AI+X선도인재양성기초프로젝트

9. 비모수검정 & 상관 및 회귀분석

Acknowledgement
한서경, 상관 및 회귀분석, 의학통계론, 서울대학교
박병주, 비모수적 통계 분석법, 의학통계론, 서울대학교
박수경, 반복측정 자료 분석, 의학통계론, 서울대학교

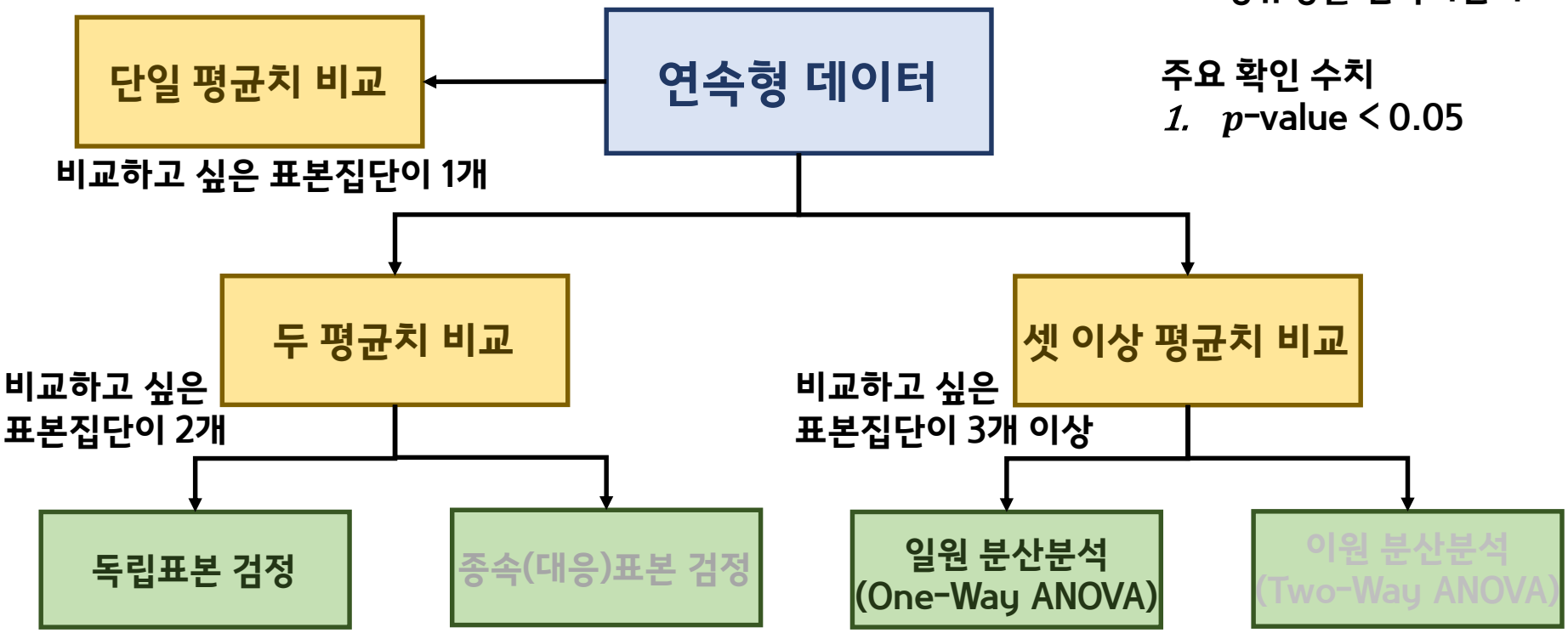
Heenam Yoon

Department of
Human-Centered Artificial Intelligence

E-mail) h-yoon@smu.ac.kr
Room) 0112

Contents

연속형 데이터 분석 review



기본적으로 자료가 정규분포일 때 적용 (or N > 30이상)

연속형 데이터 분석 review

두 평균치 비교

- 2020년 ~ 2021년 100주의 목요일 지하철 이용객 수는 1503.5±600.5명이었고, 월요일 지하철 이용객 수는 1003.5±200.5명이었다. 두 집단의 차이를 인정할 수 있겠는가 (유의수준 0.05)?

연속형 데이터 분석 review

두 평균치 비교

- $(\mu_{thr}, \sigma_{thr})$에서 n_{thr}개씩 뽑아서 나온 $\bar{x}_{thr}$와 $(\mu_{mon}, \sigma_{mon})$에서 n_{mon}개씩 뽑아서 나온 $\bar{x}_{mon}$의 차이 $(\bar{x}_{thr} - \bar{x}_{mon})$들의 분포는

  $(\mu_{thr} - \mu_{mon}, \sqrt{\frac{\sigma_{thr}^2}{n_{thr}} + \frac{\sigma_{mon}^2}{n_{mon}}})$인 정규분포를 따름

$$\frac{(\bar{x}_{thr} - \bar{x}_{mon}) - (\mu_{thr} - \mu_{mon})}{\sqrt{\sigma_{thr}^2/n_{thr} + \sigma_{mon}^2/n_{mon}}} \qquad \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

연속형 데이터 분석 review

두 평균치 비교

- 2020년 ~ 2021년 100주의 목요일 지하철 이용객 수는 1503.5±600.5명이었고, 월요일 지하철 이용객 수는 1003.5±200.5명이었다. 두 집단의 차이를 인정할 수 있겠는가 (유의수준 0.05)?

$$\frac{(\bar{x}_{thr}-\bar{x}_{mon})-(\mu_{thr}-\mu_{mon})}{\sqrt{\sigma_{thr}^2/n_{thr}+\sigma_{mon}^2/n_{mon}}} \rightarrow 0$$

- 귀무가설 (H_0): $\mu_{thr}-\mu_{mon}=0$, 두 평균 간에는 차이가 없다

- $\frac{(\bar{x}_{thr}-\bar{x}_{mon})-(\mu_{thr}-\mu_{mon})}{\sqrt{\sigma_{thr}^2/n_{thr}+\sigma_{mon}^2/n_{mon}}}=\frac{1503.5-1003.5}{\sqrt{600.5^2/100+200.5^2/100}}=7.8978$

- $\alpha=0.05$에 해당하는 값 1.984보다 크므로 $p<0.05$
- 즉, 두 집단의 차이가 인정됨

  = 목요일과 월요일 지하철 승객수에 차이가 있다

연속형 데이터 분석 review

2019년 1월 ~ 6월 중 목요일에 지하철 승객수가 많은가?

- 자료 정리를 어떻게 할지는 누가 정하는가?
- 통계분석을 어떻게 할 것인지는 누가 정하는가?

- 단, 모두가 납득할 수 있는 검정을 하면 좋다

연속형 데이터 분석 review

2019년 1월 ~ 6월 중 목요일에 지하철 승객수가 많은가?

1월

월	화	수	목	금	토	일
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31	1	2	3

A역	1000
B역	1500
C역	5000
D역	3000
…	…
평균	3254.5

요일별로 비교하자

- 월, …, 일 비교
- ANOVA 또는 비모수 방법으로 ANOVA준하는 검정

각 요일의 지하철 승객수를 어떻게 정리할까?

- 각 요일에 지하철 역의 개수가 다르다 = 합으로 계산하면 문제가 생길 것이다

  → 평균으로 하자

연속형 데이터 분석 review

2019년 1월 ~ 6월 중 목요일에 지하철 승객수가 많은가?

월

주차	평균
1월 첫 주	
1월 둘째 주	
1월 셋째 주	
1월 넷째 주	
…	
6월 넷째 주	

화

주차	평균
1월 첫 주	
1월 둘째 주	
1월 셋째 주	
1월 넷째 주	
…	
6월 넷째 주	

…

일

주차	평균
1월 첫 주	
1월 둘째 주	
1월 셋째 주	
1월 넷째 주	
…	
6월 넷째 주	

1월 ~ 6월, 4주 * 6개월 = 24개 (요일의 개수)

- N < 30 (여기서 N은 각 요일 별 데이터 개수임)
- 즉, N이 30보다 크다라는 것은 모든 요일의 승객수 데이터가 30개 이상이라는 의미
  (총 데이터 개수 24 * 7 가 30개 보다 크다는 의미가 아님)
- 정규성 검정 후, 결정 필요

사실 각 요일은 정확히 24개가 아님. 달은 31, 30, 28일로 구성

연속형 데이터 분석 review

2019년 1월 ~ 6월 중 목요일에 지하철 승객수가 많은가?

1월

월	화	수	목	금	토	일
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31	1	2	3

A역	1000
B역	1500
C역	5000
D역	3000
…	…

A역	1100
B역	1400
C역	2500
D역	4000
…	…

이렇게 접근한다면?

요일별로 지하철 승객수를 누적

- 예. 6개월간 일요일 수 * 매 일요일의 지하철 역 수

연속형 데이터 분석 review

2019년 1월 ~ 6월 중 목요일에 지하철 승객수가 많은가?

월

1	1월 1주	A역
2		B역
3		C역
…		…
…	1월 2주	A역
…		…

화

1	1월 1주	A역
2		B역
3		C역
…		…
…	1월 2주	A역
…		…

…

일

1	1월 1주	A역
2		B역
3		C역
…		…
…	1월 2주	A역
…		…

N이 매우 많아지므로 ANOVA분석을 하면 됨

그런데, 한가지 고민사항이 있음

역 별로 승객수 편차가 클 수 있다 = 이것이 새로운 변인으로 작용

우리가 알고 싶은 것은 "2019년 1월 ~ 6월 중 목요일에 지하철 승객수가 많은가"가 알고 싶다.

이상적인 경우, 관심 변인 (변수)는 하나, 그 외의 변인은 통제해야 한다. 이를 변인 통제라고 한다.

위 경우, 승객수 편차가 하나의 변인으로 작용할 수 있다

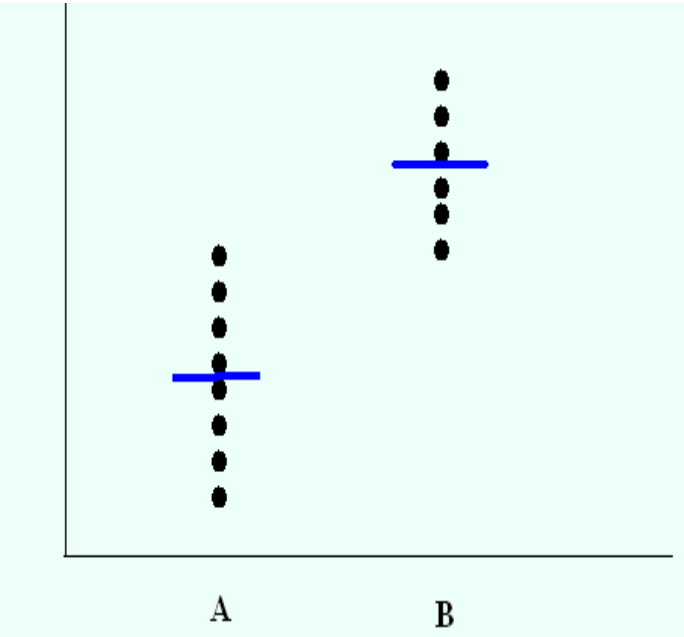
(N이 충분이 많아지면, 괜찮아 질 것 같긴하다)

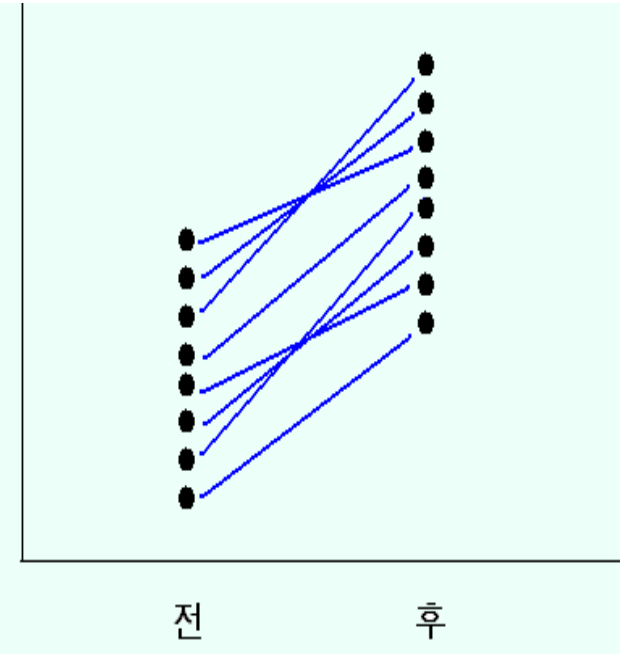
반복 측정 데이터 분석

반복 측정 데이터 분석

개념

- 두 종속 표본에서의 연속/비연속 변수 통계분석 방법
- 동일한 대상자에 대하여 어떤 사건 전/후, 측정/재측정 자료 등의 분석

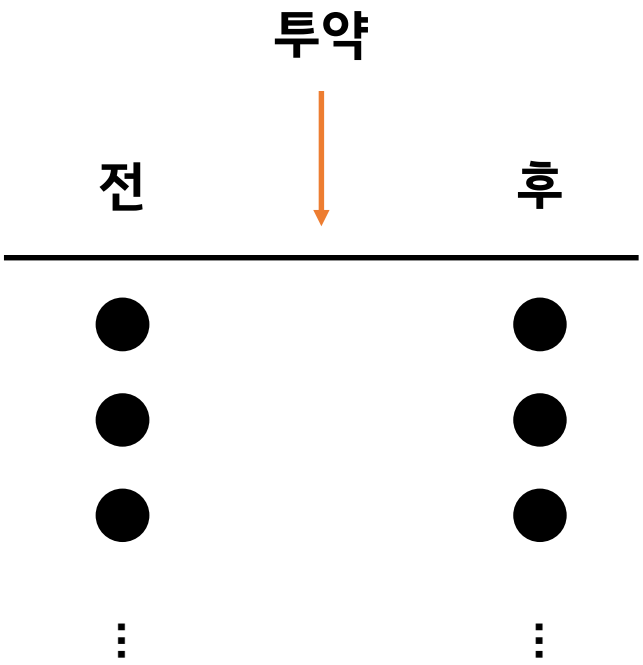




반복 측정 데이터 분석

반복 측정의 예시

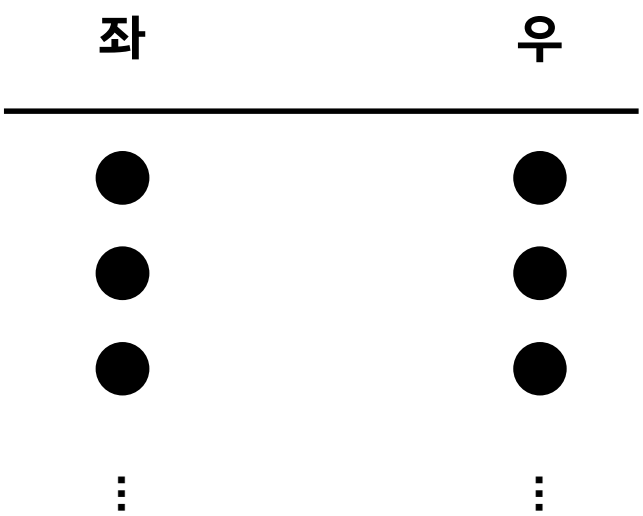
- 동일한 사람을 대상으로 특정 조건(약, 수술) 전후의 비교
- 차이 평가



반복 측정 데이터 분석

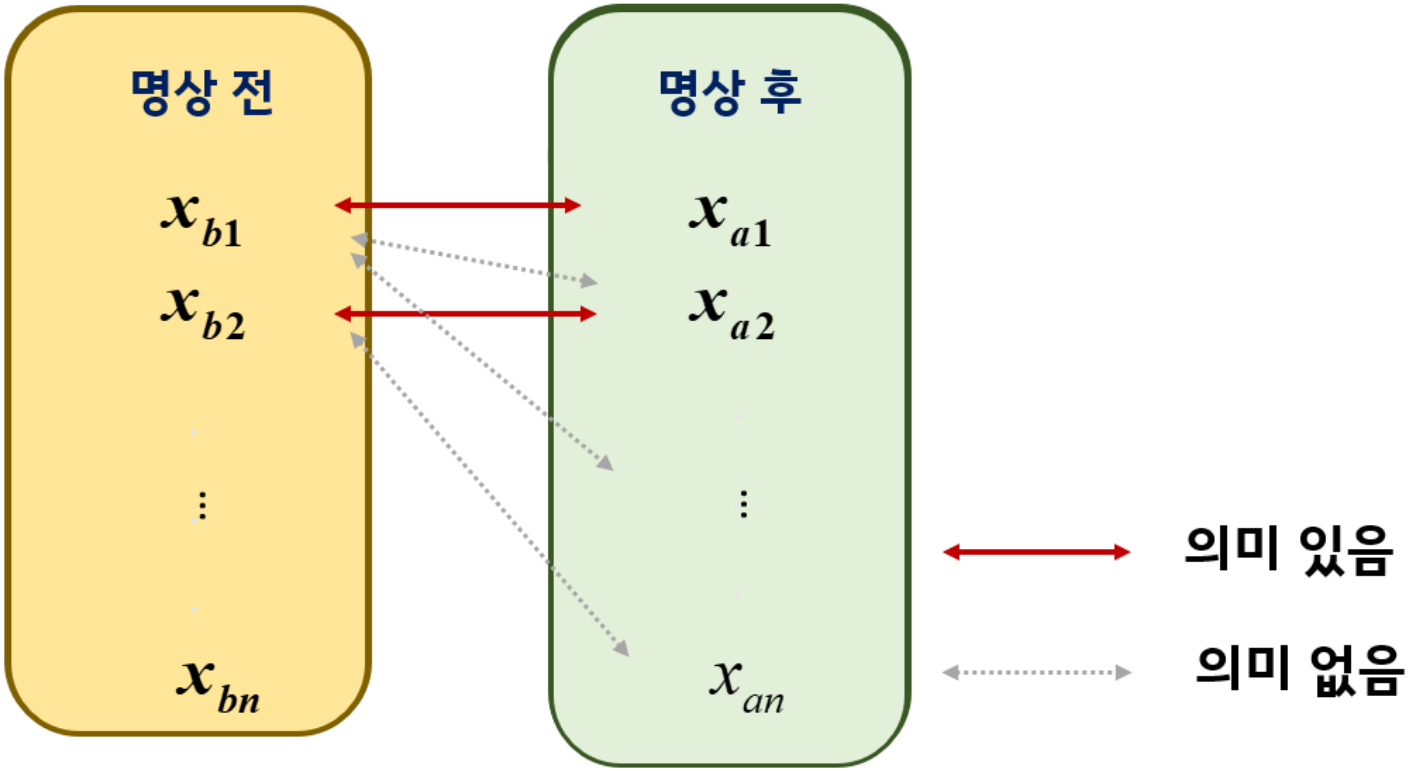
반복 측정의 예시

- 동일한 사람의 신체 두 부위 비교
- 차이 평가



반복 측정 데이터 분석

연속, 모수: paired t-test



반복 측정 데이터 분석

연속, 모수: paired t-test

- 명상 전: 98.8±11.9
- 명상 후: 90.7±15.7
- $\overline{d} = 8.1, \overline{S}_d^2 = 37.7$
- $H_0 = \overline{x}_a - \overline{x}_b = d = 0$
- $t = \frac{8.1}{\sqrt{37.7/10}} = 4.2$

	1	2	3	4	5	6	7	8	9	10	평균	분산
전	106	93	127	94	102	82	97	96	91	100	98.8	-
후	108	88	126	90	90	70	82	80	85	88	90.7	-
차	-2	5	1	4	12	12	15	16	6	12	8.1	37.7

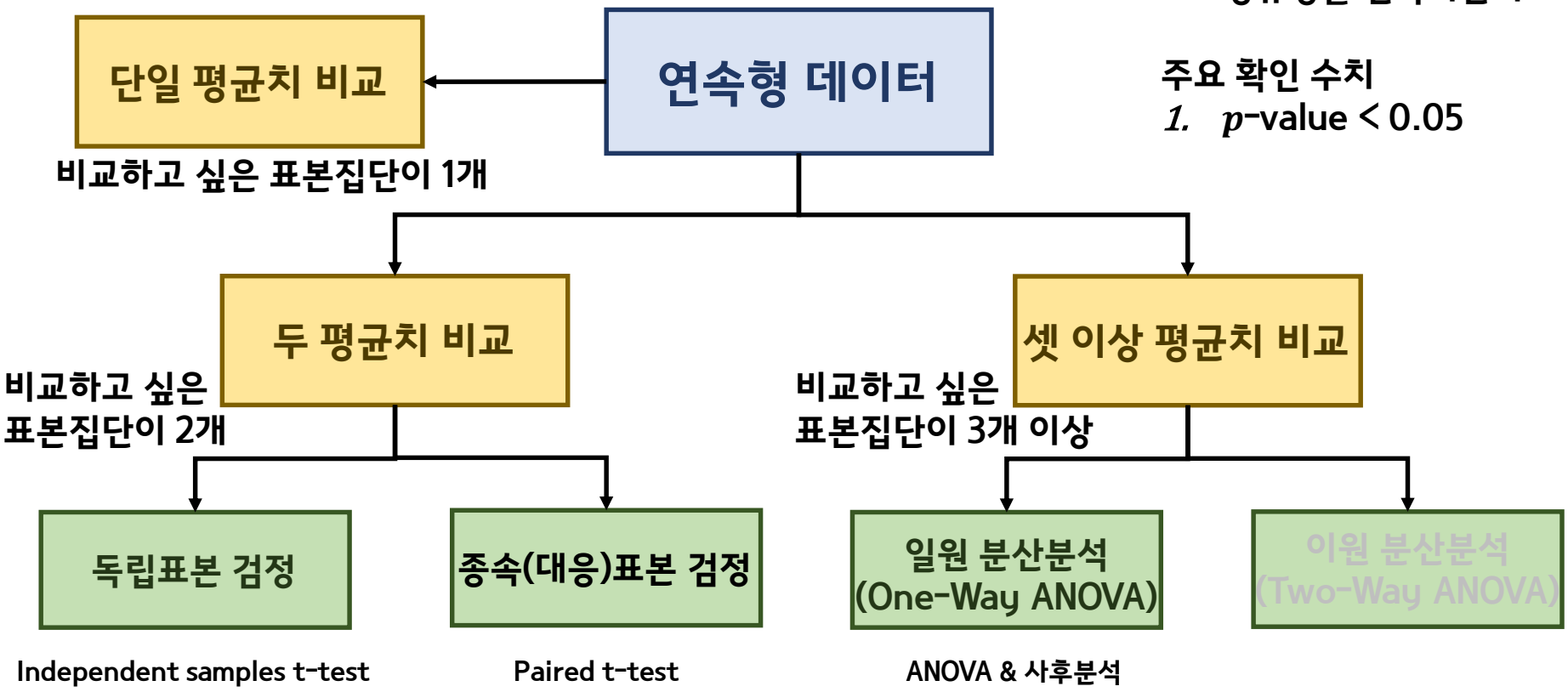
반복 측정 데이터 분석

연속, 모수: paired t-test

- 자유도 9에서 α=0.05에 해당하는 t값은 2.26이므로 $p < 0.05$로 명상이 심박수를 감소시키는데 효과가 있다고 말할 수 있다

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

연속형 데이터 분석



기본적으로 자료가 정규분포일 때 적용 (or N > 30이상)

연속형 데이터 분석 review

2019년 1월 ~ 6월 중 목요일에 지하철 승객수가 많은가?

1월

월	화	수	목	금	토	일
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31	1	2	3

A역	1000
B역	1500
C역	5000
D역	3000
…	…

평균 3254.5

목, 금을 비교해보자

각 요일의 지하철 승객수를 어떻게 정리할까?

- 각 요일에 지하철 역의 개수가 다르다 = 합으로 계산하면 문제가 생길 것이다

  → 평균으로 하자

연속형 데이터 분석 review

2019년 1월 ~ 6월 중 목요일에 지하철 승객수가 많은가?

목

주차	평균
1월 첫 주	
1월 둘째 주	
1월 셋째 주	
1월 넷째 주	
…	
6월 넷째 주	

금

주차	평균
1월 첫 주	
1월 둘째 주	
1월 셋째 주	
1월 넷째 주	
…	
6월 넷째 주	

차이

대응된 비교가 될 것 같다고 생각함

꽤 오래 생각해본 결과,

반복측정 (대응표본) 분석은 적용하지 않는 것이 좋을 것 같다고 판단

일단, 동일 대상자에 대한 자료가 아님

차이의 비교이므로 증/감에 영향을 받음

…

비모수 검정

비모수 검정 개요

비모수적 통계분석의 개념

- 모수에 관하여 특수한 분포를 전제하지 않음
- 평균치와 같은 어떤 특정값(parameter)이 아닌 분포형태 자체를 분석
- 통계적 검정에 이용되는 귀무가설 하에서의 검정통계량의 분포와 모집단의 분포함수를 전혀 무관하게 하여 분석

- Wilcoxon 검정법을 효시로 하여 비모수적 방법에 대한 연구가 본격화됨

- 응용 분야: 가설검정, 점추정, 구간추정, 분산분석, 회귀분석, 시계열분석

비모수 검정 개요

- 비모수검정법의 장점
  - 모집단 분포에 무관하게 정확한 확률 산출
  - N이 적은 경우 적용할 때 유용
  - 계산이 쉽고 간단함
  - 이해하기 쉬움
  - 척도에 제한이 없음: 명칭척도, 순위척도

- 비모수검정법의 단점
  - N이 커지면 오히려 계산이 복잡하고 어려워짐

변수의 척도

비연속 데이터 (질적 변수)	연속 데이터 (양적 변수)
명칭척도 (nominal scale) 특정 상태를 지칭 혈액형, 성별, 인종, 실험군 (대조군/치료군) 치료결과 (호전, 재발, 사망)	간격척도 (interval scale) 특정 상태 + 서열 + 측정치간 간격 온도 (0의 상태의 개념화)
순위척도 (ordinal scale) 특정 상태 + 각 범주간 서열, 교육정도, 사회경제적 수준, 병리조직학적 소견(-/±/+/++/+++), 치료 정도 (반응, 중간반응, 무반응)	비척도 (ratio scale) 간격척도 특성 + 절대 영점 연령, 혈압, 체중, 신장

비모수 검정법 요약

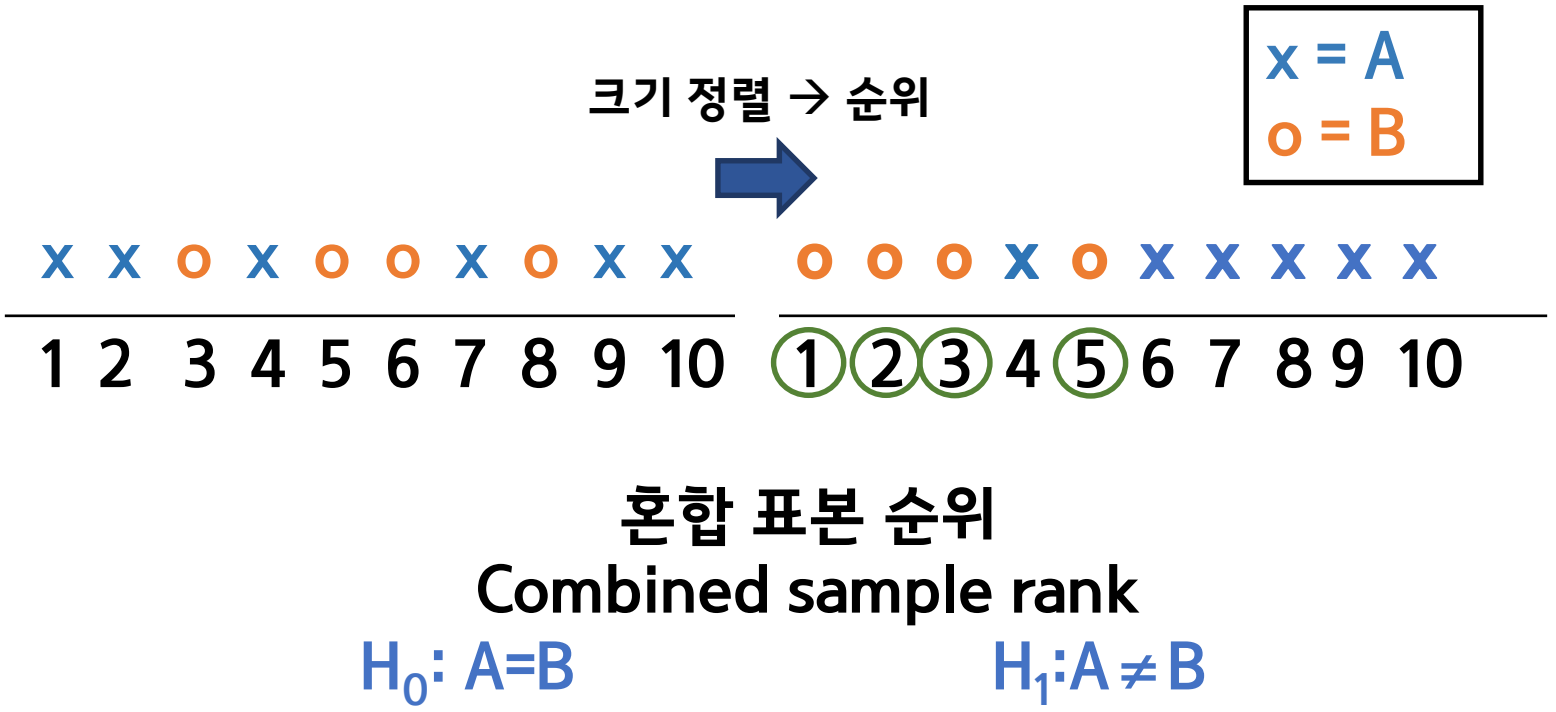
측정 수준	단일 표본	두 표본		k 표본		상관
		종속	독립	종속	독립	
명칭척도	·binomial ·x^2	·McNemar	·Fisher ·x^2	·Cochran Q	·x^2	·분할계수
순위척도	·K-S ·Runs	·sign ·Wilcoxon signed rank	·median ·M-W의 U ·K-S ·Moses ·Wilcoxon rank sum	·Friedman 2-way ANOVA	·median ·K-W의 1-way ANOVA	·Spearman의 순위상관계수 ·Kendall의 순위상관계수 ·Kendall의 편순위상관계수 ·Kendall의 일치도계수
간격척도		·Walsh ·무작위검정	·무작위검정			

다표본 비모수 검정의 핵심은 순위비교!

비모수: 두 독립표본의 검정

순위합 검정의 개념

- 두 모집단 A, B의 관측치
- 혼합
- 크기 순으로 일직선상 나열



혼합 표본 순위

Combined sample rank

H_0: A=B H_1: A≠B

비모수: 두 독립표본의 검정

예시. Wilcoxon의 순위합 검정(rank sum)

- A와 B의 두 식이요법에 차이가 있는지를 비교하기 위해 증세가 비슷한 10명의 환자를 5명씩 임의로 두 집단으로 나눈 뒤 각각에 A와 B 두 식이요법을 6주간 시행 후 체중의 증가량을 조사한 결과 다음 자료를 얻었다. 두 요법 간에 유의한 차이가 있는가?

요법 A	5.7	7.3	7.6	6.0	6.5
요법 B	4.9	7.4	5.3	4.6	6.2

  - H_0: A와 B 두 요법 간에 유의한 차이가 **없다**
  - H_1: A와 B 두 요법 간에 유의한 차이가 **있다**

비모수: 두 독립표본의 검정

예시. Wilcoxon의 순위합 검정(rank sum)

요법 A	5.7	7.3	7.6	6.0	6.5
요법 B	4.9	7.4	5.3	4.6	6.2

혼합표본	4.6	4.9	5.3	5.7	6.0	6.2	6.5	7.3	7.4	7.6
순위	1	2	3	4	5	6	7	8	9	10

- Wilcoxon의 순위합 통계량
  - W_A= A 순위 합 = 34
  - W_B = B 순위 합 = 21
  - W_S = 작은 표본의 순위 합 = 34

    (표본크기가 같으므로 선택 가능)

비모수: 두 독립표본의 검정

예시. Wilcoxon의 순위합 검정(rank sum)

- 자료의 수 5개, 5개
- 순위: 1~10, 합 55
- 양측검정 ($\alpha = 0.05 \rightarrow 0.025$)
  - $P(W_s \geq 37) = 0.028$
  - $P(W_s \leq 18) = 0.028$
  - $W_S = 34$로 H_0를 기각할 수 없음
  - 즉, 요법 간 차이 없음 ($p > 0.05$)

(Wilcoxon 순위합통계표)

Smaller sample size=5

Larger sample size

5			6			7			8		
x	P	x^*	x	P	x^*	x	P	x^*	x	P	x^*
34	.111	21	37	.123	23	41	.101	24	44	.111	26
35	.075	20	38	.089	22	42	.074	23	45	.085	25
36	.048	19	39	.063	21	43	.053	22	46	.064	24
37	.028	18	40	.041	20	44	.037	21	47	.047	23
38	.016	17	41	.026	19	45	.024	20	48	.033	22
39	.008	16	42	.015	18	46	.015	19	49	.023	21
			43	.009	17	47	.009	18	50	.015	20
									51	.009	19

9			10		
x	P	x^*	x	P	x^*
17	.120	28	51	.103	29
18	.095	27	52	.082	28
19	.073	26	53	.065	27
50	.056	25	54	.050	26
51	.041	24	55	.038	25
52	.030	23	56	.028	24
53	.021	22	57	.020	23
54	.014	21	58	.014	22
55	.009	20	59	.010	21

비모수: 두 독립표본의 검정

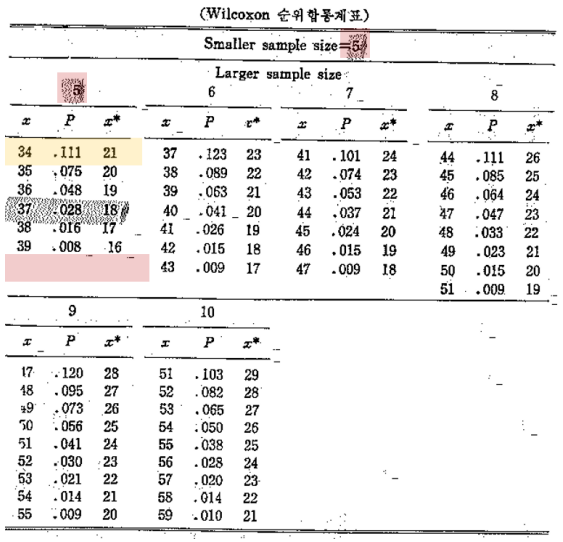
예시. Wilcoxon의 순위합 검정(rank sum)

치료 A	4.5	4.2	4.7	3.3	3.4
치료 B	1.8	2.6	2.3	1.5	3.1

혼합표본	1.5	1.8	2.3	2.6	3.1	3.3	3.4	4.2	4.5	4.7
순위	1	2	3	4	5	6	7	8	9	10

- Wilcoxon의 순위합 통계량
  - W_A= A 순위 합 = 15
  - W_B = B 순위 합 = 40
  - W_S = 작은 표본의 순위 합 = 15

    (표본크기가 같으므로 선택 가능)



(Wilcoxon 순위합통계표)

Smaller sample size=5

Larger sample size

5			6			7			8		
x	P	x^*	x	P	x^*	x	P	x^*	x	P	x^*
34	.111	21	37	.123	23	41	.101	24	44	.111	26
35	.075	20	38	.089	22	42	.074	23	45	.085	25
36	.048	19	39	.063	21	43	.053	22	46	.064	24
37	.028	18	40	.041	20	44	.037	21	47	.047	23
38	.016	17	41	.026	19	45	.024	20	48	.033	22
39	.008	16	42	.015	18	46	.015	19	49	.023	21
			43	.009	17	47	.009	18	50	.015	20
									51	.009	19

9			10		
x	P	x^*	x	P	x^*
17	.120	28	51	.103	29
18	.095	27	52	.082	28
19	.073	26	53	.065	27
50	.056	25	54	.050	26
51	.041	24	55	.038	25
52	.030	23	56	.028	24
53	.021	22	57	.020	23
54	.014	21	58	.014	22
55	.009	20	59	.010	21

비모수: 두 독립표본의 검정

Mann-Whitney U test

- U의 값은 n_B집단에서의 순위가 n_A집단에서의 순위를 앞서는 횟수에 따라 결정

- 예시
  - 젊은 5명과 나이가 많은 4명의 심박수 평균의 차이가 있는가?

E	58.1	56.5	60.3	56.2	
Y	64.7	60.9	63.5	67.4	57.2

  - H_0: 두 그룹 간의 심박수에 차이가 없다
  - H_1: 두 그룹 간의 심박수에 차이가 있다

비모수: 두 독립표본의 검정

Mann-Whitney U test

- **혼합표본**

E	58.1	56.5	60.3	56.2	
Y	64.7	60.9	63.5	67.4	57.2

…

혼합표본	56.2	56.5	57.2	58.1	60.3	60.9	63.5	64.7	67.4
순 위	1	2	3	4	5	6	7	8	9

- **E가 Y를 앞선 횟수**
  - U = 0 + 0 + 1 + 1 = 2
- **H_0하에서의 U의 표본분포를 이용하여 관찰된 U 또는 더 극단적인 값이 H_0 하에서 발생될 확률을 산출**

비모수: 두 독립표본의 검정

Mann-Whitney U test

- n_A와 n_B가 많아지면 세기 힘듦 (참고: 같은 값이 있을 땐 순위 평균)

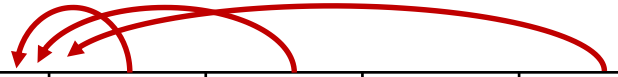
- R_A: n_A의 순위 합
- R_B: n_B의 순위 합
- U_A: $n_A \cdot n_B + \frac{n_A(n_A+1)}{2} - R_A$
- U_B: $n_A \cdot n_B + \frac{n_B(n_B+1)}{2} - R_B$

- U_A와 U_B 중 작은 것을 U로 선택

비모수: 두 독립표본의 검정

Mann-Whitney U test

E	58.1	56.5	60.3	56.2	
Y	64.7	60.9	63.5	67.4	57.2



…

혼합표본	56.2	56.5	57.2	58.1	60.3	60.9	63.5	64.7	67.4
순 위	1	2	3	4	5	6	7	8	9

- $R_A = 1 + 2 + 4 + 5 = 12$
- $R_B = 3 + 6 + 7 + 8 + 9 = 33$
- $U_A: n_A \cdot n_B + \frac{n_A(n_A+1)}{2} - R_A = 4 \times 5 + \{4 \times (4+1)\}/2 - 12 = 18$
- $U_B: n_A \cdot n_B + \frac{n_B(n_B+1)}{2} - R_B = 4 \times 5 + \{5 \times (5+1)\}/2 - 33 = 2$
- U_A, U_B 중 작은 것을 U로 선택, $U = 2$

비모수: 두 독립표본의 검정

Mann-Whitney U test

- $n_A = 4, n_B = 5, U = 2$ 일 때, p = 0.032, 양측검정이므로 p = 0.064
- $p > 0.05$ 이므로 H_0기각 불가 → 두 그룹 간 심박수 차이가 없다

$n_2 = 5$ 단측검정 값

U \ n_1	1	2	3	4	5
0	.167	.047	.018	.008	.004
1	.333	.095	.036	.016	.008
2	.500	.190	.071	.032	.016
3	.667	.286	.125	.056	.028
4		.429	.196	.095	.048
5		.571	.286	.143	.075
6			.393	.206	.111
7			.500	.278	.155
8			.607	.365	.210
9				.452	.274
10				.548	.345
11					.421
12					.500
13					.579

비모수: 두 종속표본의 검정

종류

- 부호 검정(Sign test)
- Wilcoxon signed rank test

비모수: 두 종속표본의 검정

부호검정

- 동일 개체에 행해진 두 처치 간의 차이
  - 처치 A → 결과 x_A
  - 처치 B → 결과 x_B
  - $x_A > x_B \rightarrow (+)$
  - $x_A < x_B \rightarrow (-)$

- H_0: $P(x_A > x_B) = P(x_A < x_B) = 1/2$

비모수: 두 종속표본의 검정

예시: 부호검정

- 15명의 위궤양 환자에 대한 일정기간 치료 전후에 위내시경검사를 시행하여 그 중증도에 따라 각각 0,1,2,3,4,5의 등급을 매겼다. 다음의 결과로부터 치료에 의한 상태가 호전되었다고 말할 수 있을까?

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
전	3	2	2	3	2	1	3	3	4	5	1	4	5	3	4
후	4	2	1	2	1	2	3	2	2	4	0	2	3	2	2
차	-	0	+	+	+	-	0	+	+	+	+	+	+	+	+

- H_0: 치료 전후에 차이가 없다
- H_1: 치료 전후에 차이가 있다

비모수: 두 종속표본의 검정

예시: 부호검정

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
전	3	2	2	3	2	1	3	3	4	5	1	4	5	3	4
후	4	2	1	2	1	2	3	2	2	4	0	2	3	2	2
차	-	0	+	+	+	-	0	+	+	+	+	+	+	+	+

- 전후에 차이가 없는 경우는 전체 표본수에서 제외 (n=15-2)
- Sign test의 확률표에서 n=13, x=2일 때, H_1의 p = 0.011 < 0.05
- H_1채택: 치료 전후에 차이가 있다

비모수: 두 종속표본의 검정

예시: 부호검정

20-1. Sign test 에서의 확률표

(1) 편측검정

n(표본수) \ P_r	.05 $x(P_r)$	.025 $x(P_r)$	.01 $x(P_r)$	.005 $x(P_r)$
5	0(.031)	—	—	—
6	0(.016)	0(.016)	—	—
7	0(.008)	0(.008)	0(.008)	—
8	1(.035)	0(.004)	0(.001)	0(.004)
9	1(.020)	1(.020)	0(.002)	0(.002)
10	1(.011)	1(.011)	0(.001)	0(.001)
11	2(.033)	1(.006)	1(.006)	0(.005)
12	2(.019)	2(.019)	1(.003)	1(.003)
13	3(.046)	2(.011)	1(.002)	1(.002)
14	3(.029)	2(.006)	2(.006)	1(.001)
15	3(.018)	3(.018)	2(.004)	2(.004)
16	4(.038)	3(.011)	2(.002)	2(.002)
17	4(.025)	4(.025)	3(.006)	2(.001)
18	5(.048)	4(.015)	3(.004)	3(.004)
19	5(.032)	4(.015)	4(.010)	3(.002)
20	5(.021)	5(.021)	4(.006)	3(.001)
21	6(.039)	5(.013)	4(.004)	4(.004)
22	6(.026)	5(.008)	5(.008)	4(.002)

비모수: 두 종속표본의 검정

Wilcoxon signed rank test

- 부호 + 상대적 크기를 고려
- 부호검정보다 검정력이 높음

- d_i: 짝지워진 점수들의 차이
- 가장 작은 $|d_i|$ 부터 순위를 매김
  - 각 순위에 차이의 부호를 붙임: T_+ ~T_-
  - 동점일 경우는 분석대상에서 제외시시킴 (d_i =0일 경우)
  - 차이 d_i가 같은 경우에는 평균순위를 부여

비모수: 두 종속표본의 검정

예시. Wilcoxon signed rank test

- 8명의 고혈압환자에서 특정 혈압강하제를 투여하여 아래와 같은 결과를 얻었다. 약제를 투여한 후 혈압이 유의하게 내렸다고 말할 수 있는가?

환자	전	후	차이	순위
1	160	152	-8	2
2	162	175	+13	5
3	168	150	-18	7
4	165	179	+14	6
5	170	160	-10	3
6	167	155	-12	4
7	176	149	-27	8
8	163	156	-7	1

$T_+ = 11$

비모수: 두 종속표본의 검정

예시. Wilcoxon signed rank test

- $n = 8,\ \alpha \leq 0.05$ → 3, 33 (총합: 36, 1~8까지의 순위 합)
- 즉, T_+ or T_- 값 중 작은 값이 3이하이거나 큰 값이 33이상일 확률이 5%(0.05) 이하
- $T_+ = 11 > 3, T_- = 25 < 33$
- $p > 0.05$ (H_0채택)
- 약제를 투여한 후 혈압이 유의하게 내렸다고 말할 수 **없다**

21. Wilcoxon 부호화 순위 검정확률표

	확률수준			
n	$\alpha \leqq 0.10$	$\alpha \leqq 0.05$	$\alpha \leqq 0.02$	$\alpha \leqq 0.01$
1				
2				
3				
4				
5	0, 15			
6	2, 19	0, 21		
7	3, 25	2, 26	0, 28	
8	5, 31	3, 33	1, 35	0, 36
9	8, 37	5, 40	3, 42	1, 44
10	10, 45	8, 47	5, 50	3, 52
11	13, 53	10, 56	7, 59	5, 61
12	17, 61	13, 65	9, 69	7, 71
13	21, 70	17, 74	12, 79	9, 82
14	25, 80	21, 84	15, 90	12, 93
15	30, 90	25, 95	19, 101	15, 105

비모수: 독립적인 k-표본

모수와 비모수 검정 비교

- One-way ANOVA / F-test
- Kruskal-Wallis One-way ANOVA by Ranks

비모수: 독립적인 k-표본

Kruskal-Wallis test

- k 표본들의 관측치 결합 (k > 2)
- 연속적인 순서로 배열
- 가장 작은 값부터 순위 부여
- 각 표본별 순위합
- 순위합 차이가 동일 모집단의 표본들로부터 나올 확률 계산
- k=3, 각 표본에서의 크기 ≤ 5일 때 Kruskal-Wallis의 정확한 H값을 계산
- 표본에서의 크기가 > 5 인 경우,

$$H = \frac{12}{N(N+1)} \cdot \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(N+1)$$

- k: 표본수
- n_j: j 번째 표본수
- R_j: j 번째 표본의 순위 합
- $N = \sum N_j$

비모수: 독립적인 k-표본

Kruskal-Wallis test

- 동점 관측치 출현 시, 평균순위로 계산 후 검정통계량 H 수정

$$H^* = \frac{H}{1 - \sum_{i=1}^{l} q_i(q_i^2 - 1)/N(N^2 - 1)}$$

- l: 동점 경우의 수
- q_i: i 번째 동점에서의 관측치

- df: $k - 1$

비모수: 독립적인 k-표본

예시: Kruskal-Wallis test

- 세 종류의 식이요법에 대한 효과를 조사하기 위하여 같은 종류의 쥐를 무작위로 세 집단에 배정한 뒤, 각 집단별로 다른 식이요법을 적용하였다. 일정기간 후에 쥐의 콩팥과 내장에 낀 지방의 정도는 다음과 같았다. 세 식이요법의 효과에 차이가 있는가?

식이요법1	식이요법2	식이요법3
120(22)	96(17)	98(19)
93(12.5)	62(2)	92(11)
95(15)	84(8)	81(7)
96(17)	86(9)	93(12.5)
105(20)	69(3)	75(5)
96(17)	74(4)	61(1)
110(21)	78(6)	94(14)
		87(10)
표본수(n_j) 7	7	8
순위합(R_j) 124.5	49.0	79.5

비모수: 독립적인 k-표본

예시: Kruskal-Wallis test

$$H = \frac{12}{N(N+1)} \cdot \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(N+1) \qquad H^* = \frac{H}{1 - \sum_{i=1}^{l} q_i(q_i^2 - 1)/N(N^2-1)}$$

표본수(n_j) 7	7	8
순위합(R_j) 124.5	49.0	79.5

$$H = \frac{12}{22(22+1)}\left(\frac{124.5^2}{7} + \frac{49.0^2}{7} + \frac{79.5^2}{8}\right) - 3(22+1) = 10.38$$

$$H^* = \frac{10.38}{1-\{3(3^2-1)+2(2^2-1)\}/\{22(22^2-1)\}} = 10.41$$

$$df = k - 1 = 2$$

$$\chi^2_{0.05}(2) = 5.99 < 10.41$$

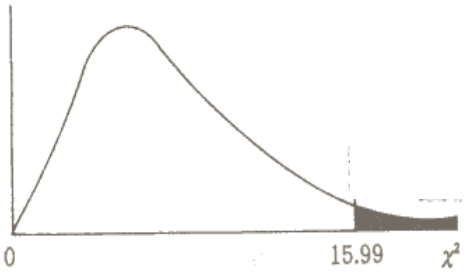
$$p < 0.05$$

세 식이요법의 효과에 차이가 없다는
귀무가설을 기각하므로
세 식이요법은 차이가 있다 → 사후검정 진행

비모수: 독립적인 k-표본

예시: Kruskal-Wallis test

6. Percentage Points of the χ^2 Distribution



ϕ \ P	.995	.99	.975	.95	.90	.75	.50	.25	.10	.05	.025	.01	.005	P / ϕ
1	0.000039	0.00016	0.00098	0.0039	0.0158	0.102	0.455	1.323	2.71	3.84	5.02	6.63	7.88	1
2	0.0100	0.0201	0.0506	0.103	0.211	0.575	1.386	2.77	4.61	5.99	7.38	9.21	10.60	2
3	0.0717	0.115	0.216	0.352	0.584	1.213	2.37	4.11	6.25	7.81	9.35	11.34	12.84	3
4	0.207	0.297	0.484	0.711	1.064	1.923	3.36	5.39	7.78	9.49	11.14	13.28	14.86	4
5	0.412	0.554	0.831	1.145	1.610	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75	5
6	0.676	0.872	1.237	1.635	2.20	3.45	5.35	7.84	10.64	12.59	14.45	16.81	18.55	6
7	0.989	1.239	1.690	2.17	2.83	4.25	6.35	9.04	12.02	14.07	16.01	18.48	20.3	7
8	1.344	1.646	2.18	2.73	3.49	5.07	7.34	10.22	13.36	15.51	17.53	20.1	22.0	8
9	1.735	2.09	2.70	3.33	4.17	5.90	8.34	11.39	14.68	16.92	19.02	21.7	23.6	9
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.55	15.99	18.31	20.5	23.2	25.2	10

비모수: 독립적인 k-표본

예시: Kruskal-Wallis test

- 분석에서 $p < 0.05$ 라면, 다른 모집단에서 관찰된 집단이 적어도 하나 이상 존재한다고 해석
- 사후 분석도 비모수 방법으로

  예. Bonferroni Correction

상관분석

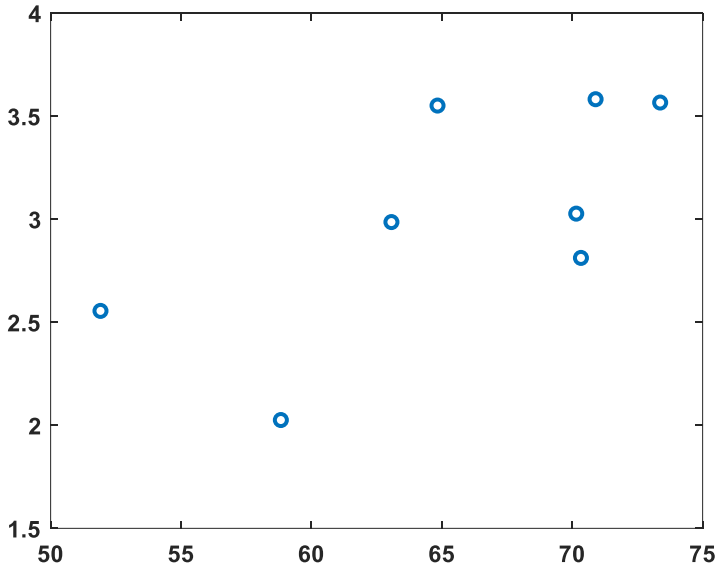
상관 & 회귀 분석 요약

- **상관 (Correlation)**
  - 두 변수 간 연관성이 있는가?

- **회귀 (Regression)**
  - 독립변수(independent variable)가 종속변수(dependent variable)를 잘 설명하고 예측하는가?

상관 분석

	X	Y
1	58	2.75
2	70	2.86
3	74	3.37
4	63.5	2.76
5	52	2.62
6	70.5	3.49
7	71	3.05
8	66	3.12



분산(Variance)과 공분산(Covariance)

- (두) 변수의 변동을 측정

Variance $$S^2 = \frac{\sum_{k=1}^{n}(x_k - \overline{x})^2}{n}$$

Covariance $$COV(x, y) = \frac{\sum_{k=1}^{n}(x_k - \overline{x})(y_k - \overline{y})}{n}$$

공분산의 해석

- X ↑ and Y ↑ / X ↓ and Y ↓ : cov (x, y) > 0
- X ↓ and Y ↑ / X ↑ and Y ↓ : cov (x, y) < 0
- No constant relationship: cov (x, y) = 0

$$COV(x, y) = \frac{\sum_{k=1}^{n}(x_k - \overline{x})(y_k - \overline{y})}{n}$$

상관 분석

공분산 계산

	x_k	y_k	$x_k - \overline{x}$	$y_k - \overline{y}$	$(x_k - \overline{x})(y_k - \overline{y})$
1	58	2.75	-6.60	-0.99	6.51
2	70	2.86	4.73	0.01	0.07
3	74	3.37	7.95	0.55	4.39
4	63.5	2.76	-2.36	-0.03	0.06
5	52	2.62	-13.51	-0.46	6.19
6	70.5	3.49	5.47	0.57	3.11
7	71	3.05	4.91	-0.20	-0.99
8	66	3.12	-0.59	0.54	-0.32

2.38 ?

상관 계수 (Correlation Coefficient)

$$-\infty \leq \boldsymbol{COV}(\boldsymbol{x}, \boldsymbol{y}) \leq \infty$$

- Covariance 자체는 해석에 어려움에 있음
  - 해법: 측정치를 표준화
- Pearson's R: 표준편차들을 이용하여 Covariance 를 표준화

상관 계수 (Correlation Coefficient) 계산

$$S^2 = \frac{\sum_{k=1}^{n}(x_k - \bar{x})^2}{n}$$

$$r = \frac{COV(x,y)}{\sqrt{var(x)var(y)}}$$

$$COV(x,y) = \frac{\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y})}{n}$$

$$-1 \le r \le 1$$

상관 분석

상관 계수 (Correlation Coefficient) 요약

- 두 연속변수 간의 선형적 연관성의 정도를 측정
- 양의 값 또는 음의 값을 가질 수 있음
- -1 에서 +1 사이의 값을 가짐
- 두 변수 간의 상관관계가 인과 관계를 의미하지는 않음

상관 계수의 검정

- 상관계수 값이 0과 유의하게 다른지 검정

$$H_0: r = 0 \qquad H_1: r \neq 0$$

상관 분석 조건

- 상관계수의 유의성은 상관계수의 크기와 표본의 크기에 의존함
- 검정의 타당성을 위해 각 관찰 값이 임의로 얻어져야 하며 적어도 하나의 변수는 정규분포를 따라야 함

$$t_0 = \frac{r_0\sqrt{n-2}}{\sqrt{1-r_0^2}}$$

상관 분석

상관 분석 예시

- 귀무가설 $H_0: r = 0$를 유의수준 1%에서 기각 ($p < 0.01$)

Correlations

		BODYFATB	BICEPS	ABDOMEN	HEIGHT	WEIGHT	AGE
BODYFATB	Pearson Correlation	1	.493**	.814**	.613**	-.089	.289**
	Sig. (2-tailed)	.	.000	.000	.000	.158	.000
	N	252	252	252	252	252	252
BICEPS	Pearson Correlation	.493**	1	.685**	.800**	.208**	-.041
	Sig. (2-tailed)	.000	.	.000	.000	.001	.515
	N	252	252	252	252	252	252
ABDOMEN	Pearson Correlation	.814**	.685**	1	.888**	.088	.230**
	Sig. (2-tailed)	.000	.000	.	.000	.165	.000
	N	252	252	252	252	252	252
HEIGHT	Pearson Correlation	.613**	.800**	.888**	1	.308**	-.013
	Sig. (2-tailed)	.000	.000	.000	.	.000	.840
	N	252	252	252	252	252	252
WEIGHT	Pearson Correlation	-.089	.208**	.088	.308**	1	-.172**
	Sig. (2-tailed)	.158	.001	.165	.000	.	.006
	N	252	252	252	252	252	252
AGE	Pearson Correlation	.289**	-.041	.230**	-.013	-.172**	1
	Sig. (2-tailed)	.000	.515	.000	.840	.006	.
	N	252	252	252	252	252	252

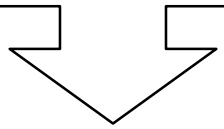
**. Correlation is significant at the 0.01 level (2-tailed).

상관 분석

비모수적 방법

- 자료가 순위성(ordinal) 자료이거나, 정규성을 만족하지 않으면,
- 순위 상관 계수인 Pearson's 상관계수 대신 Spearman's rank correlation 을 사용할 수 있음

자료에 대한 기본 가정을 만족하지 않으면
모수적 검정법을 사용은 적절하지 않음



비모수적 검정법 사용

상관 분석

예시. 비모수적 방법

- 스피어만의 순위상관계수는 피어슨의 상관계수와 동일한 방법으로 계산되지만 실측값 대신 순위를 이용하여 계산
- 범위는 -1에서 +1 사이의 값을 가지며 해석은 동일
- 자료가 정규분포를 따르지 않아도 됨 (비모수적 통계량)

예시	X	Y	Rank X	Rank Y
1	86	2	1	1
2	97	20	2	6
3	99	28	3	8
4	100	27	4	7
5	101	50	5	10
6	103	29	6	9
7	106	7	7	3
8	110	17	8	5
9	112	6	9	2
10	113	12	10	4

$$r = \frac{Cov_{XY}}{\sqrt{Var_X \; Var_Y}}$$

회귀분석

회귀분석의 예시

- 환자의 나이로부터 혈압을 예측할 수 있는가?
- 복부둘레 측정치로부터 체지방 수치를 예측할 수 있는가?

단순 회귀 분석

- 두 연속변수 간의 관계를 설명
- 두 연속변수의 관계를 가장 잘 설명할 수 있는 수식(선형적으로 표현된)을 제공
- 다른 변수의 정보로부터 결과변수 값의 예측 가능

변수 정의

- **종속변수(dependent variable):**
  - 예측하고자하는 변수 (관심있는 특정 결과)

- **독립변수(independent variable) 또는 설명변수:**
  - 특정 결과를 예측 가능하게 해주는 변수

선형 방정식과 회귀 계수

- 직선의 기울기를 나타내는 b는 회귀계수
- b는 상관계수와 같은 부호를 가짐
- x와 y 간에 상관성이 없으면 회귀계수 b의 값은 0
- 절편인 a는 x 값이 0일 때 y의 예측값

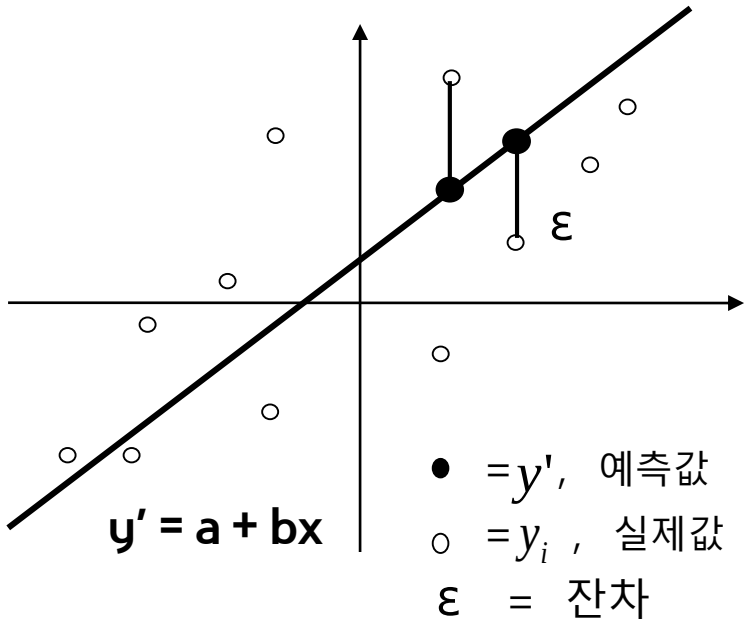
$$y' = a + b\,x$$

y'	예측값 (종속 변수)
a	절편
b	회귀선의 기울기
x	독립(설명)변수

최소 제곱법

- 관측치들과 거리를 최소로 하며 관통하는 최적의 직선을 찾아내는 것

$$\frac{\sum_{i=1}^{n}(y_i - y')^2}{n} \rightarrow \min$$



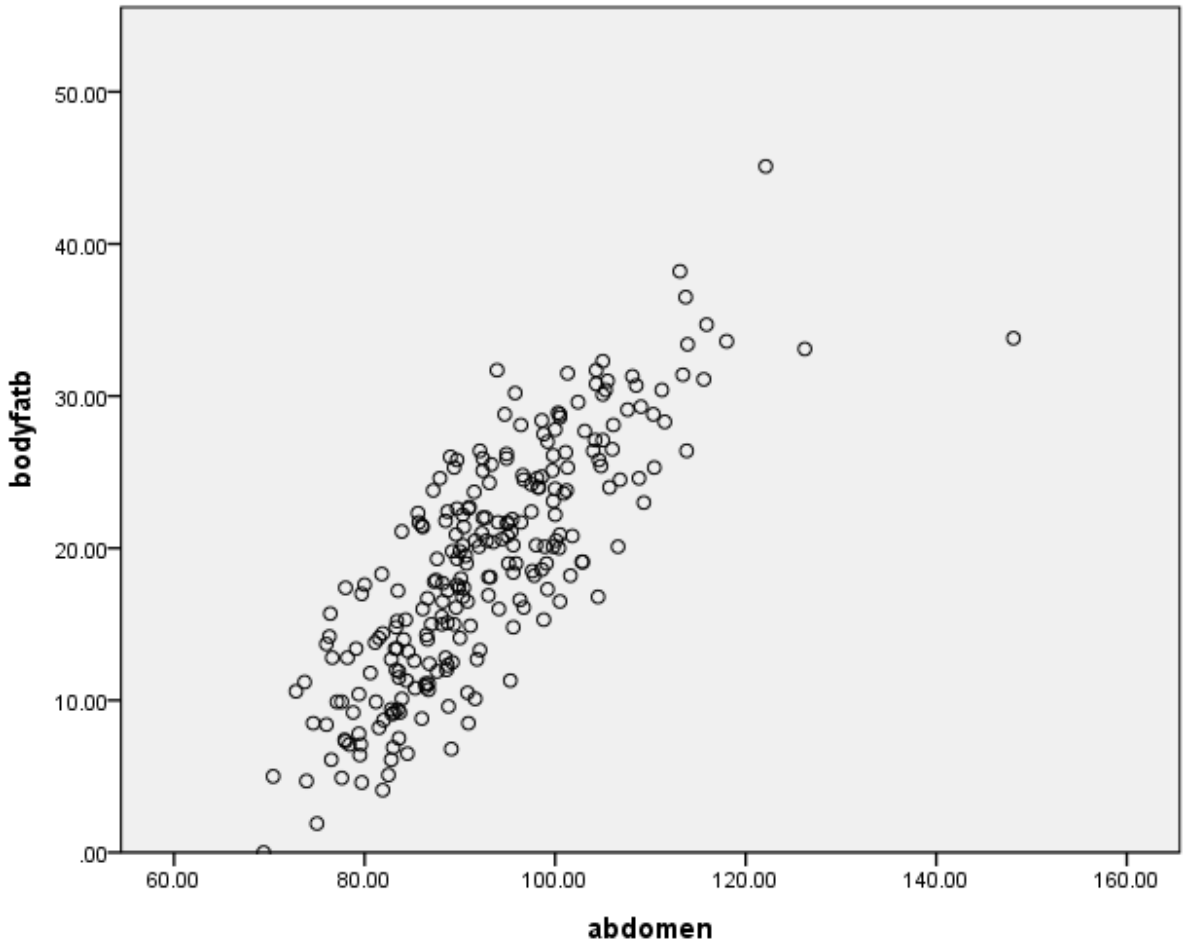
예시: 회귀 분석

- 신체 치수 등 다른 정보를 통해 체지방 정보를 추정?
- 자료: 252명을 대상으로 체지방과 복부둘레 및 신체관련 치수를 측정
- 가설: 복부둘레 치수를 통해 체지방 추정 가능?

H_0: 체지방과 복부둘레 간에 선형적 연관성이 없다.
H_1: 체지방과 복부둘레 간에 선형적 연관성이 있다.

H_0: 복부둘레는 체지방의 변동을 설명할 수 없다.
H_1: 복부둘레는 체지방의 변동의 일부를 설명할 수 있다.

예시: 회귀 분석



산점도를 통해 체지방(Y) 과 복부둘레(X) 간에 강한 양의 연관성이 있음을 파악

회귀 분석

예시: 회귀 분석 (결과)

상관계수 설명력

모형 요약[b]

모형	R	R 제곱	수정된 R 제곱	추정값의 표준오차
1	.814[a]	.662	.661	4.51439

a. 예측값: (상수), abdomen

b. 종속변수: bodyfatb

회귀 분석

예시: 회귀 분석 (해석)

- 적합된 회귀선에 의해 체지방의 변동 부분이 통계적으로 유의하게 설명이 됨을 알 수 있음($p < 0.001$)

분산분석[b]

모형		제곱합	자유도	평균 제곱	F	유의확률
1	회귀 모형	9984.086	1	9984.086	489.903	.000[a]
	잔차	5094.931	250	20.380		
	합계	15079.017	251			

a. 예측값: (상수), abdomen

b. 종속변수: bodyfatb

회귀 분석

예시: 회귀 분석 (회귀 방정식)

계수[a]

모형		비표준화 계수		표준화 계수		
		B	표준오차	베타	t	유의확률
1	(상수)	-35.197	2.462		-14.294	.000
	abdomen	.585	.026	.814	22.134	.000

a. 종속변수: bodyfatb

예측 체지방량 = 상수 + B x 복부둘레

예측 체지방량 = -35.197 + 0.585 복부둘레

예시: 회귀 분석 (예측)

- 예측에 선형 회귀식을 활용
- 회귀식으로부터 독립변수 x의 특정 값에 대해 종속변수 y값을 추정할 수 있음
- 예측 체지방량 = -35.197 + (0.585 x 복부둘레)
- 복부둘레가 100cm인 사람의 체지방량은?
- 예측 체지방량 = -35.197 + (0.585 x 100)
  = -35.197 + 58.5
  = 23.3%

- 내가 만든 모형이 잘 만들어졌을까?
- + 얼마나 설명할 수 있을까?

Review: 분산분석

기계 1	기계 2	기계 3	기계 4	기계 5	기계 6
17.5	16.4	20.3	14.6	17.5	18.3
16.9	19.2	15.7	16.7	19.2	16.2
15.8	17.7	17.8	20.8	16.5	17.5
18.6	15.4	18.9	18.9	20.5	20.1

$\overline{y_{1.}} = 17.2,\ \overline{y_{2.}} = 17.175,\ \overline{y_{3.}} = 18.175,\ \overline{y_{4.}} = 17.75,\ \overline{y_{5.}} = 18.425,\ \overline{y_{6.}} = 18.025,\ \overline{y} \fallingdotseq 17.792$

앞 식의 양변을 제곱하고 모든 i, j에 대하여 합하면 다음 결과를 얻는다.

집단간 편차의 제곱 합 **집단내 편차의 제곱 합**

$$\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\overline{y})^2}_{\text{총제곱합}} = \underbrace{\sum_{i=1}^{k} n_i(\overline{y_{i.}}-\overline{y})^2}_{\text{처리제곱합}} + \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\overline{y_{i.}})^2}_{\text{오차제곱합}}$$

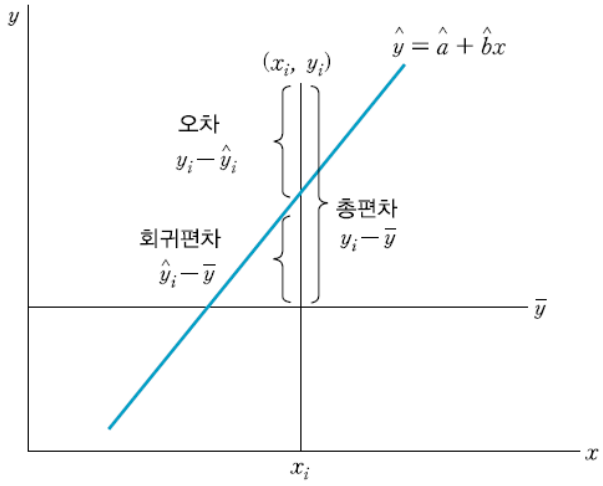
SSt **SSE**

요인	제곱합	자유도	평균제곱	검정통계량 F
처리	5.338	5	1.068 **MSt**	0.307
오차	62.64	18	3.48 **MSE**	–
합계	67.978	23	–	–

$$F = \frac{SSt/(k-1)}{SSE/(n-k)} = \frac{MSt}{MSE} \sim \mathrm{F}(k-1, n-k), \quad n = \sum_{i=1}^{k} n_i$$

회귀 분석

추정된 회귀직선의 정확도



[그림 10-4] 총편차의 분해

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

SST
Sum of squared Total

SSE
Sum of squared Error

SSR
Sum of squared Residual

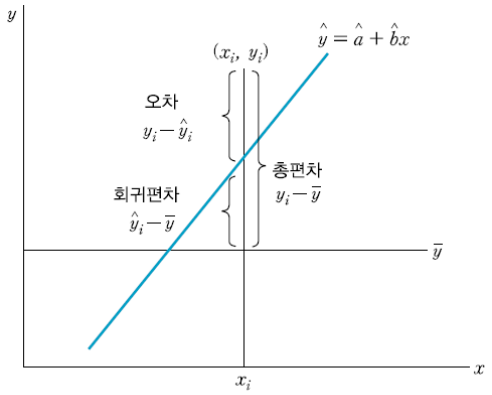
회귀 분석

추정된 회귀직선의 정확도

- **오차 (잔차) 제곱합 & 평균**

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

SST **SSE** **SSR**



[그림 10-4] 총편차의 분해

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

SST **SSR**

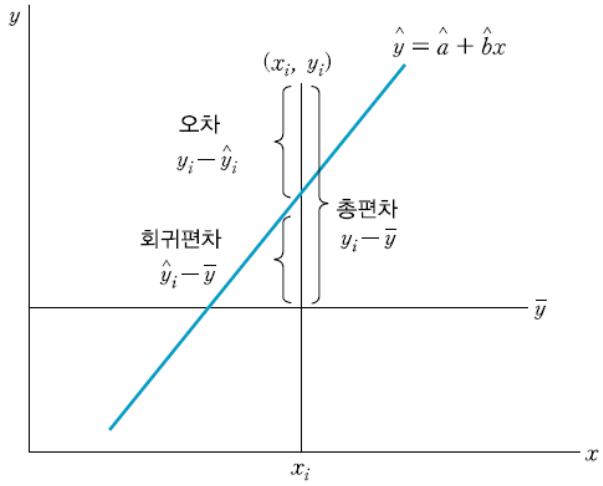
오차항의 분산 σ^2의 추정량 S^2은 오차제곱합을 이용하여 다음과 같이 구할 수 있다. 이 추정량을 **평균제곱오차**(mean square error)라고 하며, 다음과 같이 MSE로 표기한다.

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

회귀 분석

추정된 회귀직선의 정확도

- **회귀 제곱합 & 평균**



[그림 10-4] 총편차의 분해

$$\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\text{회귀제곱합}} \quad /(k\text{-}1)$$

회귀 분석

추정된 회귀직선의 정확도

- 예. X, Y (k = 2개 - 1) 자료수 75개 (75 - 2)
- F-table의 값과 비교
- $p < 0.05$라면
- 적합된 회귀선에 의해 Y의 변동 부분이 통계적으로 유의하게 설명이 됨!

모형		제곱합	자유도	평균제곱	F
1	회귀	4303.680	1	4303.680	44.693
	잔차	7029.499	73	96.295	
	전체	11333.179	74		

회귀 분석

추정된 회귀직선의 정확도

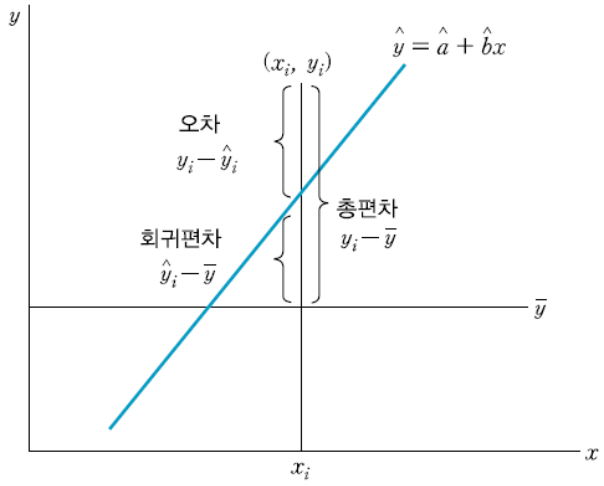
$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{\text{총제곱합}} = \underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{\text{오차제곱합}} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{\text{회귀제곱합}}$$

SST SSE SSR

- 결정계수

$$SSE = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 = \underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{\text{SST}} - \underbrace{\frac{\left[\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right]^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}_{\text{SSR}}$$

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



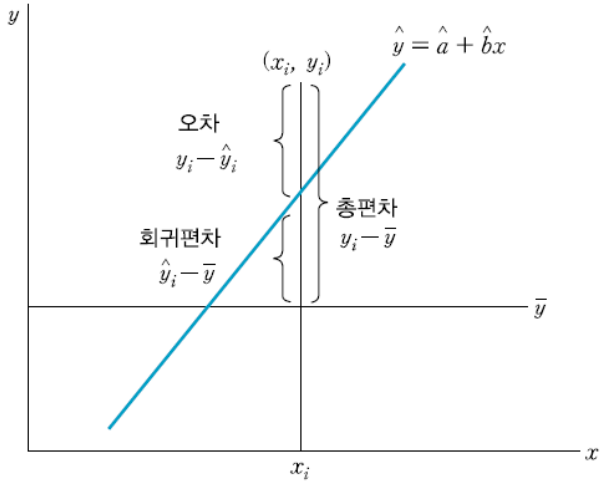
[그림 10-4] 총편차의 분해

$$\hat{\rho} = r = \frac{c_{xy}}{S_x S_y} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

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

추정된 회귀직선의 정확도

- **결정계수**
  - 결정계수는 표본상관계수의 제곱과 같으며 범위는 0과 1 사이이다. 따라서 결정계수의 값이 1에 가까울수록 추정회귀직선 주위에 자료들이 밀집되어 있으므로 자료들을 잘 대표하고 있다고 할 수 있다.
  - **통계에서 "설명력"이라고 함**



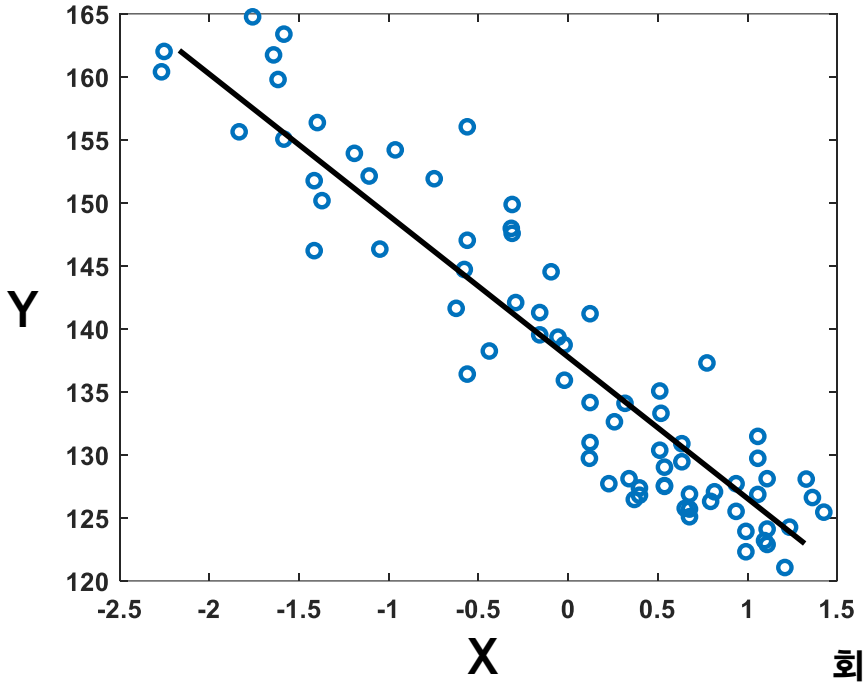
[그림 10-4] 총편차의 분해

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

SST SSE SSR

회귀 분석

추정된 회귀직선의 정확도



R=-0.934 (상관계수, $p < 0.001$)
R^2=0.872 (설명력)

X는 Y를 87.2% 설명할 수 있다

y = ax + b
a: -11.881
b: 137.663

Y = -11.881x + 137.663 (p<0.001)

회귀식에 의해 Y의 변동 부분이 통계적으로 유의하게 설명 됨 ($p < 0.001$)

중회귀 분석 요약

- 중회귀모형
- 변수의 선택
- 설명변수의 보정
- 다중공선성의 문제

중회귀 모형

- 하나의 종속변수를 설명하기 위한 두 개 이상의 설명(독립)변수에 관심을 가지는 경우
  - 관련요인 평가: 독립변수들을 한 모형에 포함하여 서로를 보정한 상태에서 종속변수에 유의하게 영향을 미치는 요인 평가
  - 예측모형의 구축: 유의한 모형을 이용 종속변수 값을 예측

- 많은 변수를 모형에 포함시킬 수 있으나 변수의 개수를 적게하여 간단한 모형을 선택하는 것이 권장

예시: 중회귀 모형

- **종속변수**
  - 체지방률(body fat)

- **설명변수**
  - age, weight, height, hip, biceps, neck, knee, forearm, abdomen circumference measurements

예시: 중회귀 모형

$$Y = a + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_kx_k$$

Y: 종속변수

a: 절편 (상수)

b_i: x_i 에 대한 회귀계수

x_i: 설명 (독립) 변수 ($i = 1, 2, \cdots, k$)

다중 공선성 (multicollinearity)

- 설명변수들이 서로 연관되어 있는 경우 발생
- 하나의 설명변수가 다른 설명변수들과 선형적 함수로 설명이 될 때 이를 다중공선성(multicollinearity) 라고 함
- b_i의 추정치에 수리적 오차가 있을 수 있으며, 표준오차 값이 커서 유의한 설명변수가 통계적으로 유의하지 않게 됨

$$Y = a + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_kx_k$$

다중 공선성 (multicollinearity)의 예

- 키와 몸무게에 대한 상관계수 계산: r=0.92
- 영양실조에 대한 지표 Pemax를 키와 몸무게에 이용하여 예측하고자 하며, 이 두 변수는 Pemax와 연관성이 높음
- Pemax = a + b1 height + b2 weight
- 모형에서 몸무게와 키에 대한 계수가 모두 유의하지 않음: $p > 0.05$
  - 두 변수의 연관성이 높아서 각각의 변수와 Pemax와의 연관성을 확인할 수 없음

분산팽창계수(VIF: variance inflation factor)

- VIF가 10이상이면 다중공선성 고려
- 변수 선택과정에서 상관계수가 높은 두 변수 중 하나만 선택
- 더 많은 데이터 수집

$$VIF_j = \frac{1}{1 - R_j^2}$$

회귀 분석

설명변수 선택법

- **회귀 모형에 포함되는 자동 변수 선택 방법에는 다음과 같은 것들이 있음**
  - 전진 선택법 (Forward selection)
  - 후진 제거법 (Backwards elimination)
  - 단계별 선택법 (Stepwise selection)
- **변수 선택 방법에 따라서 같은 자료에 대해 다른 회귀 모형이 수립될 수 있음**

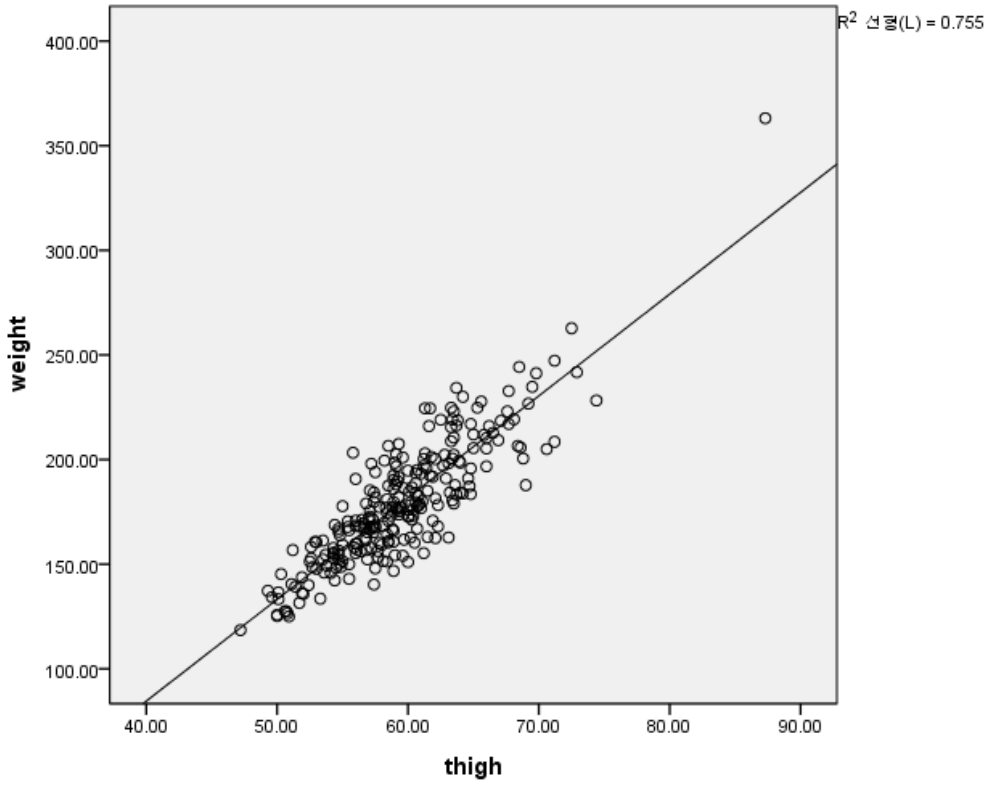
모형에 포함되는 유의수준(예.0.15)

Step1: x4 (p=0.01)
Step2: x4(0.01), x10(0.03)
Step3: x4(0.01), x10(0.2), x2(0.12)
Step4: x4, x2
Step5: x4, x2, x5
…

회귀 분석

예시: 중회귀 분석

- 설명변수 허벅지둘레와 몸무게 간에 선형관계가 발견됨
- 그러나 이로 인해 각 변수의 유의성 검정에는 영향 없었음



회귀 분석

예시: 중회귀 분석 (중회귀 방정식)

- 단계별 선택법에 의한 설명변수 선정

계수[a]

모형		비표준화 계수		표준화 계수		
		B	표준오차	베타	t	유의확률
3	(상수)	-48.039	3.987		-12.049	.000
	abdomen	.917	.052	1.276	17.578	.000
	weight	-.170	.025	-.643	-6.834	.000
	thigh	.209	.100	.142	2.100	.037

a. 종속변수: bodyfatb

$$\hat{y} = -48.04 + 0.917\,\text{ABDOMEN} - 0.17\,\text{WEIGHT} + 0.21\,\text{THIGH}$$

회귀 분석

예시: 중회귀 분석 (예측)

- 복부둘레 100cm와 168lbs의 체중, 허벅지 둘레 57cm 로 예측된 체지방률은 다음과 같음

$$\hat{y} = -48.04 + 0.917\ \text{ABDOMEN} - 0.17\ \text{WEIGHT} + 0.21\ \text{THIGH}$$

$$\hat{y} = -48.04 + 0.917 \times 100 - 0.17 \times 168 + 0.21 \times 57$$

$$= 27.1\%$$

Thank you.