SAS 기초통계분석

Contents

T 검정

모평균에 대한 검정
(Test for population Means)

1. 단일 모집단의 평균에 대한 검정

2. 독립표본에 의한 모평균 차의 검정

3. 짝표본에 의한 모평균 차의 검정

가설검정(hypothesis testing)

; 모집단의 특성인 모수에 대하여 가설을 세우고 표본에 근거하여 그 가설을 기각할 것인지 채택할 것인지를 결정하는 통계적 방법

가설검정 절차

1. 가설설정
2. 검정통계량 값 또는 유의확률의 계산
3. 유의수준 결정
4. 가설의 기각여부 결정

귀무가설(null hypothesis)

영가설(零假說). 통계학에서 처음부터 버릴 것을 예상하는 가설이다. 차이가 없거나 의미 있는 차이가 없는 경우의 가설이며 이것이 맞거나 맞지 않다는 통계학적 증거를 통해 증명하려는 가설이다

대립가설(alternative hypothesis)

귀무가설과 대립되거나 부정하는 가설.

일반적으로 대립가설은 분석자가 주장하고자 하는 내용이 포함되도록 설정.

유의수준

귀무가설이 참임에도 불구하고 귀무가설을 기각함으로써 발생되는 오류를 범할 확률의 최대허용한계를 유의수준이라 한다.

독립 이표본 t-검정

: 비교하고자 하는 그룹이 두 개인 경우

- 연구목적
  - 관심이 있는 변수의 모평균이 독립적인 두 집단 간에 서로 차이가 있는가?

- 가정
  - 해당 변수는 각 집단 별로 정규분포를 따른다 (정규성).
  - 두 집단의 분산은 서로 동일하다 (등분산성).
  - 표본의 수는 정규성 가정 및 등분산성 가정을 검토할 수 있을 정도로 충분히 크다.

독립 이표본 t-검정

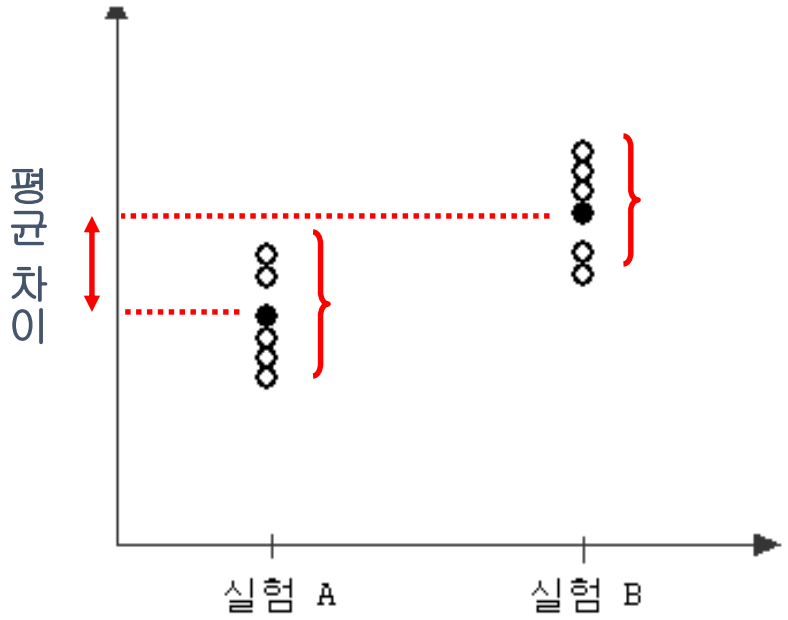
: 비교하고자 하는 그룹이 두 개인 경우

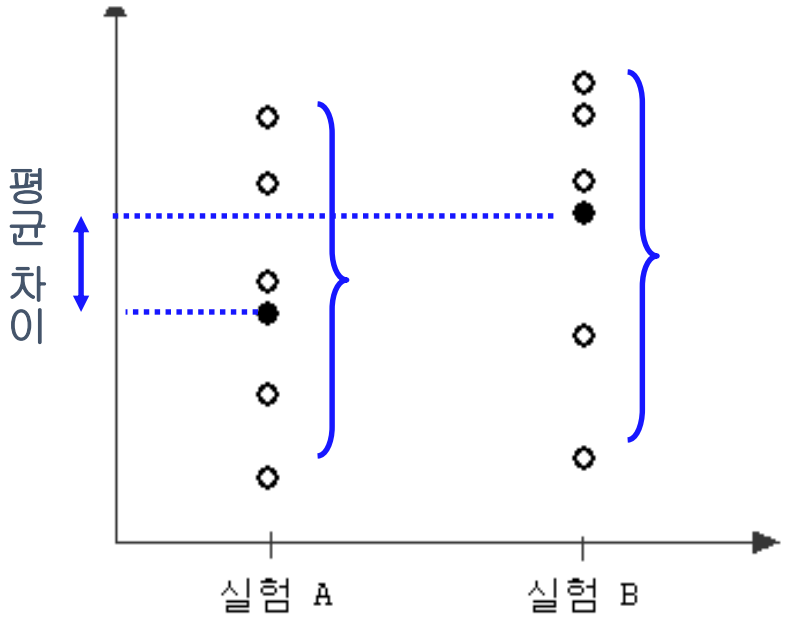
- 이론적 배경
  - 두 집단의 모평균이 동일하다면 차이는 0일 것이다.
  - 두 집단의 모평균이 동일하다는 가정 하에서 표본평균들의 차이를 계산한다.
  - 표본평균들의 차이로부터 계산된 검정통계량은 t-분포를 따른다.

독립 이표본 t-검정

: 비교하고자 하는 그룹이 두 개인 경우

- 평균차이가 통계적으로 유의한가?





독립 이표본 *t*-검정

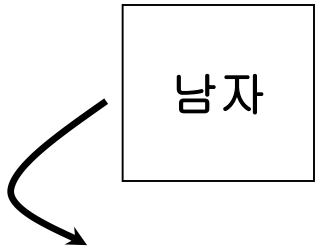
: 비교하고자 하는 그룹이 두 개인 경우

- 검정절차
  - 해당연구에 관한 귀무가설과 대립가설의 설정
    - 두 집단의 모평균은 동일하다 (귀무가설).
    - 두 집단의 모평균은 동일하지 않다 (대립가설).
  - 귀무가설 하에서의 검정통계량 값을 계산
    - 두 집단의 해당변수에 대한 분산이 동일한가?
    - 검정통계량 = 표본평균의 차이 / 표준오차
  - 검정통계량으로부터 계산된 *P*-값을 유의수준과 비교
  - 최종적인 의사결정 수립

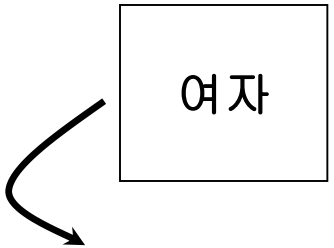
독립 이표본 *t*-검정

: 비교하고자 하는 그룹이 두 개인 경우

예) 사회현상에 대한 인식 정도의 성별 비교 (차이)



$$X_1, X_2, \cdots, X_n \sim N(\mu_1, \sigma^2)$$



$$Y_1, Y_2, \cdots, Y_m \sim N(\mu_2, \sigma^2)$$

$H_0: \mu_1 = \mu_2$ vs $H_1: \mu_1 \neq \mu_2$ (두 집단의 평균에 차이가 있는가 ?)

$$t = \frac{\left(\bar{X} - \bar{Y}\right)}{\sqrt{s_p^2/n + s_p^2/m}} \sim t(n+m-2)$$

두 모집단의 분산이 동일;

$$s_p^2 = \frac{(n-1)s_1^2 + (m-1)s_2^2}{n+m-2}$$

통합분산추정량(pooled variance estimator)

독립 이표본 *t*-검정

: 비교하고자 하는 그룹이 두 개인 경우

- *P*-값
  - 귀무가설 하에서 계산된 검정통계량보다 더 극단적인 값이 나올 확률
  - 만약 P-값이 0.001이라면
    - 귀무가설 하에서는 도저히 일어나기 힘든 차이가 발생
    - 귀무가설을 기각한다. (대립가설을 채택한다.)
  - 만약 P-값이 0.351이라면
    - 귀무가설 하에서 충분히 일어날 수 있는 차이가 발생
    - 귀무가설을 기각할 수 없다.
  - *P*-값=0.06

독립 이표본 t-검정

: 비교하고자 하는 그룹이 두 개인 경우

- 유의수준(alpha) 5%의 의미

가설검정 절차에서 5%의 유의수준을 선택한다면, 그것은 실제 채택하여야 하는데도 불구하고 우리가 그것을 기각할 경우는 100번 중 5번 정도임을 의미한다. 즉, 우리가 올바른 의사결정을 할 것을 약 95% 확신한다는 것이다. 이러한 경우에, 우리는 그 가설은 유의수준 0.05에서 기각되었다라고 말하는데, 이것은 곧 우리가 잘못된 의사결정을 내릴 확률이 0.05라는 것을 의미한다.

- 합리적인 의사결정
  - P-값 > 유의수준 : 주어진 유의수준 하에서 귀무가설을 기각할 만한 충분한 증거가 없으므로 귀무가설을 기각할 수 없다.
  - P-값 < 유의수준 : 주어진 유의수준 하에서 귀무가설을 기각할 만한 충분한 증거가 있으므로 귀무가설을 기각하고 대립가설을 채택한다.

SAS PROC TTEST 프로시저

- 단일모집단의 모평균 검정 1

```
PROC TTEST DATA=sas파일명 H0=모평균값;
   VAR   변수명;
RUN;
```

- 단일모집단의 모평균 검정 2

```
PROC UNIVARIATE DATA=sas파일명 MU0=모평균 값  ALPHA= 유의수준;
 VAR 변수명;
RUN;
```

- 독립표본의 모평균 차의 검정

```
PROC TTEST DATA=sas파일명;
 CLASS 변수명;
 VAR 변수명;
RUN;
```

- 짝표본의 모평균 차의 검정

```
PROC TTEST DATA=sas파일명;
  PAIRED 변수명1*변수명2;
RUN;
```

일표본 t 검정 예제1

임의 추출된 고대 주변 하숙집 15곳의 월세(단위 10,000원)가 다음과 같다.

35, 45, 40, 37, 38, 42, 44, 42, 38, 40, 36, 44, 39, 41, 36

이 자료를 이용하여 고대 주변의 월세의 평균이 38만원이라는 귀무가설을 검정하여 보자.

1) 가설(귀무가설, 대립가설)을 세운다. $H_0 : \mu = 38$ vs $H_1 : \mu \neq 38$

2) 유의수준 α 를 정한다. $\alpha = 0.05$

3) 검정통계량을 결정. 검정통계량: $$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$$

4) 관측된 자료에 대한 p-값을 계산한다.

$\bar{X}$ =39.8, s=3.17, 관측된 t 값= 2.20, p-값=P(|T|>2.20)=0.045

5) p-값=0.045 < α=0.05 ➔ 귀무가설 기각

결론 : 즉, 고대 주변의 월세의 평균은 38만원이 아니라고 할 수 있다.

일표본 t 검정 예제1 SAS

```
TITLE "One sample t-test for rent around KU";
DATA kurent;
 INPUT rent @@;
 CARDS;
35 45 40 37 38 42 44 42 38 40 36 44 39 41 36
;
RUN;
PROC TTEST DATA = kurent H0=38;
 VAR rent;
RUN;
PROC UNIVARIATE DATA = kurent MU0=38;
 VAR rent;
RUN;
```

일표본 t 검정 예제1 SAS 결과

One sample t-test for rent around KU

The TTEST Procedure

Statistics										
Variable	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper CL Std Dev	Std Err	Minimum	Maximum
rent	15	38.046	39.8	41.554	2.3185	3.1668	4.9943	0.818	35	45

T-Tests			
Variable	DF	t Value	Pr > |t|
rent	14	2.20	0.0450

PROC UNIVARIATE RESULTS

위치모수 검정: Mu0=38				
검정	통계량		p 값	
스튜던트의 t	t	2.201398	Pr > |t|	0.0450
부호	M	2.5	Pr >= |M|	0.2668
부호 순위	S	27.5	Pr >= |S|	0.0549

일표본 t 검정 예제2

수학시험에서 임의의 8명의 학생의 성적이 다음과 같을 때,
60, 62, 67, 69, 70, 72, 75, 80
평균이 65보다 크다고 할 수 있는가? (α=0.05)

1) 가설(귀무가설, 대립가설)을 세운다. $H_0 : \mu = 65$ vs $H_1 : \mu > 65$

2) 유의수준 α 를 정한다. α=0.05

3) 검정통계량을 결정. 검정통계량: $t = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$

4) 관측된 자료에 대한 p-값을 계산한다.
$\bar{X}$ =69.375, s=6.5452, 관측된 t 값= 1.89, p-값=P(|T|>1.89)=0.1006

5) p-값=0.1006 /2=0.0503> α=0.05 ➔ 귀무가설 기각 할 수 없음

결론 : 즉, 평균이 65보다 크다고 할 수 없다.

일표본 t 검정 예제2 SAS

```
TITLE "One sample t-test for math score";

data math;
  input score @@;
  testscore=score-65;
cards;
60   62   67   69   70   72   75   80
;
run;

proc ttest data=math H0=65;
var score;
run;

proc ttest data=math;
var testscore;
run;
```

일표본 t 검정 예제2 SAS 결과

One sample t-test for rent around KU

The TTEST Procedure

Statistics										
Variable	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper CL Std Dev	Std Err	Minimum	Maximum
score	8	63.903	69.375	74.847	4.3275	6.5452	13.321	2.3141	60	80

T-Tests			
Variable	DF	t Value	Pr > |t|
score	7	1.89	0.1006

일표본 t 검정 예제3

적포도주를 적당량 마시는 것이 심장마비를 막아줄 수 있다고 알려져 있다. 적포도주에 들어있는 폴리페놀이 그 역할을 하는 것으로 보인다.

적포도주를 적당량 마시는 것이 혈중 폴리페놀의 농도를 올리는 지 확인하기 위해 9명의 랜덤추출된 건강한 남자들에게 2주간 적포도주를 매일 반 병씩 마시도록 하였다. 그들의 혈중 폴리페놀 농도를 연구 시작과 끝에 측정하였는데, 그 차이 (% change)가 다음과 같이 나타났다.

0.7 3.5 4 4.9 5.5 7 7.4 8.1 8.4

자료가 근사적으로 정규분포를 따르는가?

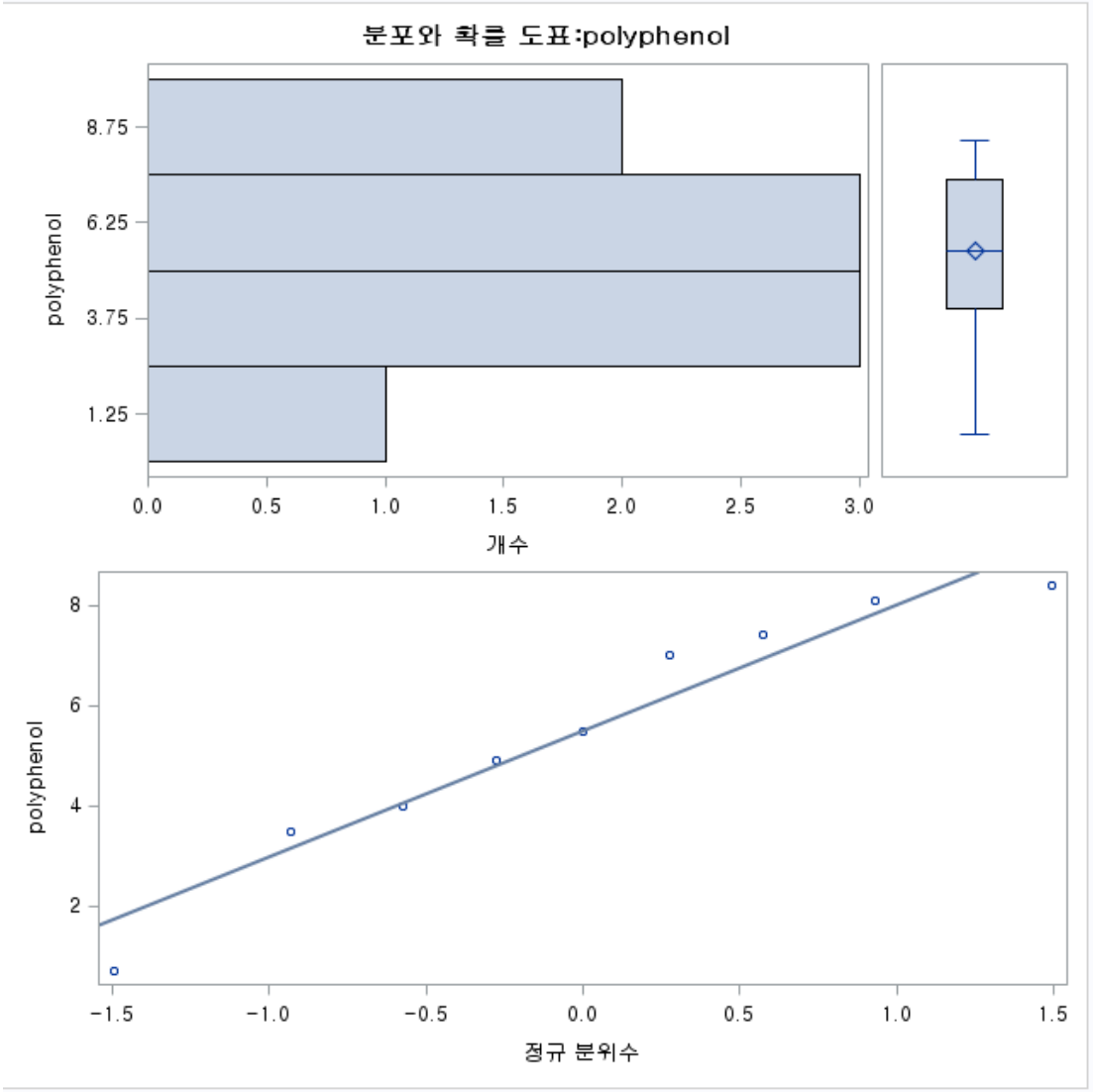
일표본 t 검정 예제3 SAS

```
* SAS program;

DATA wine;
        INPUT polyphenol @@;
CARDS;
0.7 3.5 4 4.9 5.5 7 7.4 8.1 8.4
;
RUN;

/* Histogram & Normal plot */
PROC UNIVARIATE DATA=wine NORMAL PLOT;
   VAR polyphenol;
RUN;
```

일표본 t 검정 예제3 SAS 결과



작은 값이 하나 있지만, 전체적으로 정규성을 가정해도 좋을 듯 하다.

이제, 퍼센트 차이의 평균에 대한 95% 신뢰구간을 구하시오.

일표본 t 검정 예제3 SAS

```
* SAS program;

PROC TTEST DATA=wine;
   VAR polyphenol;
RUN;
```

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
5.5000	3.5653	7.4347	2.5169	1.7001	4.8219

DF	t Value	Pr > |t|
8	6.56	0.0002

비율 검정 예제 SAS

동전을 4040번 던져서 1992번의 뒷면이 나왔다고 한다. 이 동전의 앞면과 뒷면이 나올 확률이 같은지 유의수준 5%에서 검정해보자.

```
DATA coin;
 INPUT headtail $ count;
 DATALINES;
 tail 1992
 head 2048
 ;
RUN;


PROC FREQ DATA=coin;
 WEIGHT count;
 TABLES headtail / BINOMIAL (p=0.5);
RUN;
```

비율 검정 예제 SAS 결과

FREQ 프로시저

headtail	빈도	백분율	누적 빈도	누적 백분율
head	2048	50.69	2048	50.69
tail	1992	49.31	4040	100.00

이항비	
headtail = head	
비율	0.5069
ASE	0.0079
95% 신뢰하한	0.4915
95% 신뢰상한	0.5223
정확 신뢰한계	
95% 신뢰하한	0.4914
95% 신뢰상한	0.5225

H0:P = 0.5 의 검정	
H0 하에서의 ASE	0.0079
Z	0.8810
단측 Pr > Z	0.1891
양측 Pr > |Z|	0.3783

표본 크기 = 4040

독립 이표본 t-검정 예제1 cars data

- 미국산 차와 일본산 차의 연비에 차이가 있는 지 검정하여 보자.

```
* SAS program;

TITLE "Two sample t-test to compare mpg
between American and Japanese cars";

PROC TTEST DATA=ex.cars;

 WHERE origin = 1 or origin = 3;

 CLASS origin;

 VAR mpg;

RUN;
```

독립 이표본 t-검정 예제1 SAS 결과

Two sample t-test to compare mpg between American and Japanese cars

The TTEST Procedure

Statistics

Variable	origin	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper CL Std Dev	Std Err	Minimum	Maximum
mpg	1	248	19.331	20.128	20.926	5.8606	6.3768	6.9935	0.4049	10	39
mpg	3	79	29.087	30.451	31.815	5.2661	6.09	7.222	0.6852	18	46.6
mpg	Diff (1-2)		-11.93	-10.32	-8.719	5.8592	6.3092	6.8346	0.8151		

T-Tests

Variable	Method	Variances	DF	t Value	Pr > |t|
mpg	Pooled	Equal	325	-12.66	<.0001
mpg	Satterthwaite	Unequal	137	-12.97	<.0001

Equality of Variances

Variable	Method	Num DF	Den DF	F Value	Pr > F
mpg	Folded F	247	78	1.10	0.6425

두 모비율 검정 예제 1 SAS

술을 마시는 대학생에 대해 폭음을 하는지 여부를 조사하였다. 이 자료를 바탕으로 하여 성별로 폭음의 비율이 같은지 다른지를 유의수준 5%에서 검정하고자 한다.

```
DATA binge;
 INPUT gender $ drinker $ count;
 DATALINES;
men   yes 1630
men   no  5550
women yes 1684
women no  8232
 ;
RUN;

PROC FREQ DATA=binge;
 TABLES gender*drinker
   / NOPERCENT NOROW NOCOL CHISQ;
 WEIGHT count;
RUN;
```

두 모비율 검정 예제1 SAS 결과

테이블:gender * drinker

gender	drinker		
	no	yes	합계
men	5550	1630	7180
women	8232	1684	9916
합계	13782	3314	17096

gender * drinker 테이블에 대한 통계량

통계량	자유도	값	Prob
카이제곱	1	87.1718	<.0001
우도비 카이제곱	1	86.3958	<.0001
연속성 수정 카이제곱	1	86.8062	<.0001
Mantel-Haenszel 카이제곱	1	87.1667	<.0001
파이 계수		-0.0714	
우발성 계수		0.0712	
크래머의 V		-0.0714	

Fisher의 정확 검정	
(1,1) 셀 빈도(F)	5550
하단측 p값 Pr <= F	<.0001
상단측 p값 Pr >= F	1.0000
테이블 확률 (P)	<.0001
양측 p값 Pr <= P	<.0001

표본 크기 = 17096

대응짝 t-검정 예제1

왼손으로 글을 쓰는 사람 10명에 대해 오른손과 왼손의 악력에 대한 측정을 한 결과가 다음과 같다. 이 자료에 의하면 왼손으로 글을 쓰는 사람들은 왼손의 악력이 오른손보다 강하다고 할 수 있는가? (D: 왼-오른, H_1: $\mu_D > 0$)

id	1	2	3	4	5	6	7	8	9	10
왼손	140	90	125	130	95	121	85	97	131	110
오른손	130	87	110	132	96	120	86	90	129	100

- t = 2.41, ~ t df=9
- p-값=0.0394/2=0.0197

⇨ **귀무가설 기각**

대응짝 t-검정 예제1 SAS

```
TITLE "Paired t-test example1";
DATA hand;
 INPUT id left right @@;
 CARDS;
 1 140 130 2 90 87  3 125 110 4 130 132 5 95 96
 6 121 120 7 85 86 8 97 90 9 131 129 10 110 100
 ;
RUN;

PROC TTEST DATA=hand;
 PAIRED left*right;
RUN;
```

대응짝 t-검정 예제1 SAS 결과

Paired t-test example1

The TTEST Procedure

Difference: left - right

N	Mean	Std Dev	Std Err	Minimum	Maximum
10	4.4000	5.7774	1.8270	-2.0000	15.0000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
4.4000	0.2671	8.5329	5.7774	3.9739	10.5472

DF	t Value	Pr > |t|
9	2.41	0.0394

대응짝 t-검정 예제2

Does lack of caffeine increase depression?

카페인 중독인 사람들을 대상으로 카페인이 많은 음식을 피하게 하고 대신 매일 알약을 먹게 하였다. 어떤 알약에는 카페인이 들어있고 어떤 알약에는 아무것도 들어있지 않다. (위약,placebo) 얼마간의 기간 동안 우울증세가 있는지 없는지 조사하였다.

- 각 사람에 대해 2종류의 count가 있지만, 실은 차이를 봐야함.
- 정규성을 가정.
- 11 "difference" → df = $n - 1 = 10$
- $\bar{x} = 7.36$; $s = 6.92$

Subject	Depression with Caffeine	Depression with Placebo	Placebo - Cafeine
1	5	16	11
2	5	23	18
3	4	5	1
4	3	7	4
5	8	14	6
6	5	24	19
7	0	6	6
8	0	3	3
9	2	15	13
10	11	12	1
11	1	0	-1

$$H_0: \mu_{\text{difference}} = 0 \; ; \; H_1: \mu_{\text{difference}} > 0$$

$$t = \frac{\bar{x} - 0}{s/\sqrt{n}} = \frac{7.36}{6.92/\sqrt{11}} = 3.53$$

p-값=0.0054/2=0.0027

카페인 기피가 우울증의 증가에 유의한 영향을 미친다고 할 수 있다.

대응짝 t-검정 예제2 SAS

```
TITLE "Paired t-test example2";
DATA caffeine;
 INPUT id caff placebo @@;
 CARDS;
 1 5 16 2 5 23 3 4 5 4 3 7 5 8 14 6 5 24
 7 0 6 8 0 3 9 2 15 10 11 12 11 1 0
 ;
RUN;

PROC TTEST DATA=caffeine;
 PAIRED caff*placebo;
RUN;
```

대응짝 t-검정 예제2 SAS 결과

Paired t-test example2

The TTEST Procedure

Difference: caff - placebo

N	Mean	Std Dev	Std Err	Minimum	Maximum
11	-7.3636	6.9177	2.0858	-19.0000	1.0000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
-7.3636	-12.0110	-2.7163	6.9177	4.8335	12.1401

DF	t Value	Pr > |t|
10	-3.53	0.0054

분산분석

두 집단 이상의 평균 간의 차이를 검증하는 것으로 t검정을 일반화한 분석 방법이다. 독립변수가 한 개일 때 일원분산 분석, 독립변수가 두 개 이상일 때 다원분산 분석이라고 한다.

분산 분석은 각 집단의 분산을 분석하나 실제로는 각 집단의 평균이 동일하다는 가설을 검정하는 것이 된다. 분산 분석은 각각의 모집단은 정규분포를 가정하고 있으며 분산은 모두 같은 값을 가진다고 가정하고, 귀무가설과 대립가설을 비교 검증하는 방법이다. 예를 들어, 가족유형에 따른 아동의 사회성에 차이가 있는지를 살펴볼 경우 가족유형을 양부모 가정, 편부편모 가정, 조손 가정, 다문화 가정 등으로 구분하여 분산 분석을 통해 집단별 아동의 사회성의 차이를 알 수 있다. 이때 t검정은 두 집단만을 비교할 수 있기 때문에 세 집단 이상을 비교할 경우에는 F 분산 분석을 사용해야 한다.

로널드 에일머 피셔 Ronald Aylmer Fisher

1890년~1962년 영국의 농학자/통계학자
Fisher는 1924년 F분포로 검정하는 분산분석 방법을 제안

"고교 수학교사로 일하던 그가 기회를 맞은 것은 '로담스테드 농업실험연구소'에 취직하면서부터이다. 비료회사가 운영하던 연구소는 퇴비더미 속에 있었지만 90년 동안 강수량과 온도, 비료의 종류와 수확량, 토양 상태에 관한 방대한 자료가 쌓여 있었다. 3년 후 그는 '작물 수확량 변동에 관한 연구'를 내놓으며 화려하게 학계에 복귀했다. 1962년 7월29일 사망하기까지 그는 내놓은 5편의 관련 논문 시리즈에서 제시된 '분산 분석'과 '자유도 개념' '신뢰확률 이론' 등은 20세기 통계학을 열었다."

기본 개념

- 특성치의 산포를 제곱합으로 나타내고 이 제곱합을 요인마다의 제곱합으로 분해하여 오차에 비해 특히 큰 영향을 주는 요인이 무엇인지 찾아내는 분석 방법

- 독립변수를 몇 개의 범주나 수준으로 나누고, 그에 따라 나뉜 집단 간의 평균차이 검정

이론적 배경

- 전체변동 (Total Variation)
  - 개별 반응값이 전체 평균으로부터 얼마나 퍼져 있나?
  - 전체변동 = 집단간 변동 + 집단내 변동

- 집단간 변동 (Between-group Variation)
  - 전체 변동 중 모형에 의해 설명되어지는 변동
  - 각 수준의 평균이 전체평균으로부터 얼마나 퍼져 있나?

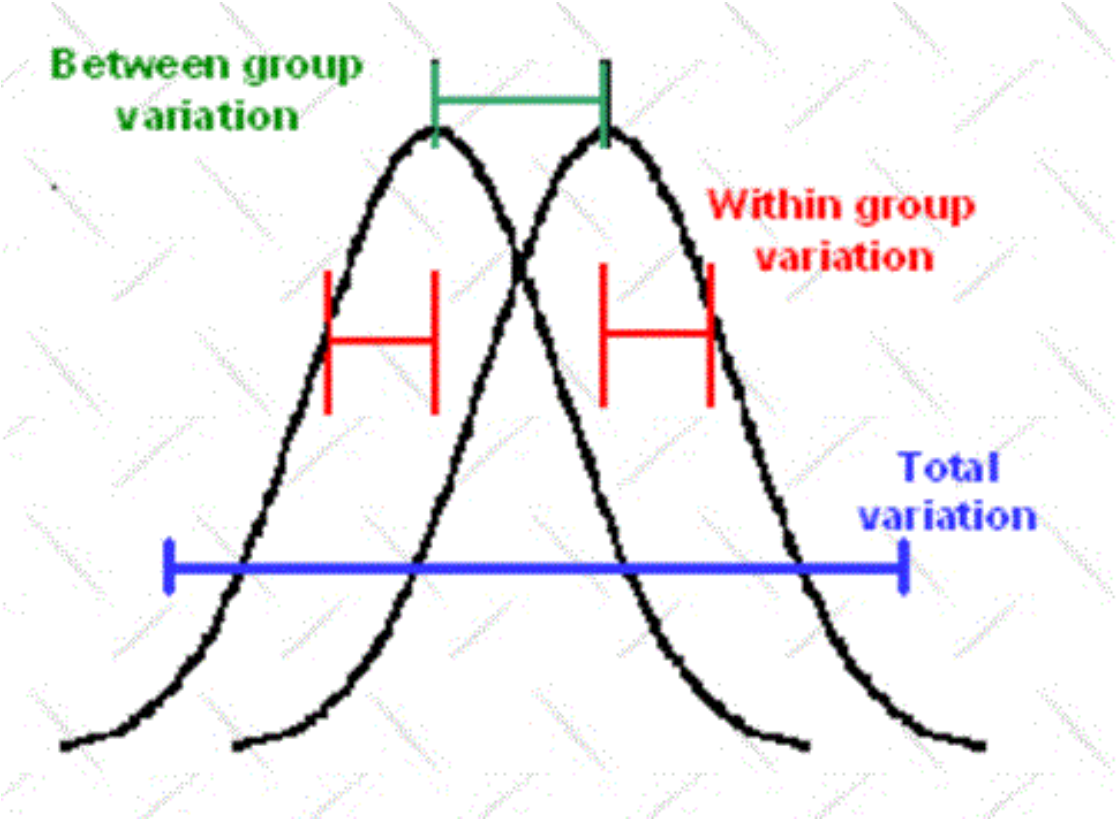
- 집단내 변동 (Within-group Variation)
  - 전체 변동 중 모형에 의해서 설명되어지지 않는 변동
  - 개별 반응값이 각 수준의 평균으로부터 얼마나 퍼져 있나?

이론적 배경 (계속)

- 집단간 변동 ↑ & 집단내 변동 ↓
  - 집단 간에 평균차이가 존재할 것이다.

- 집단간 변동 ↓ & 집단내 변동 ↑
  - 집단 간에 평균차이가 존재하지 않을 것이다.

- 검정통계량 = 집단간 평균변동 / 집단내 평균변동
  - 검정통계량 ↑ ⇔ 집단 간 평균차이가 유의할 가능성 ↑



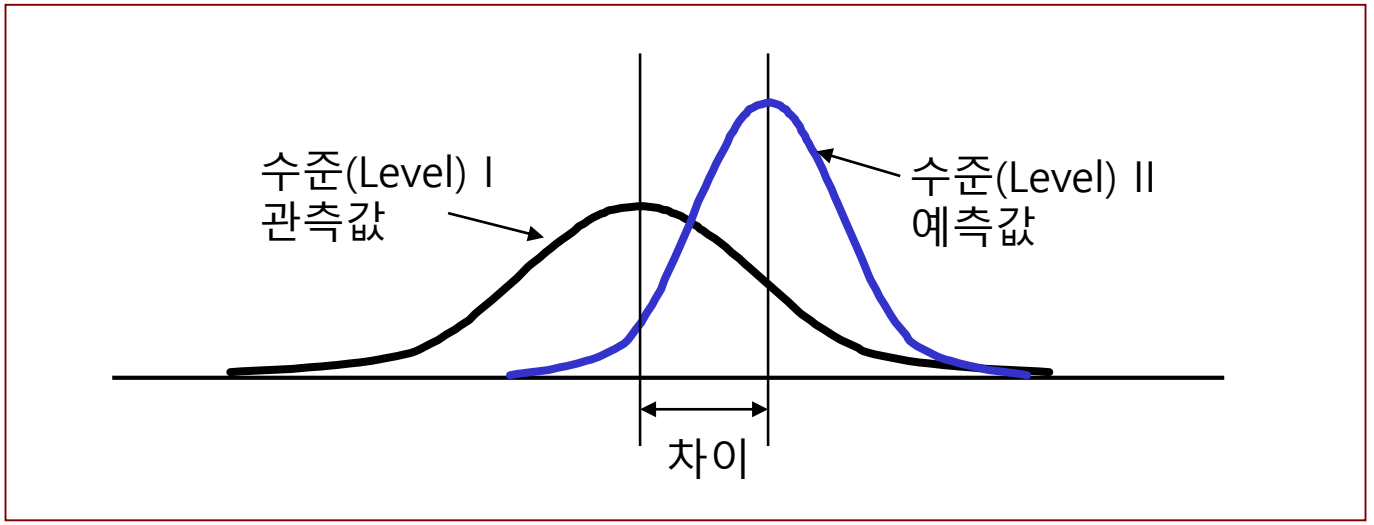
- total variation(총변동) = between group variation(집단간 변동) + within group variation(집단내 변동)

분산분석의 기본가정

- 분산분석의 기본가정

  - 각 집단에 해당되는 모집단의 분포가 정규분포이다.
  - 각 집단에 해당되는 모집단의 분산이 같다.(σ^2)
  - 각 모집단 내에서의 오차나 모집단 간의 오차는 서로 독립이다.

  ⇒ 분산이 다르다면 어떤 수준에서는 다른 인자 등의 영향을 받아 특성이 다른 모집단이 되므로, 모양이 다른 모집단의 분산이나 평균을 비교하는 것은 의미가 없음



수준 I 의 산포가 넓어서 수준 I 과 수준 II 의 평균의 차이를 구별 하기가 어려움

분산분석의 종류

- 분산분석의 종류

- One Way ANOVA(일원분산분석)
  - ✓ 요인이 하나인 경우, 두 개 이상의 모집단들의 평균이 서로 동일한 지 여부를 검정하고자 할 때 사용
  - ✓ 일원분산분석은 모집단의 수에 제한이 없으며, 각 표본의 수가 같지 않아도 됨
  - ✓ 단, 데이터의 형태 중 연속형 데이터인 경우 사용

  ⇒ 2 Sample t 검정과 동일한 목적으로 사용
  - ✓ 2-Sample t 검정이 두 모집단의 평균을 비교하는 데 반해, ANOVA분석은 여러 모집단의 평균 비교를 위해 사용

- Two-Way ANOVA(이원분산분석)
  - ✓ 요인이 두개인 경우 요인 2개가 결과치에 미치는 영향을 알아내기 위하여 사용됨.
  - ✓ 이를 통해 다양한 요인 수준 및 교호 작용에 의해 발생하는 영향을 살펴 볼 수 있음

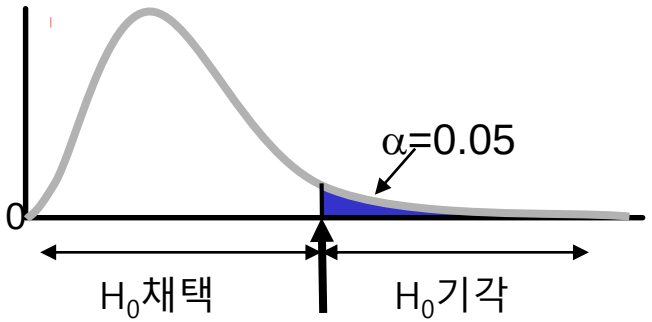
❑ 분산분석표

요인	제곱합	자유도	평균제곱합	F
처리	SSt	k-1	MSt=SSt/(k-1)	F=MSt/MSE
오차	SSE	n-k	MSE=SSE/(n-k)	
전체	SST	n-1		

❑ 가설의 검정

k개의 모집단의 평균이 같다면 SSt가 작아질 것이고 SSE가 커짐

반대의 경우 SSt가 커지고 SSE가 작아짐



H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$

H_1: 적어도 하나 이상 집단의 평균은 다르다.

$F > F_\alpha(k-1, n-k)$일때, 귀무가설 H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ 기각

일원분산분석

❑ 개별집단 평균차에 대한 사후검정

- 귀무가설이 기각되었을 경우 구체적으로 어떤 집단간에 유의한 차이가 있는지 알기 위해 사후 분석이 추가로 요구됨
- **그룹간 평균차이가 인정되는 경우, 어느 그룹간에 평균차이가 유의한 지를 알아보기 위한 검정**

- **If H_0 is rejected → H_1 : $\mu_1 \neq \mu_2 = \mu_3$, H_1 : $\mu_1 = \mu_2 \neq \mu_3$, or H_1 : $\mu_1 \neq \mu_2 \neq \mu_3$**
  → 판단하지 못함

- 사후 비교(post hoc comparison)이라고 함
- 다중비교(multiple comparison)를 수행
  - ✓ LSD, TUKEY, DUNCAN, SCHEFFE 등이 있음

❑ 다중비교 방법

1) LSD(최소유의차)

- 두 집단간 평균차이를 검정하는 독립표본 t-검정을 반복 실시

(비교별 오류율을 제어하므로 실험별 오류율이 커지게 됨)

→ 보수적인 방법으로 유의한 결과를 얻기 힘든 방법

2) Bonferroni 검정법

- 실험별 오류율을 제어 (예, 4그룹 : 총 6번의 비교)

6번 비교의 총 유의수준을 0.05 → 개별비교시 유의수준=0.05/6

→ 개별비교시 유의수준이 지나치게 작게 되므로 검정력이 떨어짐

→ 일반적으로 쓰이는 방법으로 상당히 보수적인 방법 중의 하나

❑ 다중비교 방법

3) Tukey 검정법

- Tukey's Honestly Significant Difference 라 불리우는 검정법
- 모든 집단의 크기가 동일한 경우에 적용 가능

→ 가장 보수적인 방법

$$HSD = q\sqrt{\frac{MSW}{n}}$$

MSW : 집단내 평균제곱
n : 각 집단내 사례수
q : α수준과 MSW의 df, 집단의 수 k에 의한 통계값

❑ 다중비교 방법

4) Scheffe 방법

- 여러 개의 대비들을 동시에 검정하는 방법
- 쌍체 비교시 Tukey 검정법보다 열등하나 여러 개의 처리 평균이 개재된 대비들에 대한 동시 검정에 유효하게 사용

5) Duncan의 다중범위 검정법(Multiple Range Test)

- 매 단계마다 최소유의범위(Least Significant Range; LSR)을 구하고 평균차이와 비교하여 결론을 내림.

step 1. 가장 큰 평균 반응값과 가장 작은 평균 반응값을 가지는 두 평균차이를 LSR과 비교 (예, $\bar{Y}_2 - \bar{Y}_4 = 85 - 62 = 23$)

만일 이 값이 LSR보다 작으면 다중비교 절차는 끝나고, 만일 LSR보다 크면 $\mu_2 - \mu_4$ 의 유의성 인정 → step 2로

step 2. 그 다음으로 평균 반응값 차이가 큰 처리들 간의 비교

검정절차

- 해당연구에 관한 귀무가설과 대립가설의 설정
  - 모든 집단의 모평균은 동일하다 (귀무가설).
  - 적어도 한 집단의 모평균은 다른 집단들과 다르다(대립가설).

- 귀무가설 하에서의 검정통계량 값을 계산
  - 집단간 평균변동 = 집단간 변동 / (집단 수 - 1)
  - 집단내 평균변동 = 집단내 변동 / (전체개체 수 - 집단의 수)
  - 검정통계량 = 집단간 평균변동 / 집단내 평균변동
  - 검정통계량은 귀무가설 하에서 F-분포를 따른다.

검정절차 (계속)

- 검정통계량으로부터 계산된 P-값을 유의수준과 비교
- 최종적인 의사결정 수립
  - 귀무가설의 기각 ⇒ 과연 어떤 집단들끼리 평균차이가 있는가?
- 다중비교 (혹은 사후검정)
  - Bonferroni의 t-검정, Fisher의 LSD, Scheffe의 다중비교, Duncan의 다중범위 검정, Tukey의 HSD 검정 등
  - Duncan (Liberal) ↔ Tukey (Conservative)

일원분류 분산분석 모형

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} = \mu_i + \varepsilon_{ij} \qquad i = 1, \cdots, t \quad j = 1, \cdots, n$$

y_{ij} : i번째 처리의 j번째 반응치

μ : 전체 평균

α_i : i번째 처리의 평균

ε_{ij} : 오차(error) $\sim N(0, \sigma^2)$

처리의 효과에 대한 검정 – $H_0 : \alpha_1 = \text{-- } = \alpha_t$ vs $H_1 :$ not H_0

❖ 제곱합 분리 :

$$\bar{y}_{i.} = \frac{\sum_{j=1}^{n} y_{ij}}{n} \qquad \bar{y} = \frac{\sum_{i=1}^{t}\sum_{j=1}^{n} y_{ij}}{tn}$$

$$\sum_{i=1}^{t}\sum_{j=1}^{n}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{t}\sum_{j=1}^{n}(\bar{y}_{i.} - \bar{y})^2 + \sum_{i=1}^{t}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2$$

SST = SSB + SSW

If SSB↑ SSW↓ → 그룹간 평균차이 존재

If SSB↓ SSW↑ → 그룹간 평균차이가 존재하지 않음

$\frac{SSB}{SSW}$ 를 이용하여 그룹간 평균차이 유무 검정

$$F = \frac{SSB/t-1}{SSW/t(n-1)} = \frac{MSB}{MSW}$$

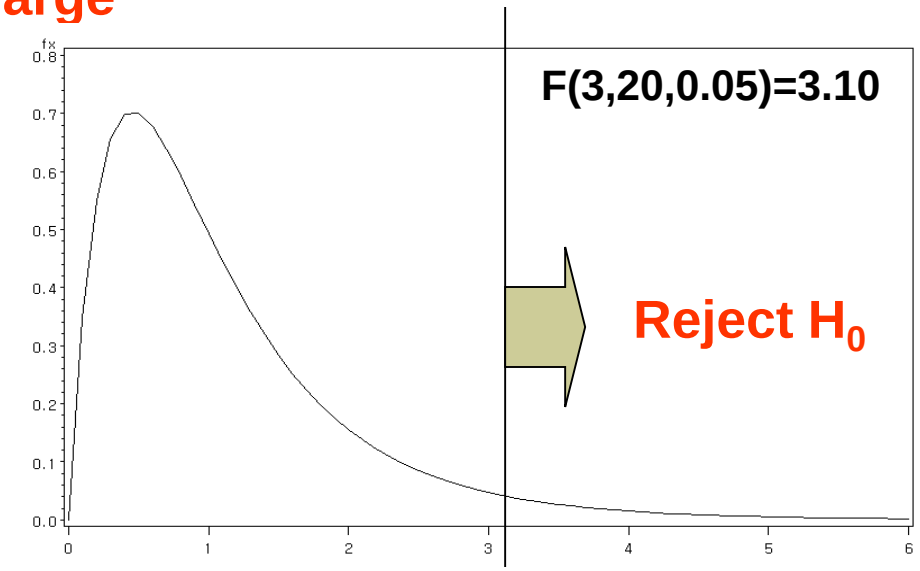
Under H_0

~ F(t-1, t(n-1))

자유도(degree of freedom, df)

Reject H_0 if F is too large

EX : F(3,20)



❑ ANOVA Table

변동의 원인	자유도	제곱합	평균제곱	F 값	P 값
처 리	*t*-1	SSB	MSB= SSB/(t-1)	F*= MSB/ MSW	
오 차	*t*(*n*-1)	SSW	MSW= SSW/t(n-1)		
전 체	*nt*-1	SST			

$P\text{-}value = \mathrm{pr}(F^* > F(t-1, t(n-1)))$

❑ Example : 상품포장색깔에 따라 판매량에 차이가 있는가?

상품포장색깔			
반복수	Red	Blue	Yellow
1	14	8	8
2	10	14	6
3	11	3	5
4	9	7	1
평균	11	8	5

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

i=1,2,3, j=1,2,3,4

t=3, n=4

전체평균 $\overline{Y}_{..}$ = 8

➢ 제곱합 분리

$$SST=\sum_{i=1}^{3}\sum_{j=1}^{4}(Y_{ij}-\overline{Y}..)^2 = (14-8)^2+(10-8)^2+(11-8)^2+(9-8)^2+\ldots+(1-8)^2=174 \quad df=3*4-1=11$$

$$SSB=\sum_{i=1}^{3}\sum_{j=1}^{4}(\overline{Y}_{i.}-\overline{Y}..)^2 = 4*\{(11-8)^2+(8-8)^2+(5-8)^2\}=72 \quad df=3-1=2$$

$$SSW=\sum_{i=1}^{3}\sum_{j=1}^{4}(Y_{ij}-\overline{Y}_{i.})^2 = (14-11)^2+(10-11)^2+\ldots+(5-5)^2+(1-5)^2=102 \quad df=3(4-1)=9$$

➢ **ANOVA Table**

변동의 원인	자유도	제곱합	평균제곱	*F* 값	*P* 값
집단간(색 깔)	2	72	36	*F*=3.18	0.0904
집단내(오 차)	9	102	11.33		
전 체	11	174			

(0.0904> 0.05= α **: 귀무 가설을 기각 못함: 상품포장색깔에 따라 판매량에 차이가 없다.)**

SAS 분산분석

- PROC ANOVA(일원분산분석/균형자료)

```
PROC ANOVA DATA=sas 파일명;
  CLASS 수준변수;
  MODEL 종속변수 = 수준변수;
  MEANS 수준변수 / LSD  DUNCAN TUKEY SCHEFFE;
RUN;
```

SAS를 이용한 일원분산분석

❑ 혈청항원의 농도자료의 분산분석 SAS프로그램-***등분산일 경우***

```
data anova;
input group $ y @@;
cards;
1 755 1 343 1 820 1 345 1 170 1 460 1 325 1 440 1 380 1 360 1 400 1 450 1 415
1 410 1 225 1 400 1 435 1 360 1 365 1 900 1 300 1 385 1 215 2 220
2 165 2 390 2 290 2 435 2 235 2 345 2 320 2 330 2 205 2 375 2 345 2 305 2 220
2 270 2 355 2 360 2 335 2 305 2 325 2 245 2 285 2 370 2 345 2 345
2 230 2 370 2 285 2 315 2 195 2 270 2 305 2 375 3 380 3 510 3 315 3 565 3 715
3 380 3 390 3 245 3 155 3 335 3 295 3 200 3 105 3 105 3 245
;
proc anova;
      class group;
      model y=group;
      means group/ tukey duncan lines;
run;
```

SAS를 이용한 일원분산분석

❑ 결과

SAS 시스템

The ANOVA Procedure

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	185159.319	92579.660	5.09	0.0087
Error	68	1236697.159	18186.723		
Corrected Total	70	1421856.479			

R-Square	Coeff Var	Root MSE	y Mean
0.130224	38.82305	134.8582	347.3662

Source	DF	Anova SS	Mean Square	F Value	Pr > F
group	2	185159.3195	92579.6597	5.09	0.0087

❑ 결과

- 귀무가설 : 자폐아, 정상아, 지진아 집단의 혈청항원의 농도의 평균이 같다.

☞ **기각**

- 이 경우 구체적으로 어떤 집단간에 유의한 차이가 있는가를 알기 위해 사후 분석이 추가로 요구됨
- 즉, 두 집단씩 묶어 t-검정을 하게 되는 경우 비교 횟수에 따라 1종 오류가 커짐
- 다중비교(multiple comparison)를 수행
  - LSD, TUKEY, DUNCAN, SCHEFFE 등
  - SAS 프로그램 중 means group/ tukey duncan lines;

SAS를 이용한 일원분산분석

❑ 결과

The ANOVA Procedure

Duncan's Multiple Range Test for y

Note: This test controls the Type I comparisonwise error rate, not the experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	68
Error Mean Square	18186.72
Harmonic Mean of Cell Sizes	21.36023

Note: Cell sizes are not equal.

Number of Means	2	3
Critical Range	82.34	86.63

Means with the same letter are not significantly different.

Duncan Grouping	Mean	N	group
A	419.91	23	1
B	329.33	15	3
B			
B	305.00	33	2

자폐아는 (정상아와 지진아)와 차이가 있음

SAS를 이용한 일원분산분석

❑ 결과

Tukey's Studentized Range (HSD) Test for y

Note: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha	0.05
Error Degrees of Freedom	68
Error Mean Square	18186.72
Critical Value of Studentized Range	3.38847
Minimum Significant Difference	98.873
Harmonic Mean of Cell Sizes	21.36023

Note: Cell sizes are not equal.

Means with the same letter are not significantly different.

Tukey Grouping	Mean	N	group
A	419.91	23	1
A			
B A	329.33	15	3
B			
B	305.00	33	2

자폐아와 정상아는 차이가 있음

SAS를 이용한 일원분산분석

❑ 혈청항원의 농도자료의 분산분석 SAS프로그램-***이분산일 경우***

```
data anova;
input group $ y @@;
cards;
1 755 1 343 1 820 1 345 1 170 1 460 1 325 1 440 1 380 1 360 1 400 1 450 1 415
1 410 1 225 1 400 1 435 1 360 1 365 1 900 1 300 1 385 1 215 2 220
2 165 2 390 2 290 2 435 2 235 2 345 2 320 2 330 2 205 2 375 2 345 2 305 2 220
2 270 2 355 2 360 2 335 2 305 2 325 2 245 2 285 2 370 2 345 2 345
2 230 2 370 2 285 2 315 2 195 2 270 2 305 2 375 3 380 3 510 3 315 3 565 3 715
3 380 3 390 3 245 3 155 3 335 3 295 3 200 3 105 3 105 3 245
;
proc glm; /* glm : 불규형자료나 선형모형 분석 */
        class group;
        model y = group;
        means group/ hovtest=levene welch dunett;
        run;
```

분산의 동일성 검정

평균 비교(등분산 x)

사후검정

SAS를 이용한 일원분산분석

❑ 결과- ***이분산일 경우***

SAS 시스템

The GLM Procedure

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	185159.319	92579.660	5.09	0.0087
Error	68	1236697.159	18186.723		
Corrected Total	70	1421856.479			

R-Square	Coeff Var	Root MSE	y Mean
0.130224	38.82305	134.8582	347.3662

Source	DF	Type I SS	Mean Square	F Value	Pr > F
group	2	185159.3195	92579.6597	5.09	0.0087

Source	DF	Type III SS	Mean Square	F Value	Pr > F
group	2	185159.3195	92579.6597	5.09	0.0087

❑ 결과- ***이분산일 경우***

The GLM Procedure

Levene's Test for Homogeneity of y Variance
ANOVA of Squared Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
group	2	1.127E10	5.6354E9	3.83	0.0265
Error	68	9.997E10	1.4702E9		

Welch's ANOVA for y

Source	DF	F Value	Pr > F
group	2.0000	4.34	0.0237
Error	25.8865		

❑ 결과- ***이분산일 경우***

Alpha	0.05
Error Degrees of Freedom	68
Error Mean Square	18186.72
Critical Value of Dunnett's t	2.26100

Comparisons significant at the 0.05 level are indicated by ***.

group Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
3 - 1	-90.58	-191.77	10.62	
2 - 1	-114.91	-197.74	-32.09	***

➢ **자폐아는 정상아와 차이가 있음**

(참고) ANOVA 프로시저

```
data anova;
input group $ y @@;
cards;
1 755 1 343 1 820 1 345 1 170 1 460 1 325 1 440 1 380 1 360 1 400 1 450
1 415 1 410 1 225 1 400 1 435 1 360 1 365 1 900 1 300 1 385 1 215 2 220
2 165 2 390 2 290 2 435 2 235 2 345 2 320 2 330 2 205 2 375 2 345 2 305
2 220 2 270 2 355 2 360 2 335 2 305 2 325 2 245 2 285 2 370 2 345 2 345
2 230 2 370 2 285 2 315 2 195 2 270 2 305 2 375 3 380 3 510 3 315 3 565
3 715 3 380 3 390 3 245 3 155 3 335 3 295 3 200 3 105 3 105 3 245
;
proc anova;
        class group;
        model y = group;
        means group/ hovtest=bartlett welch dunett;
run;
```

/* 분산의 동일성 검정 : hovtest=Bartlett */

The ANOVA Procedure

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	185159,319	92579,660	5,09	0,0087
Error	68	1236697,159	18186,723		
Corrected Total	70	1421856,479			

R-Square	Coeff Var	Root MSE	y Mean
0,130224	38,82305	134,8582	347,3662

Source	DF	Anova SS	Mean Square	F Value	Pr > F
group	2	185159,3195	92579,6597	5,09	0,0087

Bartlett's Test for Homogeneity of y Variance			
Source	DF	Chi-Square	Pr > ChiSq
group	2	28,3997	<,0001

Welch's ANOVA for y			
Source	DF	F Value	Pr > F
group	2,0000	4,34	0,0237
Error	25,8865		

The ANOVA Procedure

Dunnett's t Tests for y

Note: This test controls the Type I experimentwise error for comparisons of all treatments against a control.

Alpha	0,05
Error Degrees of Freedom	68
Error Mean Square	18186,72
Critical Value of Dunnett's t	2,26100

Comparisons significant at the 0,05 level are indicated by ***.

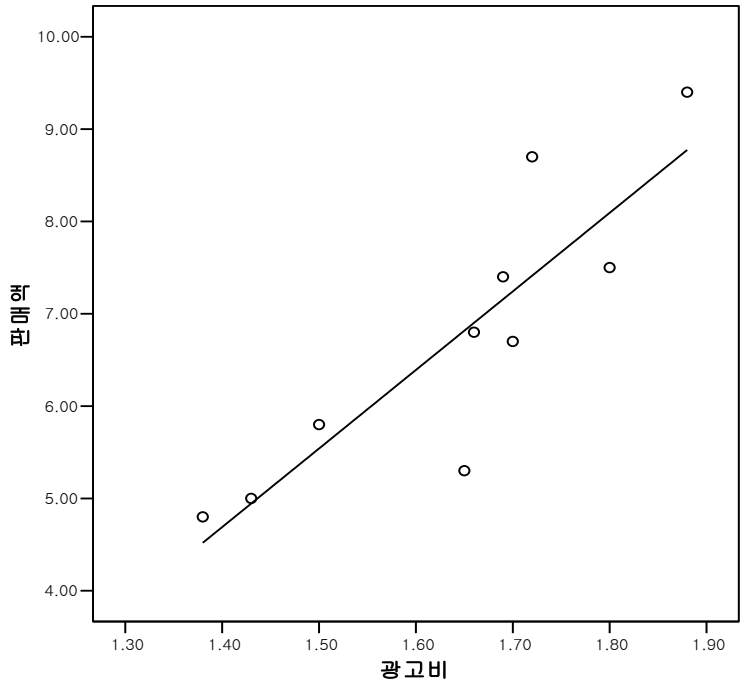
group Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
3 - 1	-90,58	-191,77	10,62	
2 - 1	-114,91	-197,74	-32,09	***

상관분석

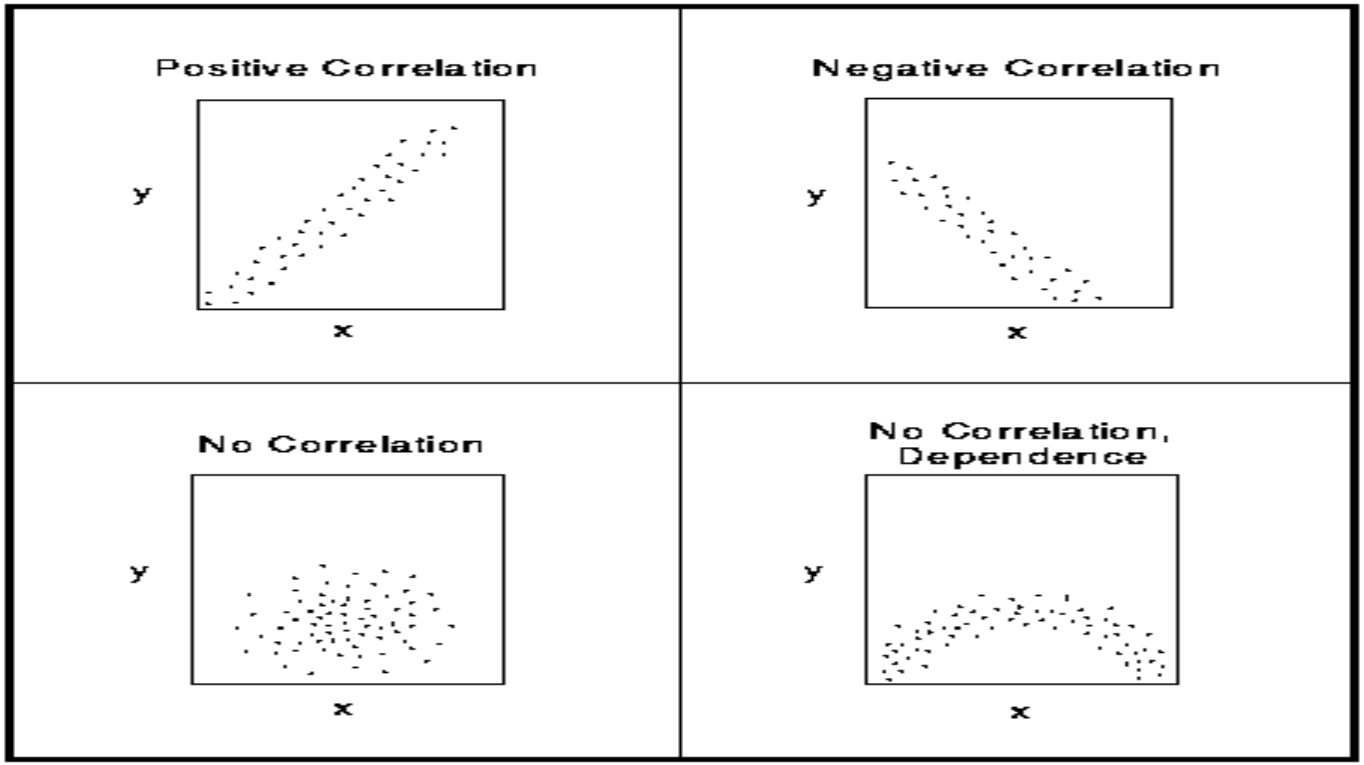
상관분석 (Correlation Analysis)

❖ 두 연속형 변수들간의 관련성 정도를 분석

광고비 X_i	판매액 Y_i
1.70	6.7
1.43	5.0
1.65	5.3
1.38	4.8
1.80	7.5
1.69	7.4
1.88	9.4
1.66	6.8
1.50	5.8
1.72	8.7



상관분석 : 상관의 형태


Positive Correlation
y
x
Negative Correlation
y
x
No Correlation
y
x
No Correlation,
Dependence
y
x

상관분석

❖ 양의 상관 (positive correlation) : 한 변수(X) 의 값이 증가하면 다른 변수(Y)의 값도 증가한다.
- 예: 나이가 증가할 수록 혈압이 증가한다.

❖ 음의 상관 (negative correlation): 한 변수(X) 의 값이 증가하면 다른 변수(Y)의 값은 감소한다.
- 예: 나이가 증가할 수록 기억력이 감소한다.

❖ 상관이 영(zero): 두 변수 사이에 선형적인 관련이 없다.

❖ 상관계수는 Pearson's correlation coefficient로 추정한다.

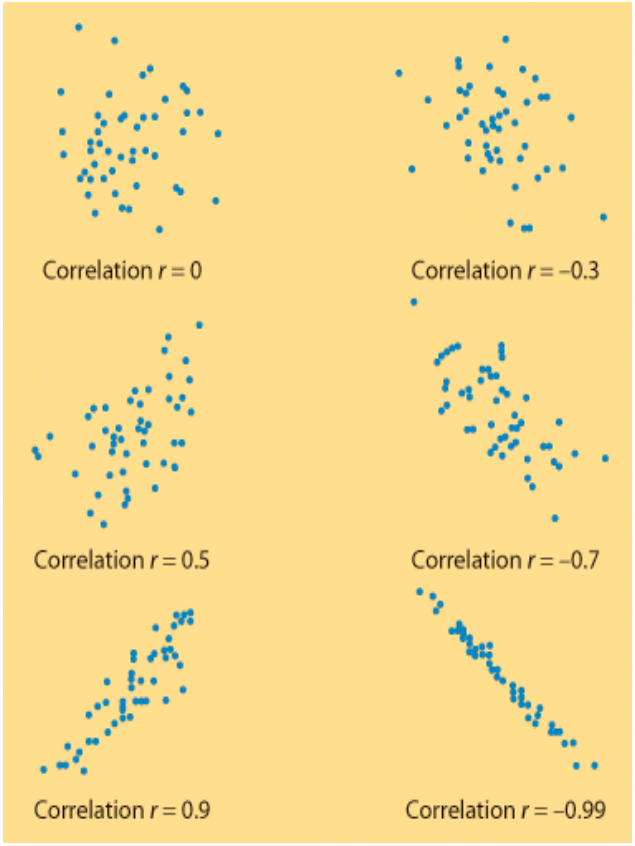
상관계수

표준화를 통해 정의되므로 단위가 없고 항상 -1과 1 사이의 값만 가짐

두 변수의 측정단위를 바꾸어도 상관계수는 불변

관측치가 정확히 직선상에만 존재하면 상관계수는 -1 or 1

상관계수의 절대값은 선형관계의 강약을 나타내고, 부호는 선형관계의 방향을 나타냄



해석의 주의점

- Y와 X 간에 상관이 있다는 것을 입증했다 하더라도, 이것이 반드시 Y의 변동이 X의 변동에 의해서 초래되었다는 것을 의미하지는 않는다. X와 Y 모두에 변동을 초래하는 제3의 변수가 "숨어" 있을 수 있다.
- 두 변수 간에 관계가 있다는 결론이 인과관계를 의미하는 것은 아니다. ➔ 상관은 인과관계를 파악하는 것이 아니다!
- 표본상관계수의 값이 "0"에 가깝다는 것은 두 변수 사이의 직선관계가 약하다는 뜻이지, 반드시 두 변수 사이에 관계가 없음을 뜻하는 것은 아니다.
- 항상 산점도를 먼저 그려서 선형성이 있는지 확인.

다른 형태의 상관계수

- 순위를 이용한 상관
  - 스피어만의 rho
    - Spearman's rank correlation coefficient
    - 자료의 순위를 이용하여 표본상관계수구함

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

  - 켄달의 타우
    - Kendall's tau
    - 두 변수간 순위의 일치쌍과 불일치쌍의 수를 이용

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{\frac{1}{2}n(n-1)}.$$

SAS 상관분석

- PROC CORR

```
PROC CORR DATA=sas_dataset PEARSON SPEARMAN ;
  BY 변수명;
  VAR 변수명;
  WITH 변수명1 변수명2 ~~ ;
RUN;
```

✓ 두 변수간의 (선형)상관관계를 분석하는 프로시저

PROC CORR

- PROC CORR 명령어

  ✓ WITH variables : 특정 조합의 변수에 대해서만 상관계수를 구하고자 할 때 사용
  예) PROC CORR; VAR A B;WITH C D; 는 A 와 C, B 와 C, A 와 D,
  그리고 B 와 D 의 상관계수만 구함

  ✓ PARTIAL variables : Pearson's partial correlation,
  Spearman's partial rank-order correlation 또는
  Kendall's partial tau-b를 구할 때
  효과가 제거되는 (Partialed) 변수를 지정

  ✓ WEIGHT variable : WEIGHT문은 PEARSON 상관계수를 구할 때에만 사용가능

  ✓ FREQ variable : 가중치를 고려할 때 사용하는 것으로 자유도 계산에 있어
  차이가 있음

PROC CORR

- PROC CORR 옵션들
  - ✓ Dataset 옵션
    - DATA=SAS-dataset
      PROC CORR 에 사용되는 SAS-dataset을 지정
    - OUTH=SAS-dataset
      Hoeffding 통계량을 갖고 있는 새로운 SAS Dataset을 만드는 경우에 사용
    - OUTK=SAS-dataset
      Kendall 상관계수를 갖고 있는 새로운 SAS Dataset을 만드는 경우에 사용
    - OUTP=SAS-dataset
      Pearson 상관계수를 갖고 있는 새로운 SAS Dataset을 만드는 경우에 사용
      이 Dataset은 TYPE=CORR이며 평균, 표준편차, 상관계수 등을 포함
    - OUTS=SAS-dataset
      Spearman 상관계수를 갖고 있는 새로운 SAS Dataset을 만드는 경우에 사용

  - ✓ 상관계수 선택에 관한 옵션
    - HOEFFDING : Hoeffding's D 통계량을 프린트
    - KENDALL : Kendall's tau-b Coefficient 프린트
    - PEARSON : Pearson Correlation을 구함(Default)
    - SPEARMAN : Spearman Coefficients를 프린트

PROC CORR

PROC CORR 옵션들

✓ 프린트에 관한 옵션

- RANK : 각 변수와의 상관계수를 프린트할 때 절대값의 크기순으로 프린트
- BEST=n : 각 변수와의 상관계수를 프린트할 때 상관계수의 절대값의 크기순으로 n개 까지만 프린트
- COV : 공분산(Covariance)을 구할 때 사용
- NOCORR : OUTPUT Dataset에서 상관계수를 포함시키지 않을 경우 사용
- NOPRINT : 상관계수를 프린트하지 않을 경우에 사용
- NOPROB : 상관계수와 관련된 유의수준확률을 프린트하지 않을 경우에 사용
- NOSIMPLE : 간단한 기술통계량을 프린트하지 않을 경우에 사용

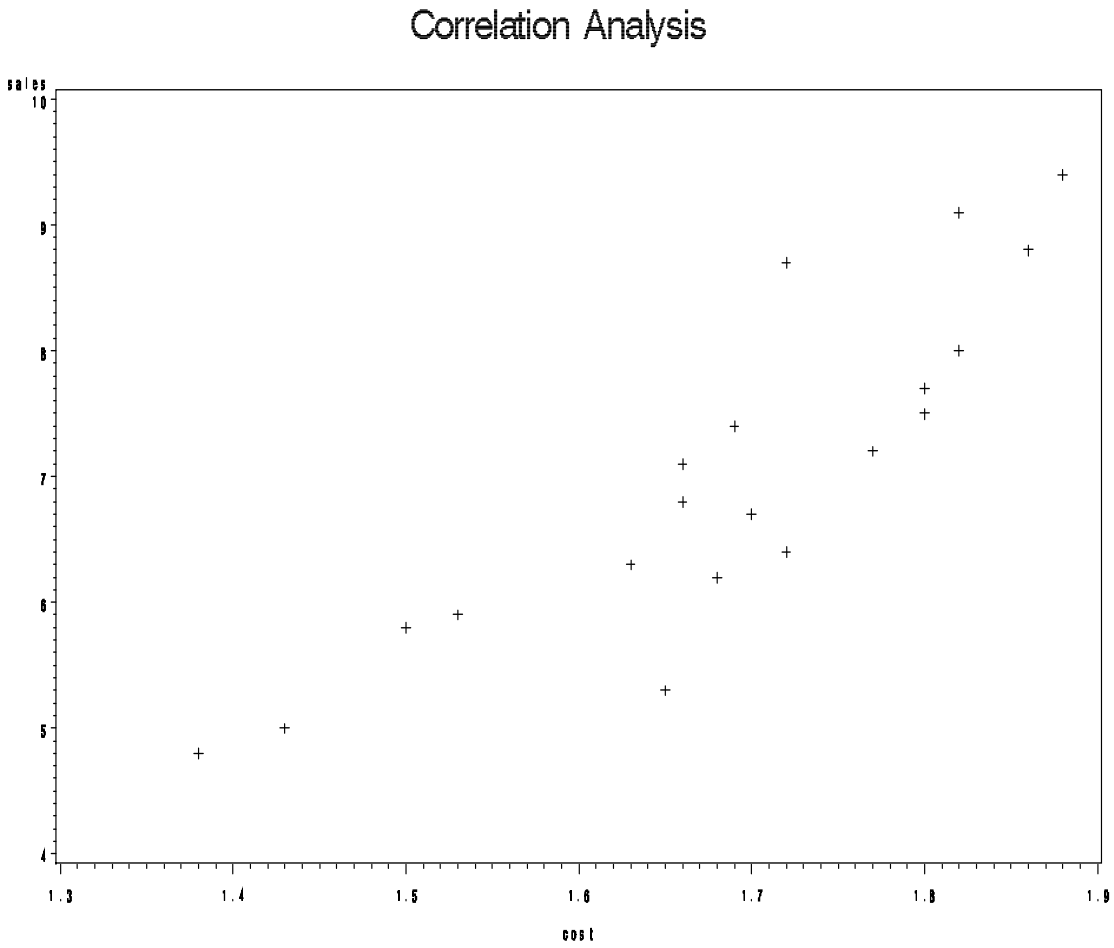
✓ 기타

- ALPHA
  Crombach's Coefficient alpha를 프린트
- NOMISS
  상관계수를 구하는 변수중 하나라도 missing이 있으면 그 관측치는 상관계수 계산에서 제외하고자 하는 경우에 사용

상관분석 예제1 광고비용과 판매액 자료 SAS

```
TITLE "Correlation Analysis";
DATA adver;
 INPUT cost sales @@;
 DATALINES;
1.70 6.7 1.77 7.2 1.65 5.3 1.38 4.8 1.80 7.5
1.69 7.4 1.88 9.4 1.66 6.8 1.82 9.1 1.72 8.7
1.80 7.7 1.53 5.9 1.63 6.3 1.68 6.2 1.43 5.0
1.72 6.4 1.66 7.1 1.86 8.8 1.82 8.0 1.50 5.8
;
RUN;
PROC GPLOT DATA=adver;
 PLOT sales*cost;
RUN; QUIT;
PROC CORR DATA=adver;
 VAR cost sales;
RUN;
```

상관분석 예제1 SAS 결과



상관분석 예제1 SAS 결과

Correlation Analysis

The CORR Procedure

2 Variables:	cost sales

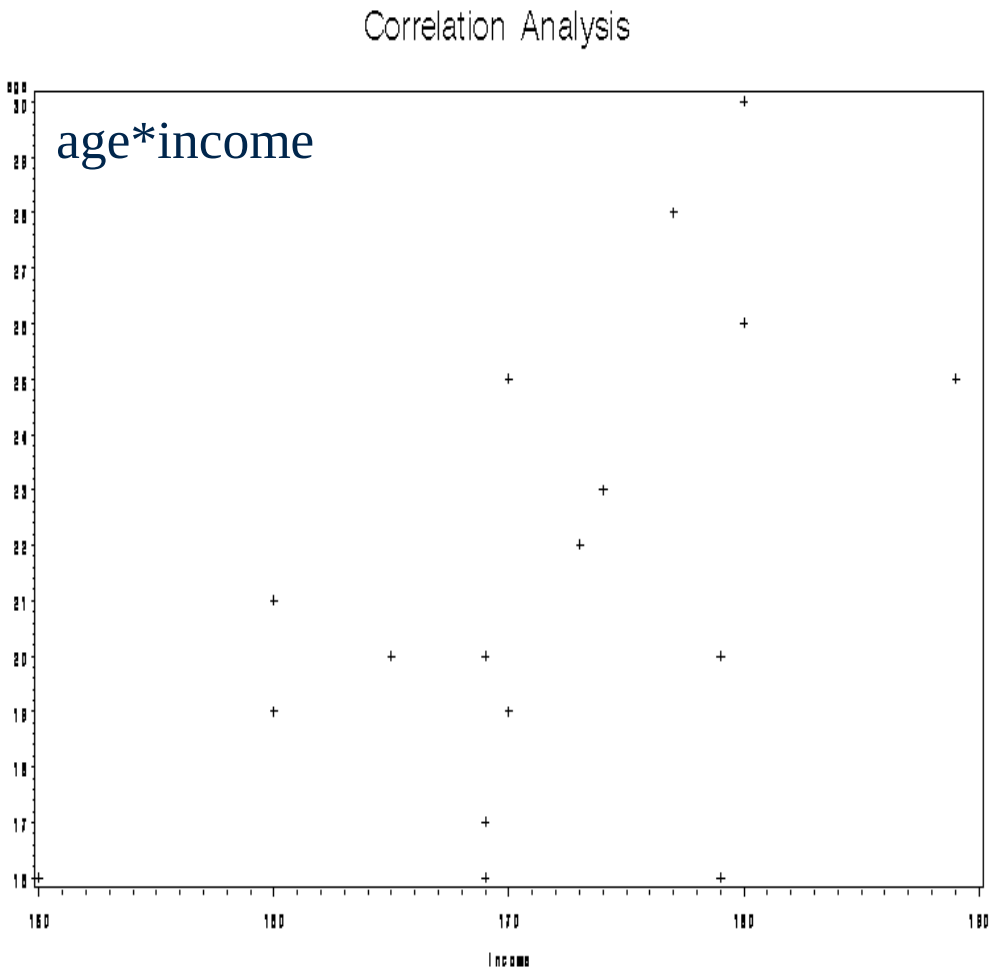
Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
cost	20	1.68500	0.13828	33.70000	1.38000	1.88000
sales	20	7.00500	1.34261	140.10000	4.80000	9.40000

Pearson Correlation Coefficients, N = 20 Prob > |r| under H0: Rho=0		
	cost	sales
cost	1.000	0.87046 <.0001
sales	0.87046 <.0001	1.000

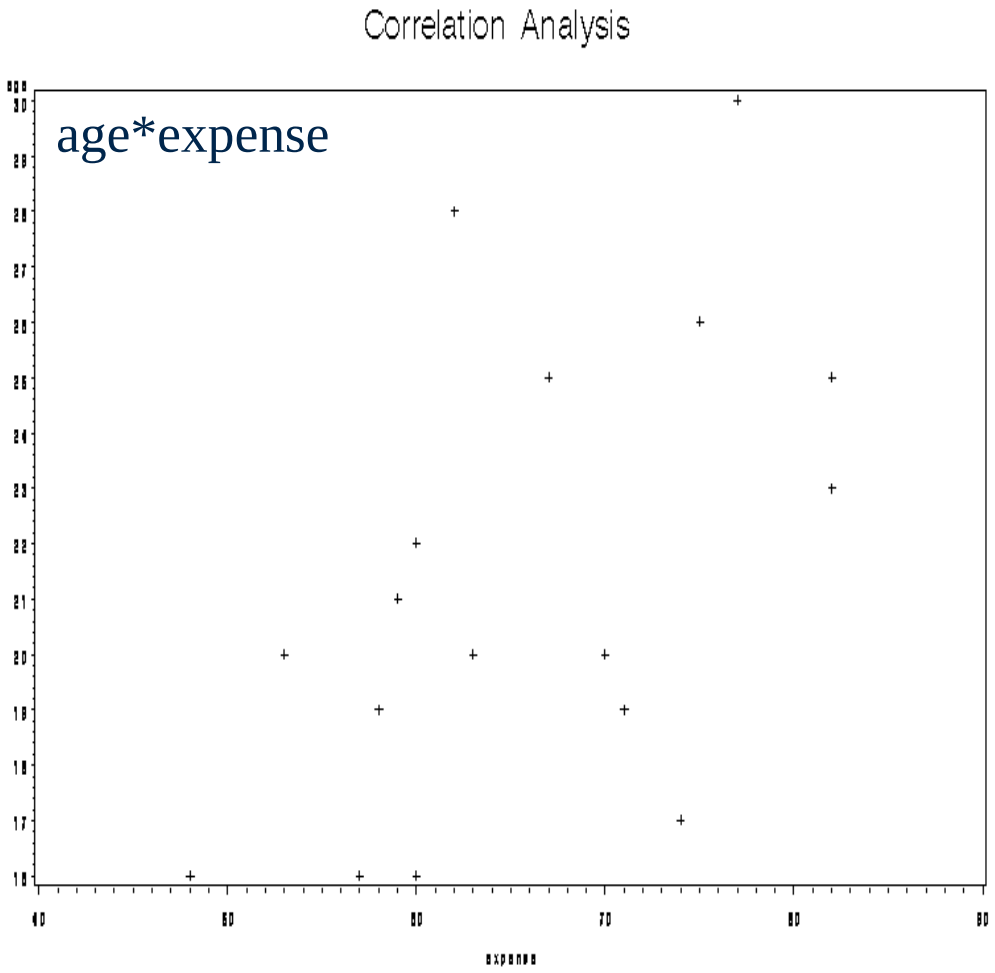
상관분석 예제2 나이, 월평균 소득, 지출 자료 SAS

```
TITLE "Correlation Analysis";
DATA student;
 INPUT age income expense @@;
 CARDS;
25 170 67 28 177 62 20 165 53 16 150 48
19 160 58 21 160 59 22 173 60 16 169 57
20 169 70 19 170 71 20 179 63 26 180 75
23 174 82 16 179 60 25 189 82 17 169 74
30 180 77
;RUN;
PROC GPLOT DATA=student;
 PLOT age*(income expense) income*expense;
RUN; QUIT;
PROC CORR DATA=student;
 VAR age income expense;
RUN;
```

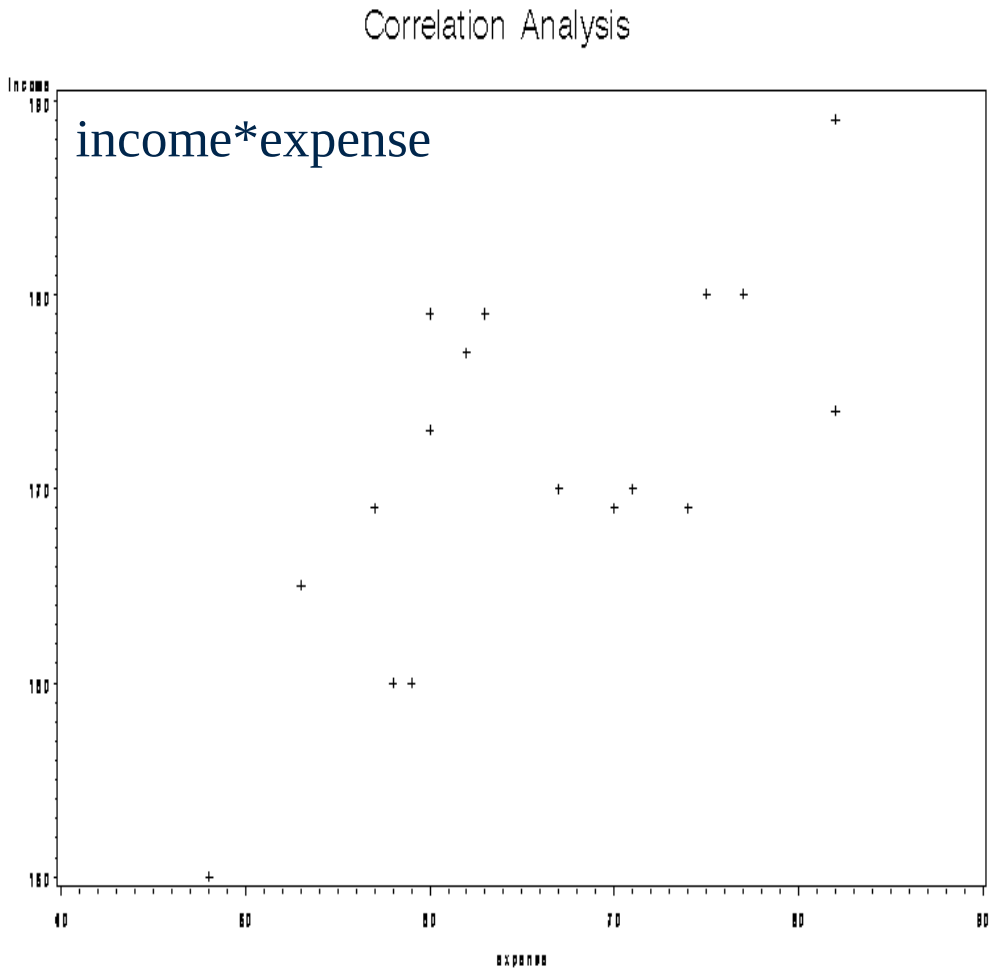
상관분석 예제2 SAS 결과



상관분석 예제2 SAS 결과



상관분석 예제2 SAS 결과



상관분석 예제2 SAS 결과

The CORR Procedure

3 Variables:	age income expense

Simple Statistics						
Variable	**N**	**Mean**	**Std Dev**	**Sum**	**Minimum**	**Maximum**
age	17	21.35294	4.27114	363.00000	16.00000	30.00000
income	17	171.35294	9.33368	2913	150.00000	189.00000
expense	17	65.76471	10.00955	1118	48.00000	82.00000

Pearson Correlation Coefficients, N = 17 Prob > |r| under H0: Rho=0			
	age	**income**	**expense**
age	**1.000**	**0.54697 0.0231**	**0.52981 0.0287**
Income	**0.54697 0.0231**	**1.000**	**0.68130 0.0026**
expense	**0.52981 0.0287**	**0.68130 0.0026**	**1.000**

회귀분석

회귀분석의 역사

TABLE 10.3.1 NUMBER OF ADULT CHILDREN OF VARIOUS STATURES BORN OF 205 MID-PARENTS OF VARIOUS STATURES (ALL FEMALE HEIGHTS HAVE BEEN MULTIPLIED BY 1.08)

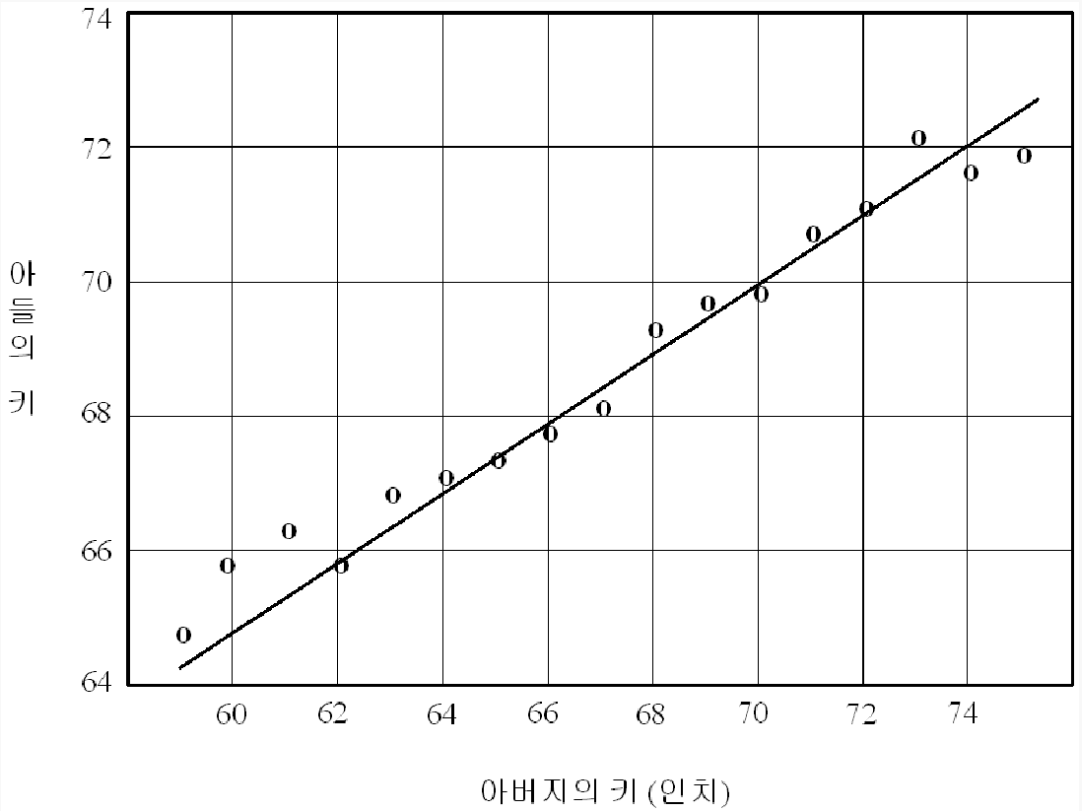
Height of the Mid-Parents (inches)	Heights of the Adult Children														Total Number of:		Medians or Values of *M*
	Below	62.2	63.2	64.2	65.2	66.2	67.2	68.2	69.2	70.2	71.2	72.2	73.2	Above	Adult Children	Mid-Parents	
Above 72.5	—	—	—	—	—	—	—	—	—	—	—	1	3	—	4	5	
72.5	—	—	—	—	—	—	—	1	2	1	2	7	2	4	19	6	72.2
71.5	—	—	—	—	1	3	4	3	5	10	4	9	2	2	43	11	69.9
70.5	1	—	1	—	1	1	3	12	18	14	7	4	3	3	68	22	69.5
69.5	—	—	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68.9
68.5	1	—	7	11	16	25	31	34	48	21	18	4	3	—	219	49	68.2
67.5	—	3	5	14	15	36	38	28	38	19	11	4	—	—	211	33	67.6
66.5	—	3	3	5	2	17	17	14	13	4	—	—	—	—	78	20	67.2
65.5	1	—	9	5	7	11	11	7	7	5	2	1	—	—	66	12	66.7
64.5	1	1	4	4	1	5	5	—	2	—	—	—	—	—	23	5	65.8
Below 64.5	1	—	2	4	1	2	2	1	1	—	—	—	—	—	14	1	
Totals	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	
Medians	—	—	66.3	67.8	67.9	67.7	67.5	68.3	68.5	69.0	69.0	70.0					

Francis Galton(1889) : 부모의 평균키와 자녀의 키의 관계를 연구

(평범으로의 회귀 → 평균값으로의 회귀)

실제 데이터는 이론상으로 추측한 값보다 평균값에 가까워진다는 의미

회귀분석의 역사



Karl Pearson(1903) : 아들의 키 ≈ 33.73 + 0.516 x 아버지의 키

1. 서론

- **회귀분석(Regression Analysis)**은 변수들 사이에 함수적 관계를 탐색하는 방법
- 관련성은 반응(혹은 종속)변수 Y 와 설명(혹은 독립)변수 X 들을 연결하는 방정식(equation) 또는 모형(model)의 형태로 표현

예) Y = 판매량, X = 광고비

Y = 컴퓨터 수리시간, X = 수리될 부품의 수

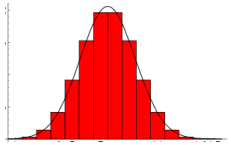
Y = 자동차 유지비용, X = 자동차 사용 년수

상관분석: $Y \leftrightarrow X$

회귀분석: $Y \leftarrow X$

1. 서론

- 변수(Variable)의 형태에 따른 구분
  - **양적(Quantitative)** 변수:
    - **연속 (Continuous)** 변수 : 무수히 많은 다른 값을 가짐
    - **이산 (Discrete)** 변수 : 몇 개의 다른 값만 가짐.
  - **질적(Qualitative)/범주형(Categorical)** 변수:
    - **명목 (Nominal)** 변수 : 순서 없는 범주를 가지는 변수
    - **순서 (Ordinal)** 변수 : 순서가 있는 범주를 가지는 변수
- 변수(Variable)의 역할에 따른 구분
  - **종속(Dependent)/반응(Response)** 변수: 결과물이나 효과
  - **독립(Independent)/설명(Explanatory)/예측(Predictor)** 변수: 입력 값이나 원인
- 반응변수와 설명변수 사이의 관계를 묘사하는 함수가 설명변수의 선형결합이 반응변수와 직접 연결되는 형태인 모형을 **선형회귀모형(Linear Regression Model)**
- 반응변수가 연속형일 때(특히, 정규분포를 따를 때) 선형회귀모형 사용

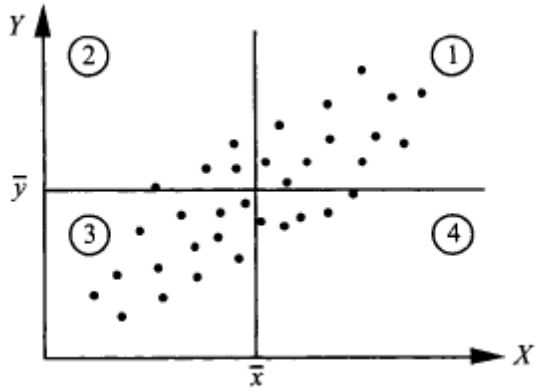


2. 단순선형회귀

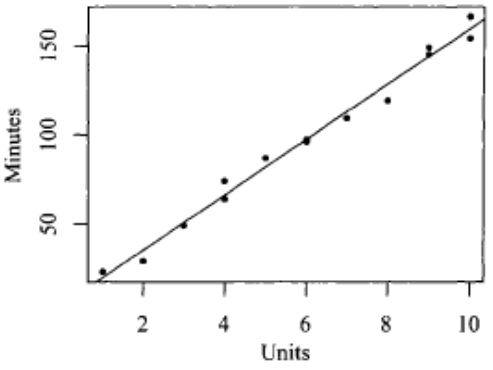
- **단순선형회귀(Simple Linear Regression)**에서는 반응변수 Y와 1개의 설명변수 X 사이에 선형적 관계 연구

Notation for the Data Used in Simple Regression and Correlation

Observation Number	Response Y	Predictor X
1	y_1	x_1
2	y_2	x_2
⋮	⋮	⋮
n	y_n	x_n



Row	Minutes	Units	Row	Minutes	Units
1	23	1	8	97	6
2	29	2	9	109	7
3	49	3	10	119	8
4	64	4	11	149	9
5	74	4	12	145	9
6	87	5	13	154	10
7	96	6	14	166	10



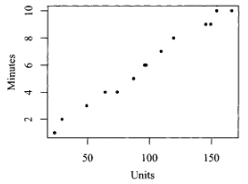
2. 단순선형회귀

- **공분산(Covariance)**과 **상관계수(Correlation Coefficient)**: 변수들 사이의 방향(direction)과 강도(strength)를 측정하는 측도

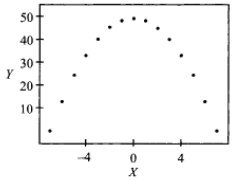
$$Cov(Y,X) = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{n-1}$$

  - 공분산 $Cov(Y,X) > 0$이면 양(positive)의 상관, $Cov(Y,X) < 0$이면 음(negative)의 상관; 강도는 알 수 없음

$$Cor(Y,X) = \frac{Cov(Y,X)}{\sqrt{\mathrm{Var}(Y)}\sqrt{Var(X)}} = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (x_i - \bar{x})^2}},\ -1 \le Cor(Y,X) \le 1$$



  - 상관계수 $Cor(Y,X) > 0$이면 양(positive)의 상관, $Cor(Y,X) < 0$이면 음(negative)의 상관; -1 또는 1에 가까울수록 관계가 더 강함
  - 상관계수는 선형(linear)관계를 측정하기 때문에 $Cor(Y,X) = 0$이 관계없음을 의미하지는 않음



2.1 단순선형회귀모형

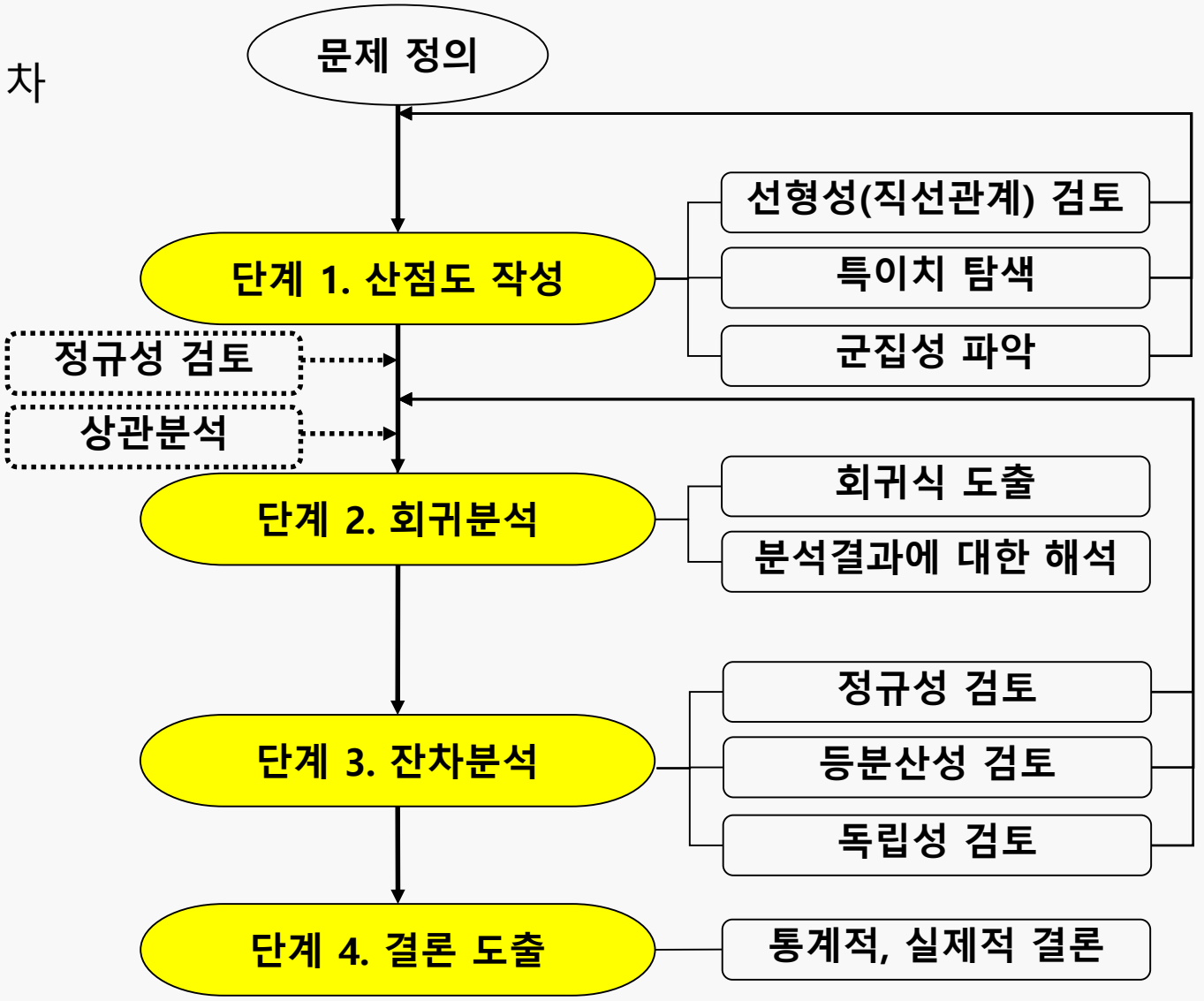
- 위 식을 각 관측개체에 따라 다음과 같이 표현

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ i = 1, 2, \ldots, n$$

- 여기서 y_i는 반응변수 Y의 i번째 값을 나타내고 x_i는 설명변수 X의 i번째 관측값을 나타냄
- 모형(model)의 형태에 대한 가정:
  - 반응변수 Y와 설명변수 X의 관계는 선형(linear)이다.
- 오차(error)에 대한 가정:
  - 오차 $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$는 정규분포를 따른다. $\epsilon_i \sim \mathrm{N}(0, \sigma^2),\ i = 1, 2, \ldots, n$
  - 오차 $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$의 각 평균은 0이다. $\mathrm{E}(\epsilon_i) = 0,\ i = 1, 2, \ldots, n$
  - 오차 $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$의 각 분산은 동일하다. $\mathrm{Var}(\epsilon_i) = \sigma^2,\ i = 1, 2, \ldots, n$
  - 오차 $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$의 서로 독립이다. $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0,\ i \neq j$

회귀분석의 절차

❑ 회귀분석의 절차



SAS 회귀분석

- PROC REG

```
PROC REG 옵션1;
  MODEL 종속변수 = 독립변수들/옵션2 ;
  BY variables;
  FREQ variable;
  ID variable;
  VAR variables;
  WEIGHT variable;
  OUTPUT OUT=SAS-dataset keyword=이름;
  PLOT 세로축변수 * 가로축변수;
RUN;
```

✓ 변수상호간의 관계를 표본으로부터 추정하는 프로시저

PROC REG

- PROC REG 명령어

```
MODEL 종속변수 = 독립변수들/옵션2 ;
```

✓ 모형선택기법에 관한 옵션

- SELECTION=name

  FORWARD 또는 F : 전진선택법(Forward Selection Method)
  BACKWARD 또는 B : 후진제거법(Backward Selection Method)
  STEPWISE : 증감법(Stepwise)
  MAXR : 최대 R2 법
  MINR : 최소 R2 법
  RSQUARE : R2 선택법
  ADJRSQ : 수정된 R2 선택법
  CP : Mallows의 Cp 선택법
  NONE (Default 옵션)

PROC REG

- PROC REG 명령어

```
MODEL 종속변수 = 독립변수들/옵션2 ;
```

✓ 모형선택의 세부적 사항 지정

- NOINT : 절편(intercept)없는 모형을 선택한다.
- INCLUDE=n :
  MODEL 문에 지정된 독립변수들 중 처음 n개의 변수를 모형에 포함
- SLENTRY=value 또는 SLE=value
  FORWARD와 STEPWISE 방법에서 모형에 변수를 추가시키고자 할 때 만족시켜야 하는 유의수준을 지정
  Default는 FORWARD의 경우에는 0.50, STEPWISE 경우에는 0.15
- SLSTAY=value 또는 SLS=value
  BACKWARD와 STEPWISE 방법에서 모형에 변수가 계속 남아있을 유의수준을 지정
  Default는 BACKWARD의 경우에는 0.10, STEPWISE 경우에는 0.15
- START=n: 독립변수 n 개 이상인 모형만 고려
- STOP=n : 독립변수 n 개 이하인 모형만 고려.

PROC REG_예제

- 예제
  - ✓ 부모의 수입과 학생의 아르바이트 수입이 학생의 용돈에 어떤 영향을 미치는가?

부모수입(fa)	아르바이트수입(ar)	학생용돈(y)
119	40	50
120	35	41
130	30	32
135	25	24
140	15	17
119	45	60
120	40	54
130	35	44
135	30	36
140	25	31
119	50	70
120	45	62
130	40	58
135	35	49
140	30	42

PROC REG_예제

- 프로그램

```
DATA a1;
 INPUT fa ar y @@;
 CARDS;
 119 40 50  120 35 41
 130 30 32  135 25 24
 140 15 17  119 45 60
 120 40 54  130 35 44
 135 30 36  140 25 31
 119 50 70  120 45 62
 130 40 58  135 35 49
 140 30 42
; RUN;
PROC REG DATA=a1;
 MODEL y=fa ar/P R;     /* P : 독립변수 값에 해당하는 종속변수의 예측값 */
 PLOT STUDENT.*(fa ar);     /* STUDENT : 표준화잔차 (회귀모형의 선형성)  */
RUN;                              /* 표준화잔차와 fa와 ar의 산점도 출력 */

  /* 오차항의 등분산성 검토 : plot student.*p.  */
```

PROC REG_예제

- 결과

SAS 시스템

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read	15
Number of Observations Used	15

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	3027.81139	1513.90570	186.28	<.0001
Error	12	97.52194	8.12683		
Corrected Total	14	3125.33333			

Root MSE	2.85076	R-Square	0.9688
Dependent Mean	44.66667	Adj R-Sq	0.9636
Coeff Var	6.38230		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-83.80563	25.42472	-3.30	0.0064
fa	1	0.47149	0.16215	2.91	0.0131
ar	1	1.95415	0.15092	12.95	<.0001

PROC REG_예제

- 결과

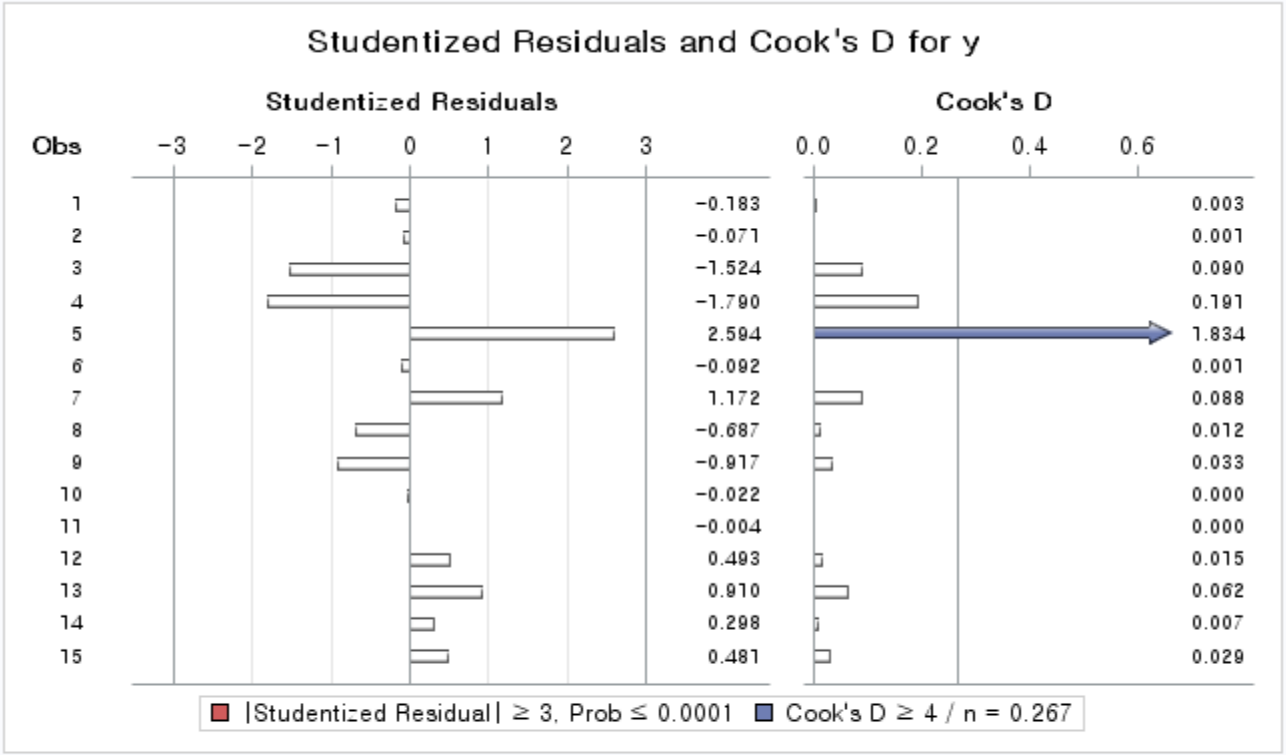
SAS 시스템

The REG Procedure
Model: MODEL1
Dependent Variable: y

Output Statistics							
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	Cook's D
1	50	50.4681	1.2573	-0.4681	2.559	-0.183	0.003
2	41	41.1689	1.5687	-0.1689	2.380	-0.071	0.001
3	32	36.1131	0.9203	-4.1131	2.698	-1.524	0.090
4	24	28.6999	1.1108	-4.6999	2.625	-1.790	0.191
5	17	11.5159	1.9121	5.4841	2.114	2.594	1.834
6	60	60.2389	1.1681	-0.2389	2.600	-0.092	0.001
7	54	50.9396	1.1444	3.0604	2.611	1.172	0.088
8	44	45.8838	0.7736	-1.8838	2.744	-0.687	0.012
9	36	38.4706	0.9312	-2.4706	2.694	-0.917	0.033
10	31	31.0573	1.2442	-0.0573	2.565	-0.022	0.000
11	70	70.0096	1.5122	-0.009596	2.417	-0.004	0.000
12	62	60.7104	1.1390	1.2896	2.613	0.493	0.015
13	58	55.6546	1.2202	2.3454	2.576	0.910	0.062
14	49	48.2413	1.2804	0.7587	2.547	0.298	0.007
15	42	40.8281	1.4839	1.1719	2.434	0.481	0.029

PROC REG_예제

- 결과

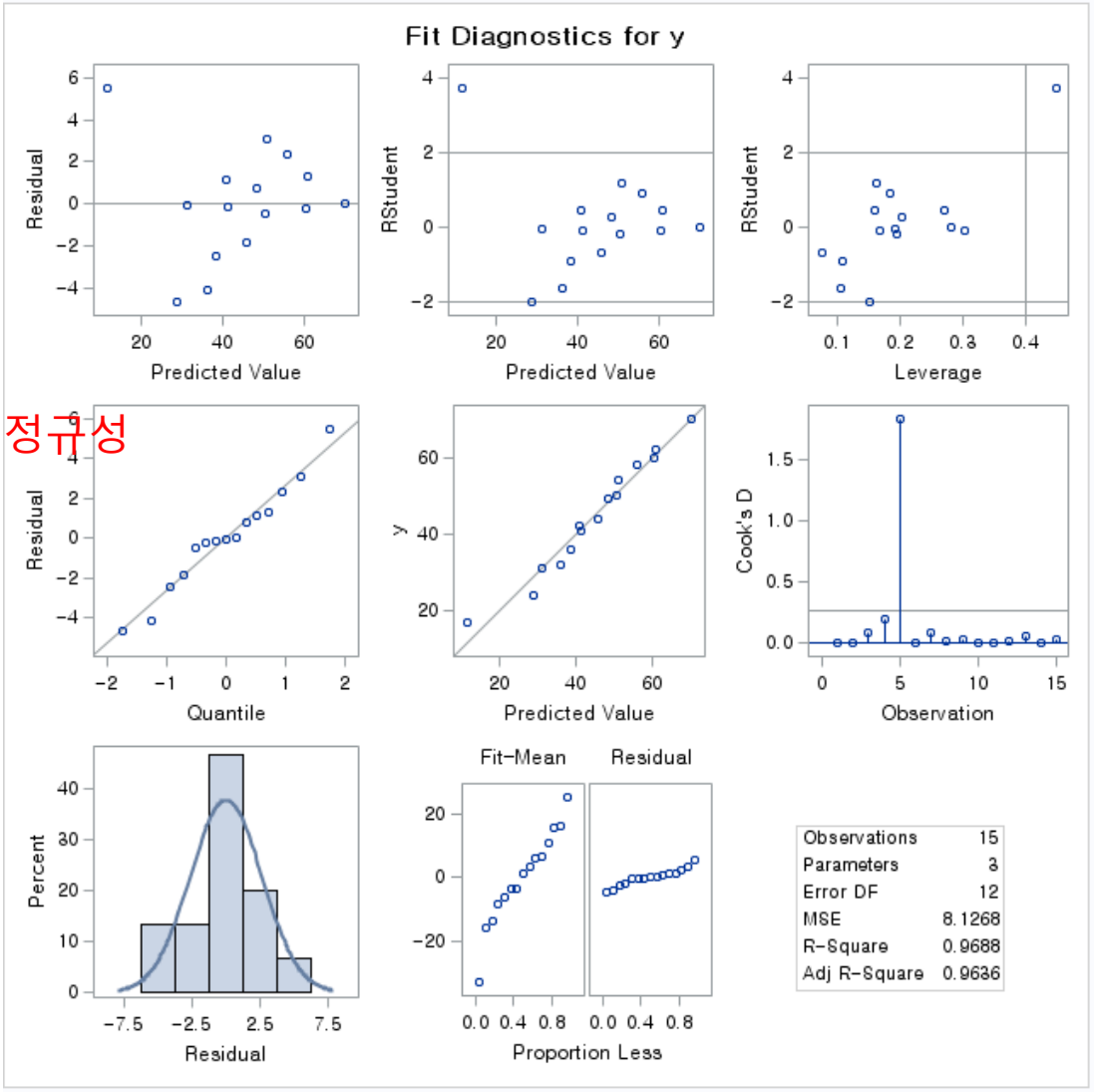


Sum of Residuals	0
Sum of Squared Residuals	97.52194
Predicted Residual SS (PRESS)	190.83813

PROC REG_예제

- 결과

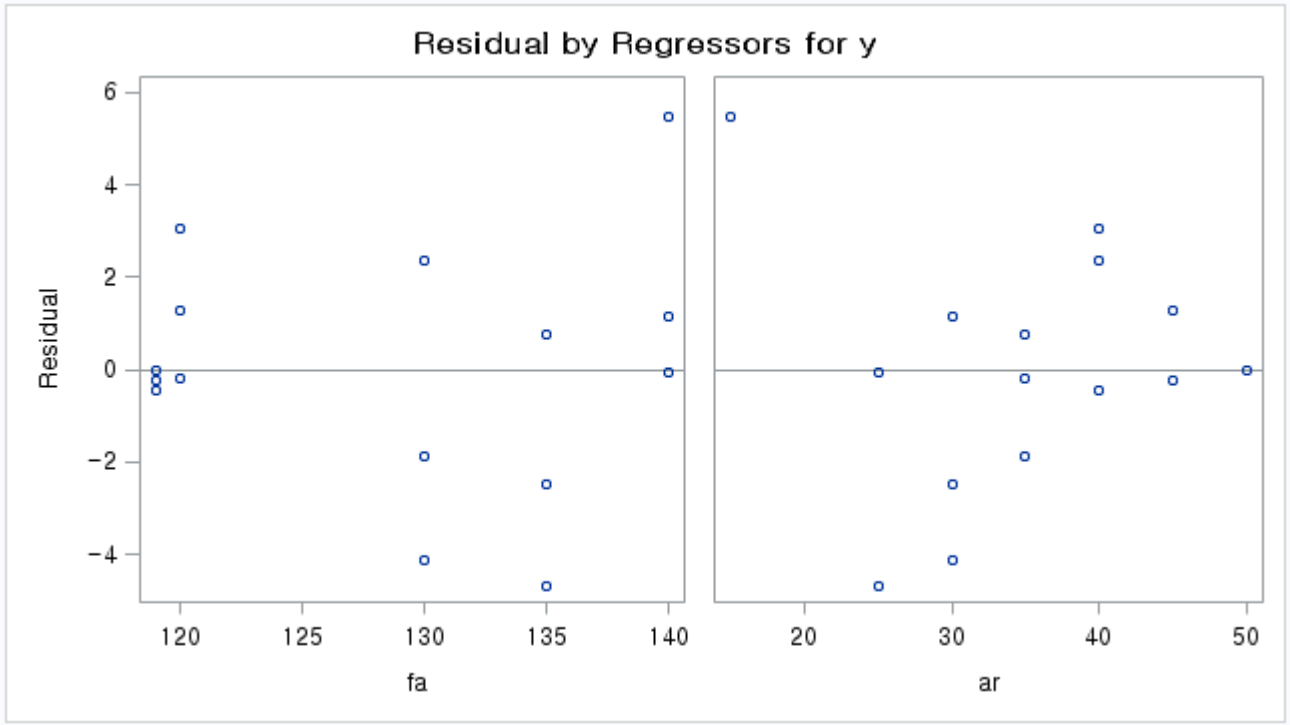
오차항의 정규성 검토



PROC REG_예제

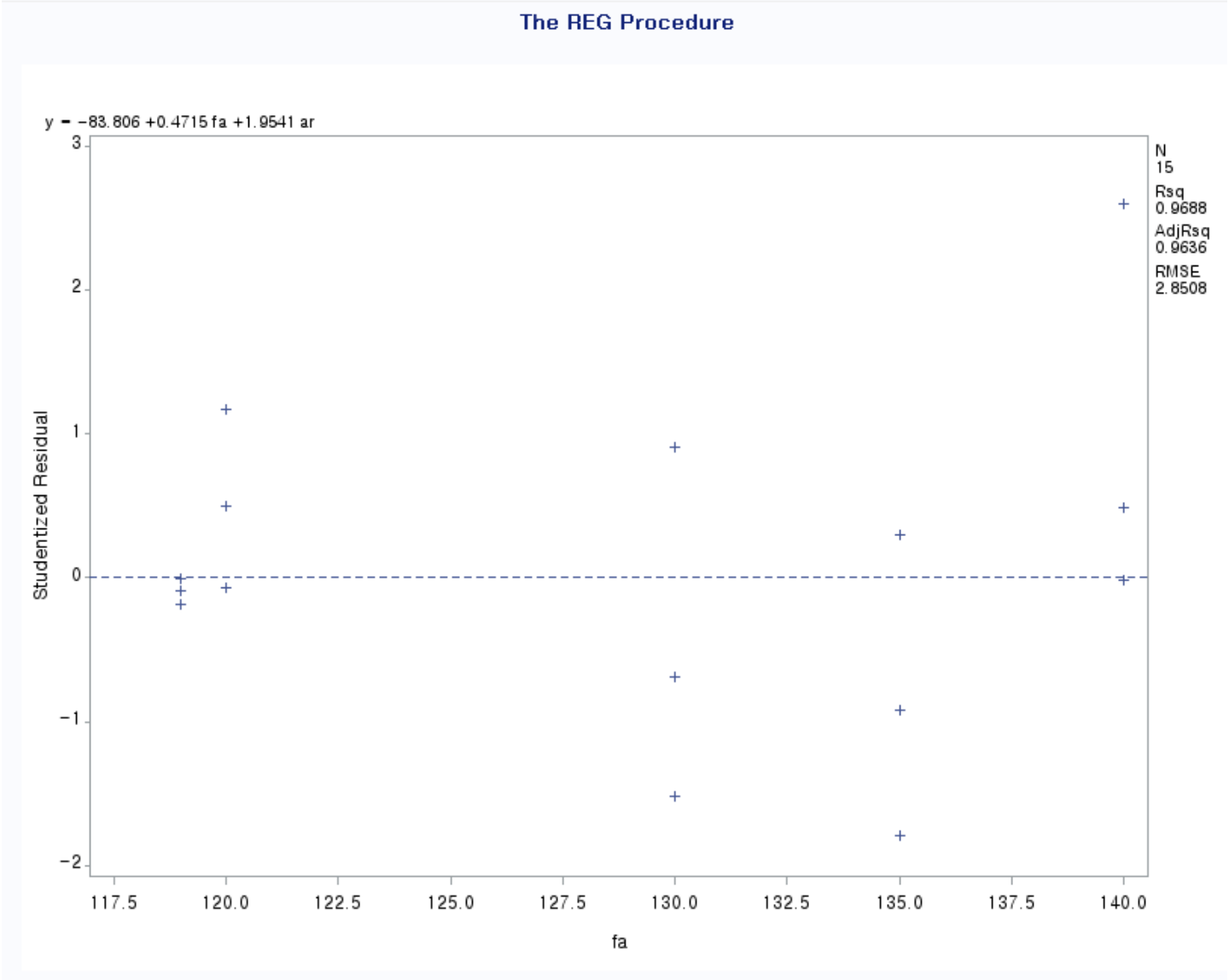
- 결과

0을 중심으로 랜덤하게
나타나면 선형성 만족



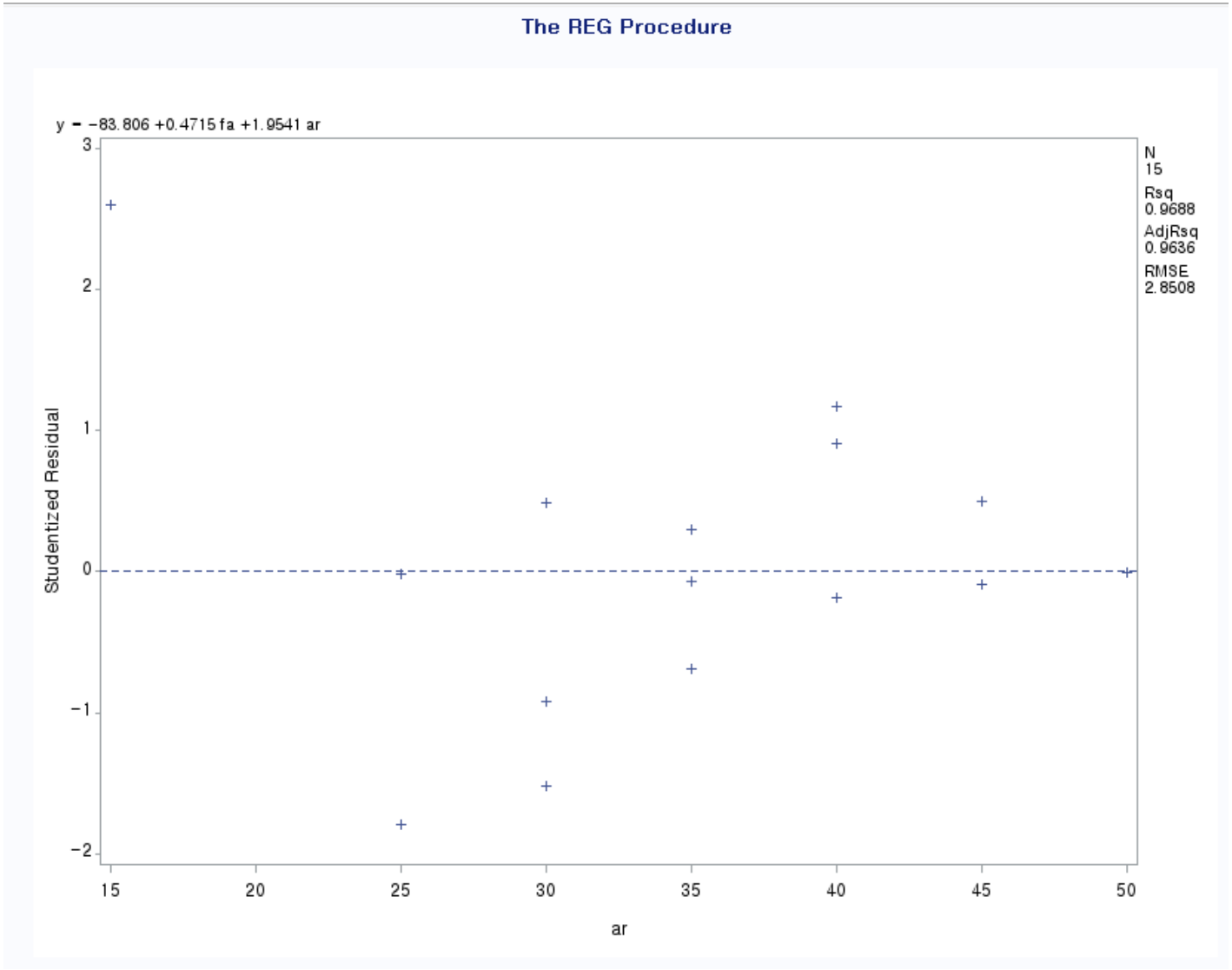
PROC REG_예제

- 결과



PROC REG_예제

- 결과



예제 : Computer Repair (단순회귀)

- 컴퓨터수리회사에서 컴퓨터수리시간(Minutes)과 교체 또는 수리되어야 할 전자부품의 수(Units) 사이의 관계 연구

Row	Minutes	Units	Row	Minutes	Units
1	23	1	8	97	6
2	29	2	9	109	7
3	49	3	10	119	8
4	64	4	11	149	9
5	74	4	12	145	9
6	87	5	13	154	10
7	96	6	14	166	10

```
data repair;
input minutes units;
datalines;
23    1
29    2
49    3
64    4
74    4
87    5
96    6
97    6
109   7
119   8
149   9
145   9
154   10
166   10
;
run;
```

```
proc corr;
  var minutes units;
run;
```

```
proc reg ;
  model minutes=units;
run;
```

예제 : Computer Repair (단순회귀)

CORR 프로시저

2 개의 변수:	minutes units

단순 통계량

변수	N	평균	표준편차	합	최솟값	최댓값
minutes	14	97.21429	46.21718	1361	23.00000	166.00000
units	14	6.00000	2.96129	84.00000	1.00000	10.00000

피어슨 상관 계수, N = 14
H0: Rho=0 가정하에서 Prob > |r|

	minutes	units
minutes	1.00000	0.99370 <.0001
units	0.99370 <.0001	1.00000

The REG Procedure
Model: MODEL1
Dependent Variable: minutes

Number of Observations Read	14
Number of Observations Used	14

Analysis of Variance

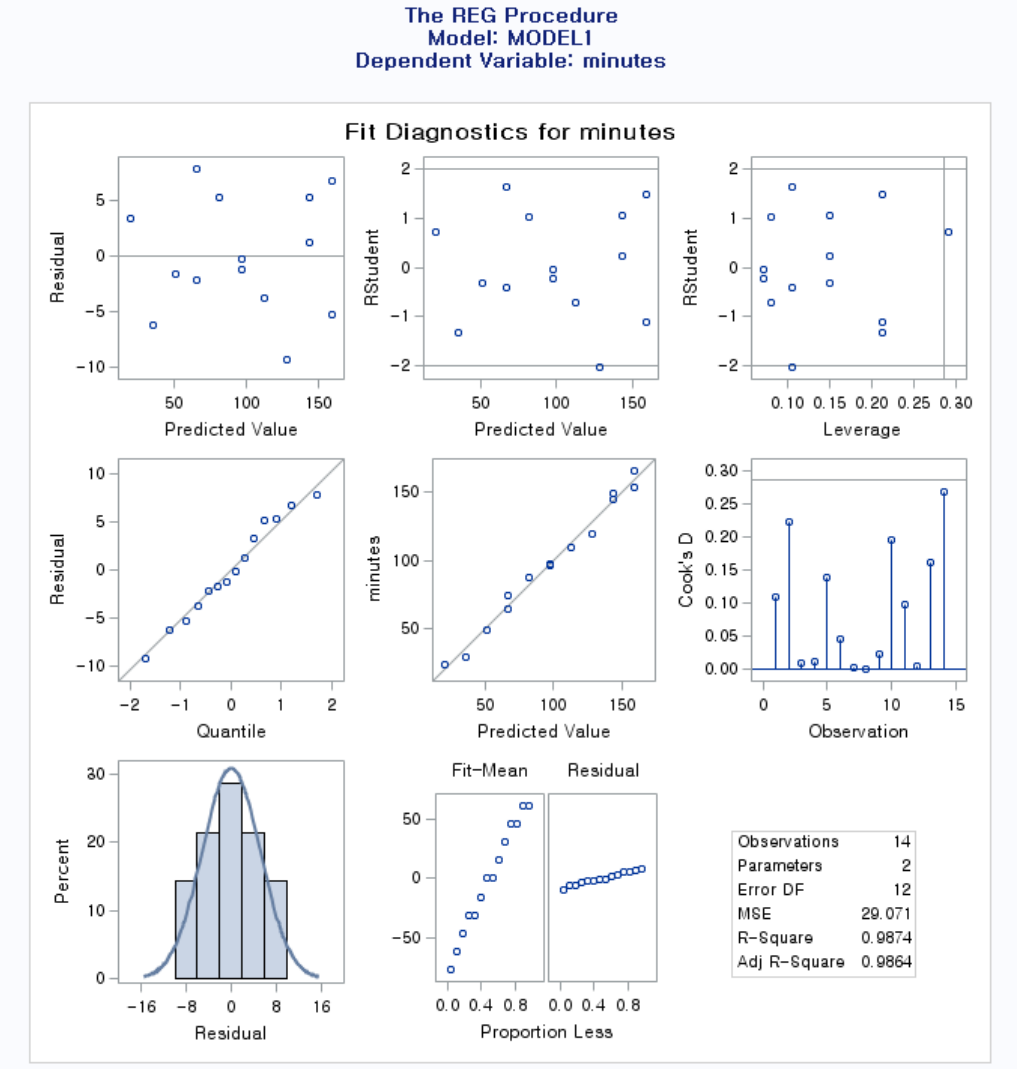
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	27420	27420	943.20	<.0001
Error	12	348.84837	29.07070		
Corrected Total	13	27768			

Root MSE	5.39172	R-Square	0.9874
Dependent Mean	97.21429	Adj R-Sq	0.9864
Coeff Var	5.54623		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.16165	3.35510	1.24	0.2385
units	1	15.50877	0.50498	30.71	<.0001

예제 : Computer Repair (단순회귀)



실습(T 검정)

- 물리학적 이론에 의하면 압축기의 냉각액으로 사용되는 물의 평균온도 증가는 5 °C이하여야 한다. 실제 압축기의 냉각액의 온도를 8번에 걸쳐 독립적으로 측정해 본 결과는 다음과 같다.

  3.4, 4.3, 5.7, 4.9, 3.5, 6.4, 5.1, 3.9

  모집단에 대해 적절한 가정을 하고, 물리학적인 이론이 옳은 가를 검정하시오.

실습(T 검정)

- 산모가 임신 중에 산부인과 검진을 잘 받는 것이 태어나는 신생아의 건강과 관련이 있는지를 연구하기 위하여 두 그룹의 산모를 조사하였다. 그룹 1은 산모가 임신중 산부인과 검진을 5회 이하 받은 경우이고 그룹 2는 산모가 임신중 산부인과 검진을 6회 이상 받은 경우이며 아래 표는 각각의 경우 태어난 신생아의 체중을 나타낸다.

	신생아의 체중 (단위는 kg)													
그룹1	1.39	3.07	3.13	2.33	2.64	3.24	3.81	3.24	2.73	1.48	2.87	3.24	3.41	3.30
그룹2	3.78	3.07	2.64	3.38	3.38	2.78	3.01	3.72	2.47	4.35	3.30	3.66	2.76	3.13

- 이 자료를 통해 그룹 2의 평균이 그룹 1의 평균보다 크다고 할 수 있는지 t 검정을 시행하시오. (유의수준 =0.05)

실습(일원분산분석)

6가지 시약

시약 1	시약 2	시약 3	시약 4	시약 5	시약 6
19.4	17.7	17.0	20.7	14.3	17.3
32.6	24.8	19.4	21.0	14.4	19.4
27.0	27.9	9.1	20.5	11.8	19.1
32.1	25.2	11.9	18.8	11.6	16.9
33.0	24.3	15.8	18.6	14.2	20.8

✓ 여섯 가지 시약을 붉은 클러브작물에 투여한 후 질소함유량을 측정한 실험의 데이터

✓ 이들 시약 종류에 따른 질소함유량의 차이가 있는가를 분석

실습(상관분석)

- 대학교 1학년생 20명을 임의 추출하여 이들의 대학 입학성적(x)과 1학년 1학기 중간고사 성적(y)간의 관계를 분석하려 한다. 자료는 grade.sas7bdat에 SAS dataset 형태로 입력되어 있다.

  - 산점도를 SAS에서 그려보고 두 변수 사이에 선형적인 관계가 있는 지 살펴보시오.

  - 상관계수를 구하고, 상관계수가 0이라는 귀무가설을 검정하시오. 또 그 결과를 설명하시오.