 1.87$ $s_y = 1.64$

$s_{xy} = 14.5/5 = 2.9$
 $r_{xy} = s_{xy} / (s_x s_y) = 0.943$

$\bar{x} = 3.5$ $\bar{y} = 2.5$
 $s_x = 1.87$ $s_y = 1.64$

$s_{xy} = 14.5/5 = 2.9$
 $r_{xy} = s_{xy} / (s_x s_y) = 0.943$

선형관계의 기울기와 상관계수

C2 $\text{=CORREL}(B5:B9,C5:C9)$

	A	B	C	D	E
1					
2			1		
3					
		X	Y		
5		1	2		
6		2	4		
7		3	6		
8		4	8		
9		5	10		

C2 $\text{=CORREL}(B5:B9,C5:C9)$

	A	B	C	D	E
1					
2			1		
3					
		X	Z		
5		1	3		
6		2	6		
7		3	9		
8		4	12		
9		5	15		

기울기는 상관
계수와 무관

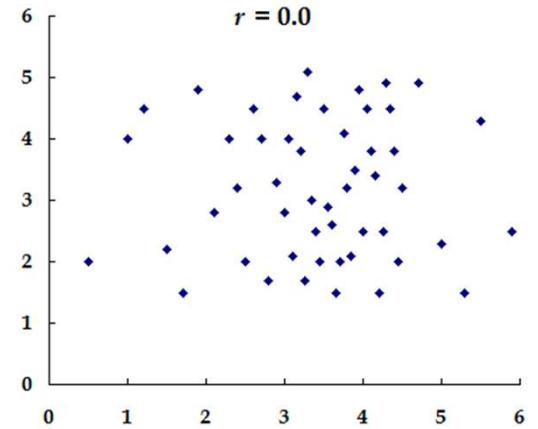
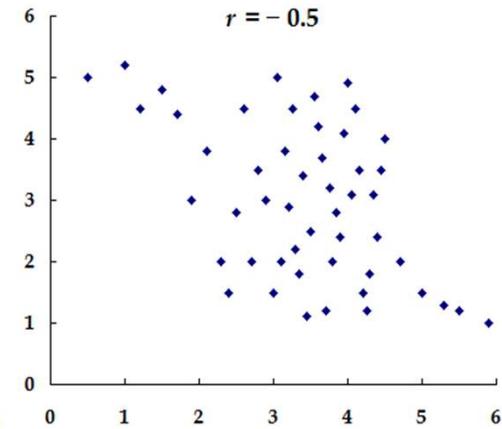
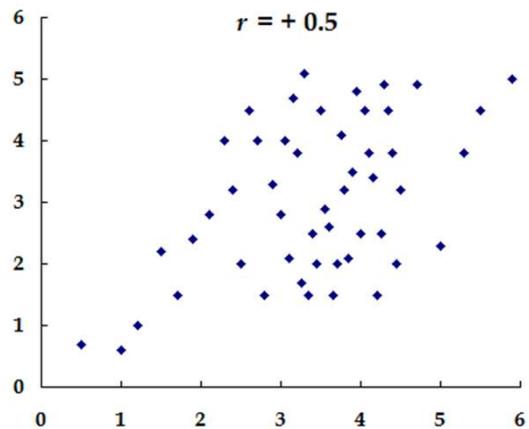
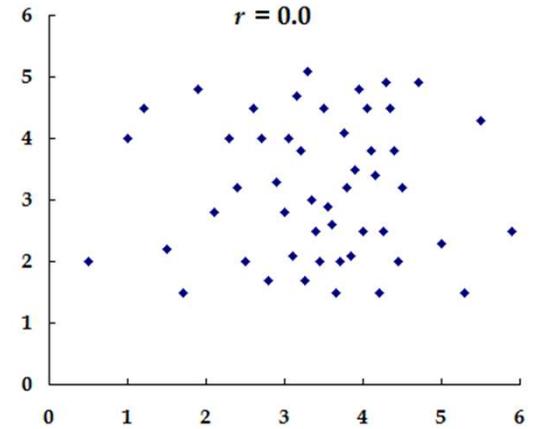
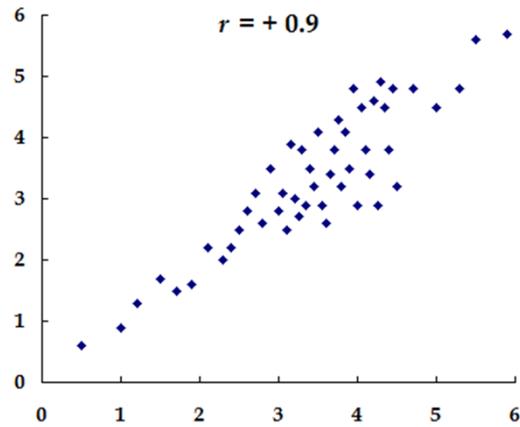
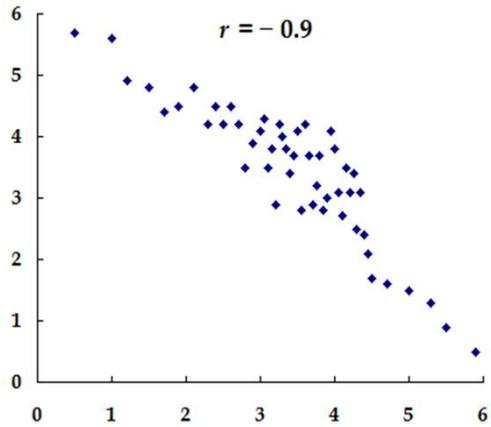
C2 $\text{=CORREL}(B5:B9,C5:C9)$

	A	B	C	D	E
1					
2			-1		
3					
		X	Y		
5		1	-2		
6		2	-4		
7		3	-6		
8		4	-8		
9		5	-10		

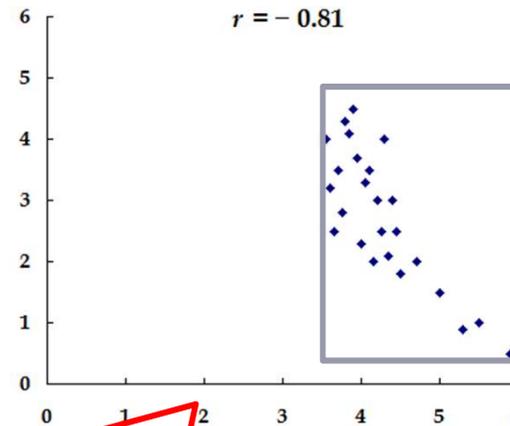
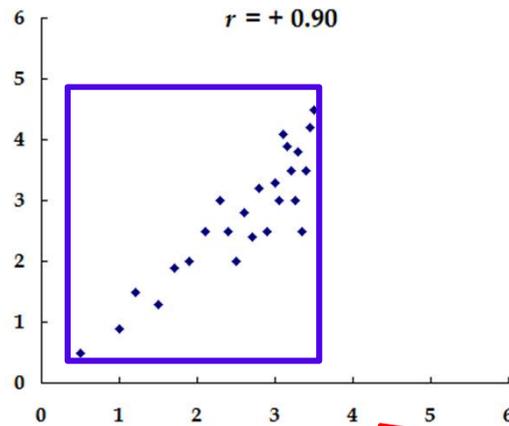
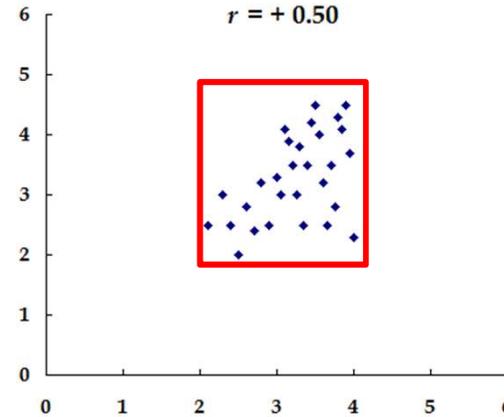
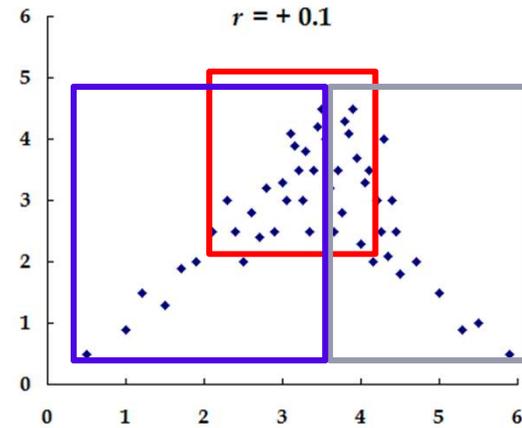
C2 $\text{=CORREL}(B5:B9,C5:C9)$

	A	B	C	D	E
1					
2			-1		
3					
		X	Z		
5		1	-3		
6		2	-6		
7		3	-9		
8		4	-12		
9		5	-15		

선형관계의 기울기와 상관계수



자료의 통합과 분리



분리하려면 Domain Knowledge가 필요

선형관계의 측정: 순위상관계수, 회귀직선과 결정계수

- ◆ 수치형 변수의 상관관계 : 피어슨 상관계수 (Pearson correlation coefficient)
- ◆ 서열척도 변수의 상관관계: 스피어만 순위상관계수(Spearman rank correlation coefficient)
- ◆ 신입사원 10명의 대학졸업평점과 영업성과 평가결과
 (4.2, 3) (3.4, 4) (3.7, 3) (2.8, 5) (3.5, 4) (4.3, 1) (3.5, 5) (3.9, 3) (3.1, 4) (2.7, 2)

C2 =CORREL(D5:D14,G5:G14)

	A	B	C	D	E	F	G
1							
2			-0.44701				
3							
4	ID	졸업평점	평점순위	x_i	성과평가	성과순위	y_i
5	1	4.2	2	2	3	6	7
6	2	3.4	7	7	4	3	4
7	3	3.7	4	4	3	6	7
8	4	2.8	9	9	5	1	1.5
9	5	3.5	5	5.5	4	3	4
10	6	4.3	1	1	1	10	10
11	7	3.5	5	5.5	5	1	1.5
12	8	3.9	3	3	3	6	7
13	9	3.1	8	8	4	3	4
14	10	2.7	10	10	2	9	9

결정계수

◆ 상관계수의 척도: 서열척도 => 곱하기(\times), 나누기(\div) 안됨

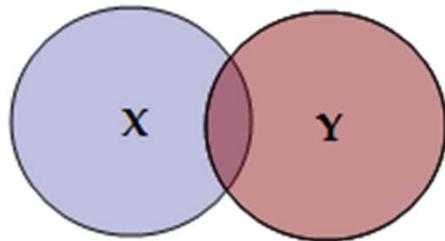
- 상관계수가 0.8인 데이터는 상관계수가 0.4인 데이터 보다 2배의 선형관계가 있는가?
No
- 상관계수가 0.8인 데이터는 상관계수가 0.4인 데이터 보다 선형관계가 강한가? Yes

◆ 상관계수보다 유용한 척도는? 결정계수(coefficient of determination)

- 결정계수 (R^2_{xy})=상관계수² = $(r_{xy})^2$
- 상관계수의 의미 : 종속변수의 변동(변화)를 독립변수가 설명할 수 있는 정도

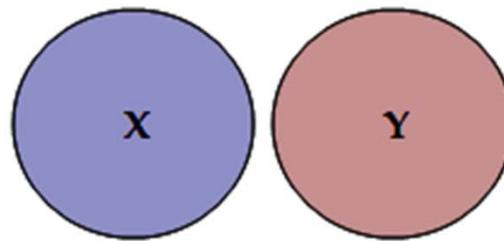
$$r_{xy} = -0.3$$

$$R^2_{xy} = 0.09 \text{ 혹은 } 9\%$$



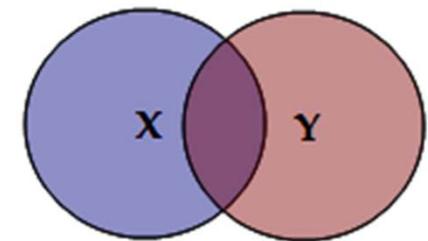
$$r_{xy} = 0.0$$

$$R^2_{xy} = 0.0 \text{ 혹은 } 0\%$$



$$r_{xy} = +0.4$$

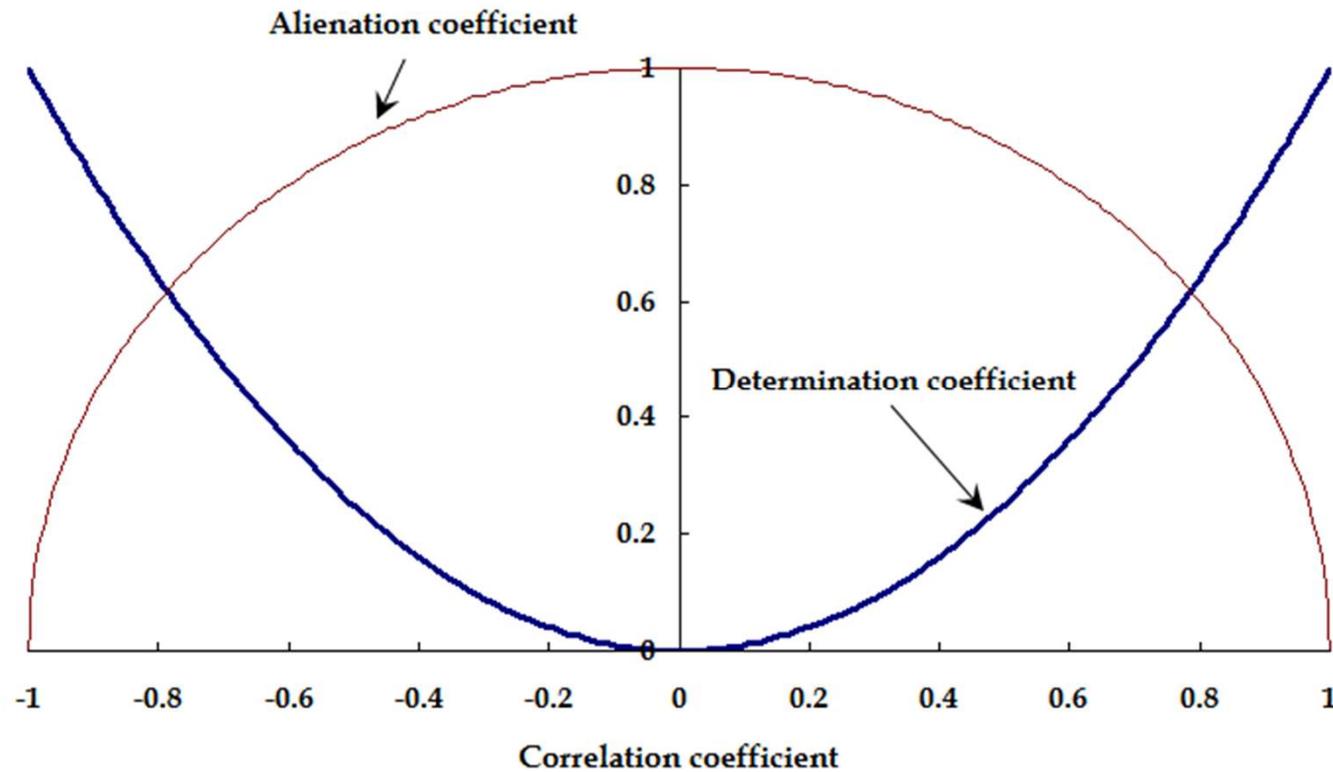
$$R^2_{xy} = 0.16 \text{ 혹은 } 16\%$$



이관계수 (COEFFICIENT OF ALIENATION)

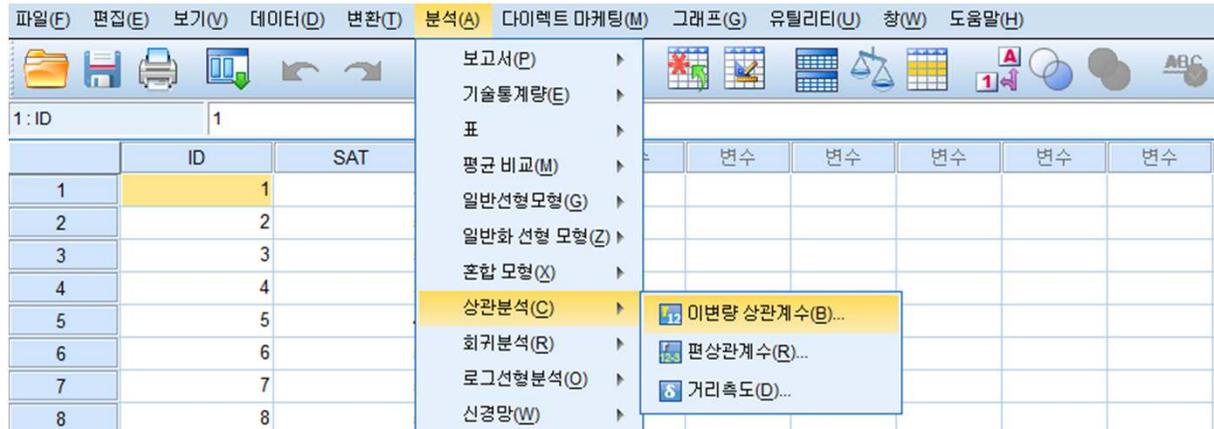
◆ 이관계수: 독립변수로 설명되지 않는 종속변수 변동의 정도

● 이관계수 = $(1 - \text{결정계수})^{1/2}$



상관관계분석 : SPSS 이용과 보고서 작성

◆ SPSS의 활용: 두 수치형(박쥐형) 변수의 관련성 분석



상관관계분석 : SPSS 이용과 보고서 작성

◆ SPSS의 활용: 두 수치형(박쥐형) 변수의 관련성 분석

상관계수

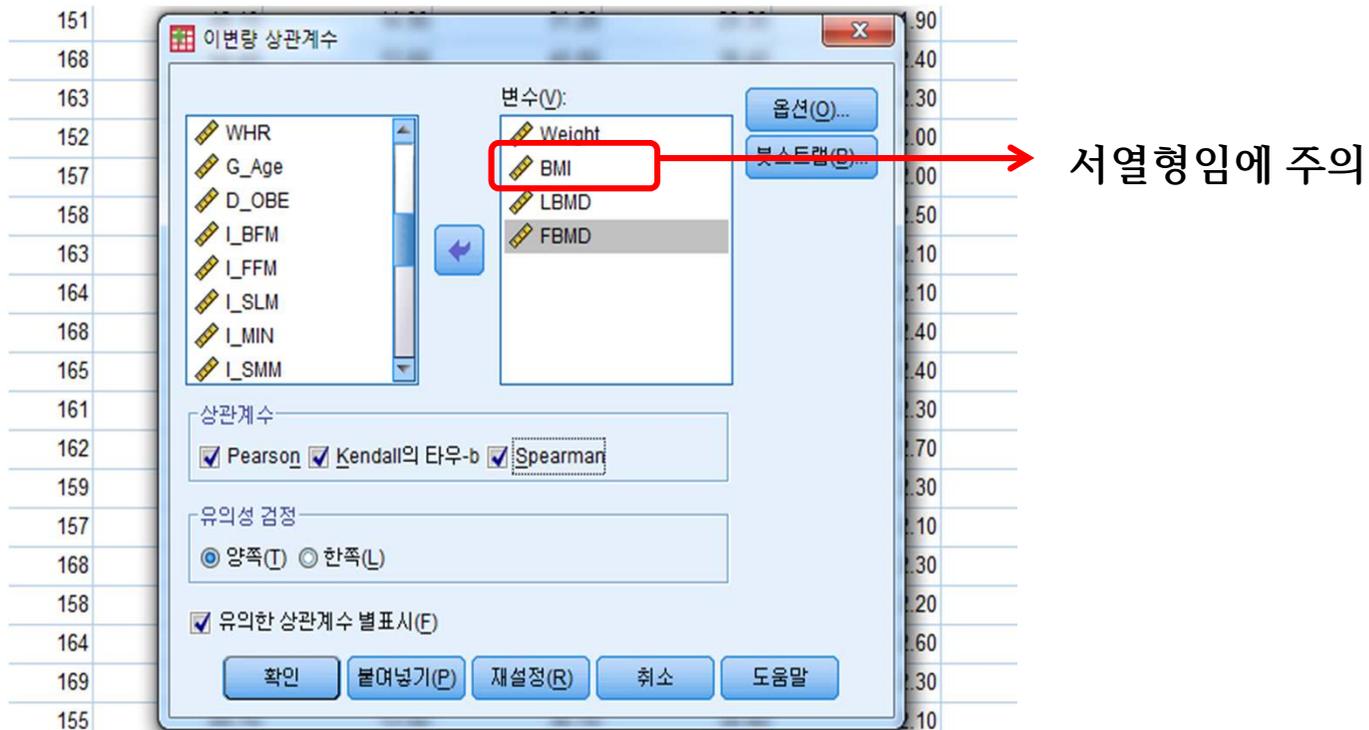
		SAT	REV
SAT	Pearson 상관계수	1	.285 ^{**}
	유의확률 (양쪽)		.001
	N	132	130
REV	Pearson 상관계수	.285 ^{**}	1
	유의확률 (양쪽)	.001	
	N	130	131

H_0 : 모상관계수는 0이다.
 H_1 : 모상관계수는 0이 아니다.

** 상관계수는 0.01 수준(양쪽)에서 유의합니다.

보고서의 작성: 두 수치형(박쥐형) 변수의 선형관계 검정

◆ SPSS의 활용: 두 수치형(박쥐형) 변수의 관련성 분석



보고서의 작성: 두 수치형(박쥐형) 변수의 선형관계 검정

◆ SPSS의 활용: 두 수치형(박쥐형) 변수의 관련성 분석

상관계수

		Weight	BMI	LBMD	FBMD
Weight	Pearson 상관계수	1	.806**	.248**	.413**
	유의확률(양측)		.000	.000	.000
	N	821	821	821	821
BMI	Pearson 상관계수	.806**	1	.124**	.273**
	유의확률(양측)	.000		.000	.000
	N	821	821	821	821
LBMD	Pearson 상관계수	.248**	.124**	1	.656**
	유의확률(양측)	.000	.000		.000
	N	821	821	821	821
FBMD	Pearson 상관계수	.413**	.273**	.656**	1
	유의확률(양측)	.000	.000	.000	
	N	821	821	821	821

** 상관계수는 0.01 수준(양측)에서 유의합니다.

BMI	0.793** (0.000)		
LBMD	0.248** (0.000)	0.098** (0.005)	
FBMD	0.413** (0.000)	0.264** (0.000)	0.656** (0.000)
	Weight	BMI	LBMD

상관계수

			Weight	BMI	LBMD	FBMD
Kendall의 tau_b	Weight	상관계수	1.000	.601**	.146**	.270**
		유의확률(양측)		.000	.000	.000
		N	821	821	821	821
	BMI	상관계수	.601**	1.000	.066**	.182**
		유의확률(양측)	.000		.005	.000
		N	821	821	821	821
	LBMD	상관계수	.146**	.066**	1.000	.470**
		유의확률(양측)	.000	.005		.000
		N	821	821	821	821
FBMD	상관계수	.270**	.182**	.470**	1.000	
	유의확률(양측)	.000	.000	.000		
	N	821	821	821	821	
Spearman의 rho	Weight	상관계수	1.000	.793**	.216**	.389**
		유의확률(양측)		.000	.000	.000
		N	821	821	821	821
	BMI	상관계수	.793**	1.000	.098**	.264**
		유의확률(양측)	.000		.005	.000
		N	821	821	821	821
	LBMD	상관계수	.216**	.098**	1.000	.640**
		유의확률(양측)	.000	.005		.000
		N	821	821	821	821
FBMD	상관계수	.389**	.264**	.640**	1.000	
	유의확률(양측)	.000	.000	.000		
	N	821	821	821	821	

** 상관 유의수준이 0.01입니다(양측).

부분상관계수: 혼동변수의 통제

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
W	21	20	19	22	23	24	27	26	25	28	29	30	33	32	31	34	35	36	39	38	37
S	40	42	44	50	52	54	60	62	64	70	72	74	80	82	84	88	90	92	94	96	98

상관계수

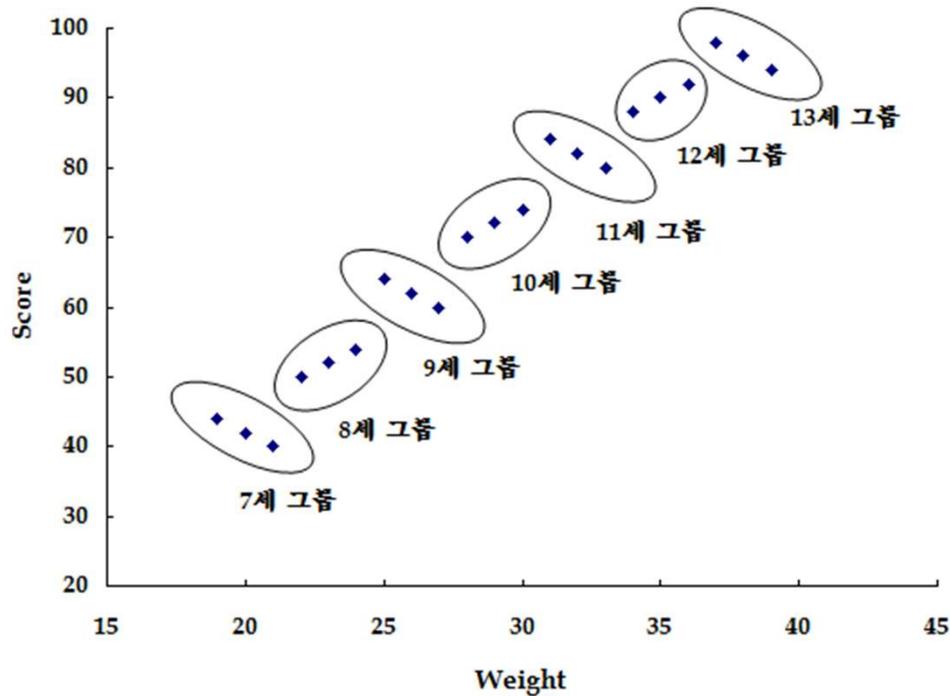
		Weight	Score
Weight	Pearson 상관계수	1	.983 ^{**}
	유의확률 (양쪽)		.000
	N	21	21
Score	Pearson 상관계수	.983 ^{**}	1
	유의확률 (양쪽)	.000	
	N	21	21

** 상관계수는 0.01 수준(양쪽)에서 유의합니다.

부분상관계수: 혼동변수의 통제

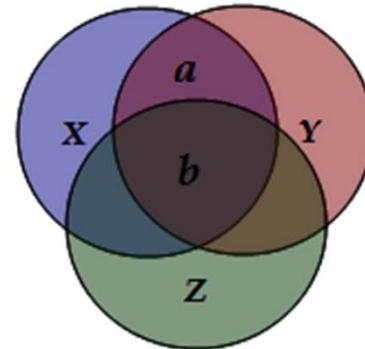
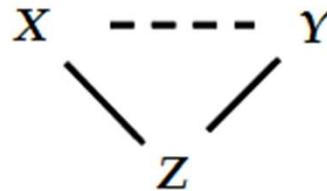
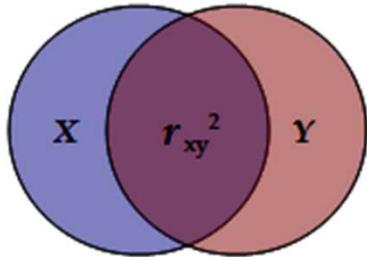
◆ 나이 변수의 추가

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
A	7	7	7	8	8	8	9	9	9	10	10	10	11	11	11	12	12	12	13	13	13
W	21	20	19	22	23	24	27	26	25	28	29	30	33	32	31	34	35	36	39	38	37
S	40	42	44	50	52	54	60	62	64	70	72	74	80	82	84	88	90	92	94	96	98

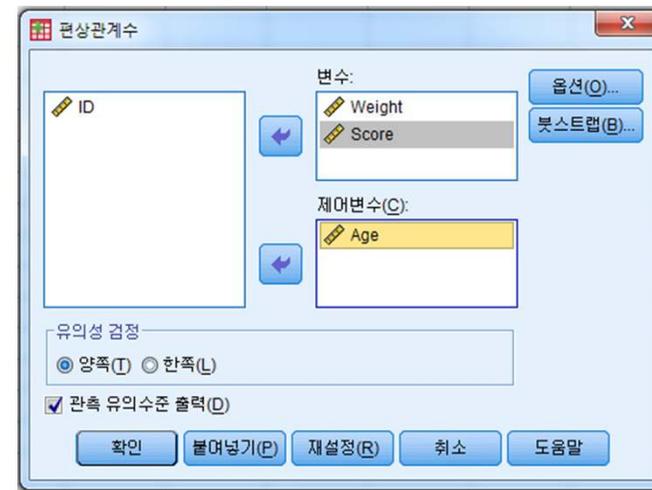
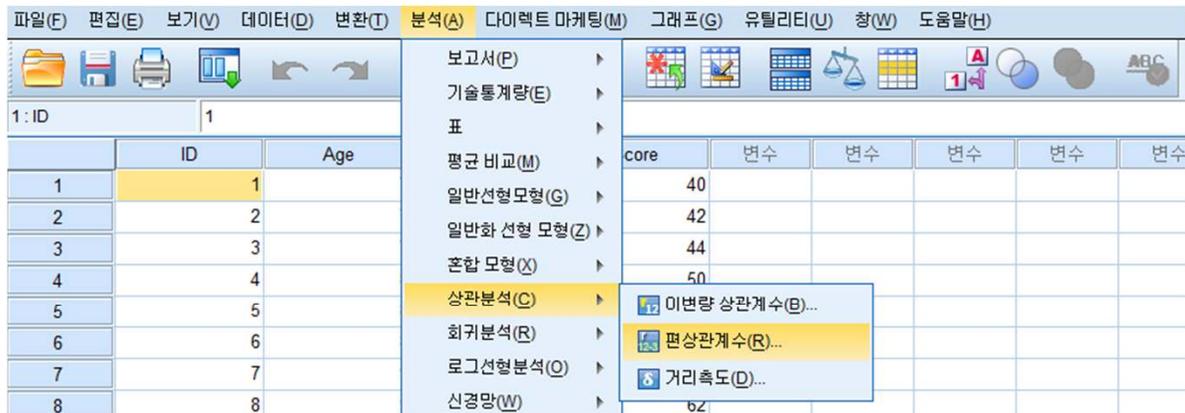


혼동변수와 통제

- ◆ 혼동변수 Z가 X와 Y의 관계에 영향을 주는 부분? b
- ◆ 혼동변수 Z의 통제=b를 제거
- ◆ 앞의 예제에서는 나이가 혼동변수
 - 나이가 동일한 사람끼리 분석



SPSS의 활용: 변수 통제와 부분상관계수



상관

통제변수			Weight	Score
Age	Weight	상관	1.000	-.109
		유의수준(양측)	.	.648
		df	0	18
Score	Score	상관	-.109	1.000
		유의수준(양측)	.648	.
		df	18	0

