SPSS를 활용한 통계분석의 이해

상관분석
회귀분석

과제 1

- DMBA 투여량이 면역장기인 흉선의 중량에 영향을 미치는지 검사하기 위하여 실험한 결과 다음과 같은 결과를 얻었다. 이 자료에서 DMBA 투여량에 따라 흉선의 무게에 차이가 발생하는지 검정하라. 차이가 있다면 Duncan을 쓴 사후분석의 결과를 해석하여 쓰라. (유의수준 5%)

DMBA 투여량 (mg/g body wt)							
0mg	47	47	43	45	49	50	45
5mg	42	42	40	43	44	41	40
50mg	26	26	28	27	30	25	26
100mg	25	25	26	25	24	26	23

과제2

- 식품을 30분간 열처리한 다음 비타민 C의 함량(mg/100mg)을 측정한 결과 다음과 같은 자료를 얻었다. 열처리 방법에 따라 유의수준 0.01에서 유의한 차이가 있는지 검정하여라. 차이가 있다면 Tukey를 쓴 사후분석의 결과를 해석하여 쓰라. (유의수준 5%)

처리온도							
10도	25.3	22.7	24.5	22.8	24.6	24.3	21.7
60 도	15.6	19.0	16.0	15.7	13.5	16.3	15.4
100 도	15.3	15.9	16.5	12.3	15.8	17.1	15.7

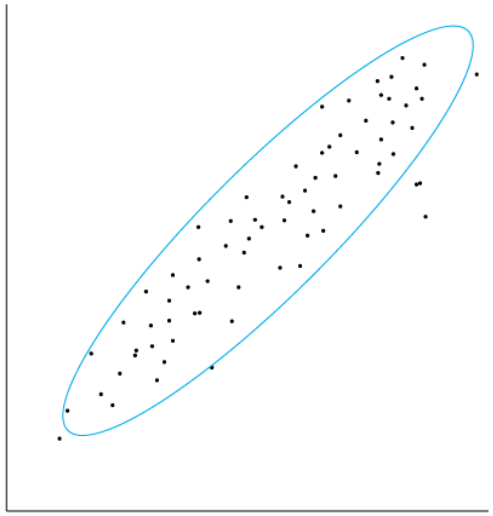
과제3-p170실습예제8-1

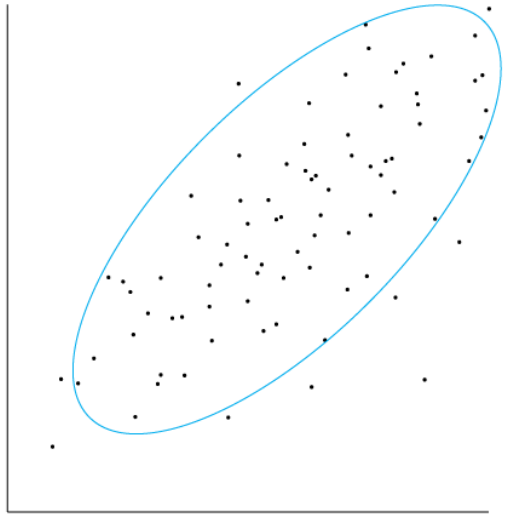
- TV시청 자세가 아동의 시력에 영향을 미치는지에 대해 실험한 자료를 통해 시청 자세에 따른 시력에 차이가 있는지 검정하고자 한다. 차이가 있다면 LSD를 사용한 사후 분석 결과를 해석하여라. (유의수준 5%)

평균시력						
앉아서	0.2	0.8	0.6	1.0	0.9	0.9
누워서	0.5	0.1	0.2	0.7	0.2	0.4
엎드려서	0.8	0.3	0.1	0.1	0.6	0.4

산점도

- 변수와 변수와의 어떤 관련성이 있다 : 상관이 있다
- 산점도
  - 두 개의 연속형 변수에 대한 관측값 패턴을 파악하기 위한 그림
  - 두 개의 연속형 변수를 각각 x축과 y축에 배치





상관분석

- 두 변수 사이의 관계
  - 종속관계
    - 어느 한 변수의 변동에 따라 다른 한 변수가 변화하는 관계
    - 예) 신생아의 체중과 생후 월령의 관계
    - 독립변수 – 시간적 또는 논리적으로 먼저 발생한 것
    - 종속변수 – 시간적 또는 논리적으로 나중에 발생한 것
  - 비종속관계
    - 어느 변수가 시간적 또는 논리적으로 우선한다고 볼 수 없는 경우
    - 예) 신장과 체중
  - 상관분석의 목적
    - 변수간의 관계의 강도 또는 정도를 측정하는 것
    - 회귀계수
      - 종속관계의 측도
    - 상관계수
      - 비종속관계의 측도

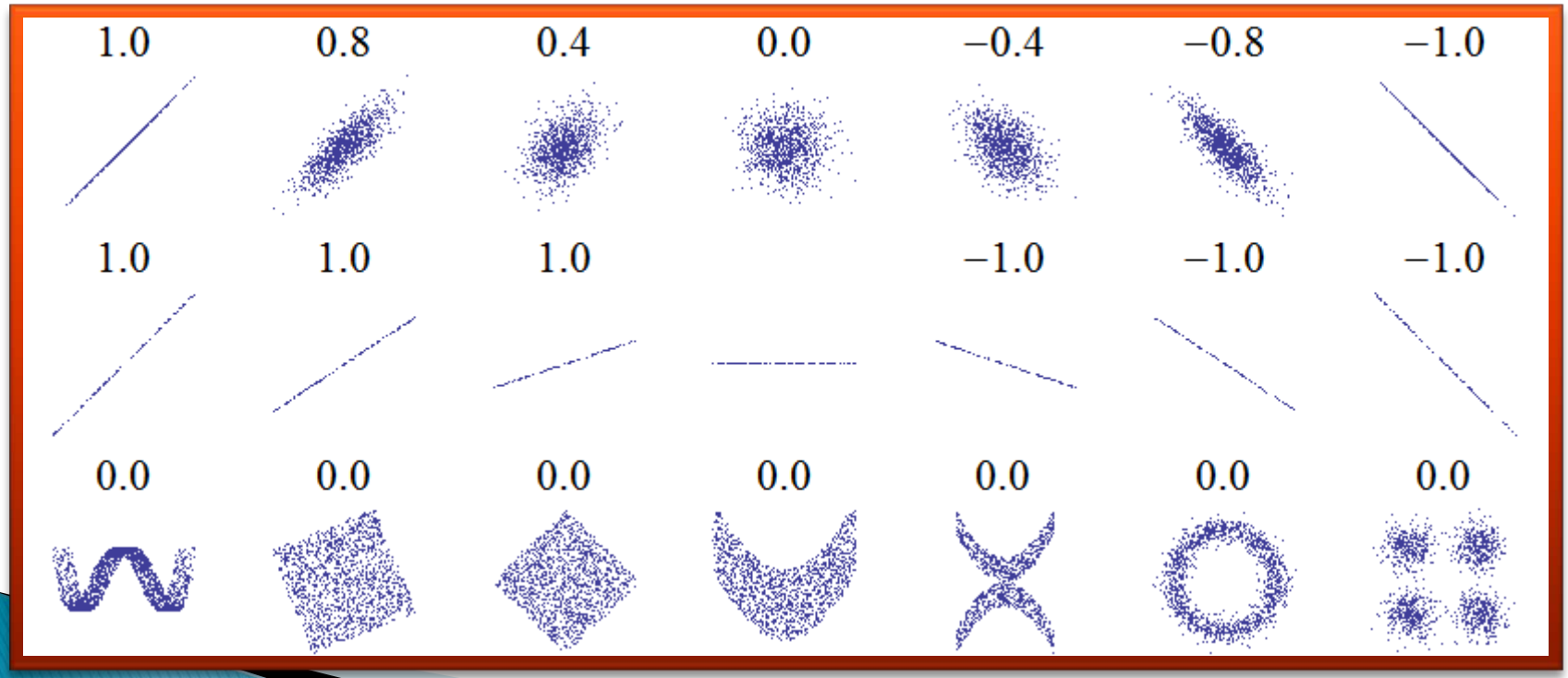
상관계수

- 표본상관계수
  - 두 변수간의 선형성의 정도를 나타내는 수치
  - 일반적으로 Pearson이 제안한 방법으로 계산 : “피어슨 상관계수”

- 상관계수
  - -1과 1사이의 값
  - 상관계수의 부호
    - 양 – 한 변수의 값이 커지면 다른 변수의 값도 커짐
    - 음 – 한 변수의 값이 커지면 다른 변수의 값은 작아짐
  - 상관계수의 절대값
    - 1에 가까울 수록 – 관측값들이 직선 주위에 몰려있으며 강한 상관관계
    - 0에 가까울 수록 – 관측값들이 기울기가 없는 직선관계로 약한 상관관계

상관분석

- 귀무가설
  - 두 변수 사이에는 선형관계가 없다. ($\rho = 0$)
- 대립가설
  - 두 변수 사이에는 선형관계가 있다. ($\rho \neq 0$)



예제

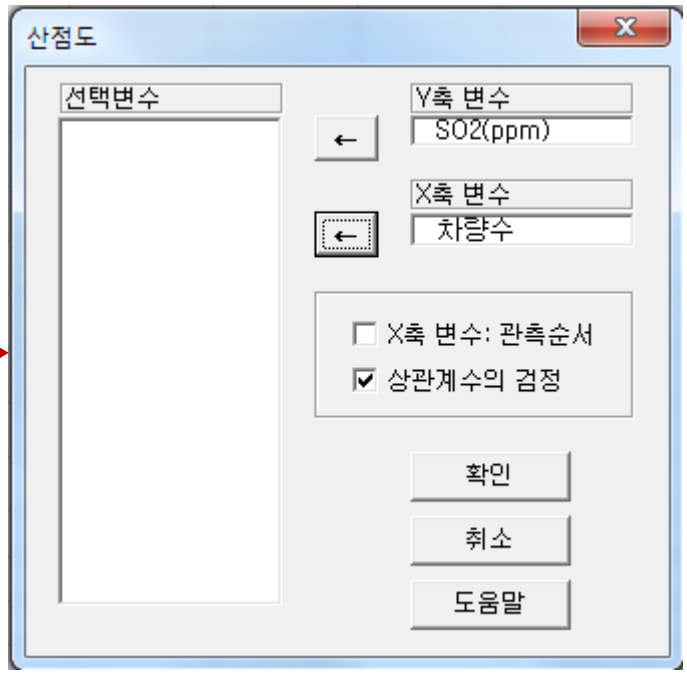
차량수	SO_2(ppm)
130	0.033
129	0.044
137	0.043
144	0.045
158	0.047
165	0.050
173	0.051
188	0.055
195	0.056
220	0.060
237	0.061
280	0.073

- 서울특별시 종로구의 공기오염도와 차량운행의 상관관계를 조사하기 위하여 아황산가스의 농도(ppm) 및 시간당 차량수를 조사하였다.
  이 자료로부터 표본상관계수를 구하라.

예제-산점도그리기

- [추가기능]-[그래프]-[산점도]
- Y-축 : 대기오염, X-축 : 차량수

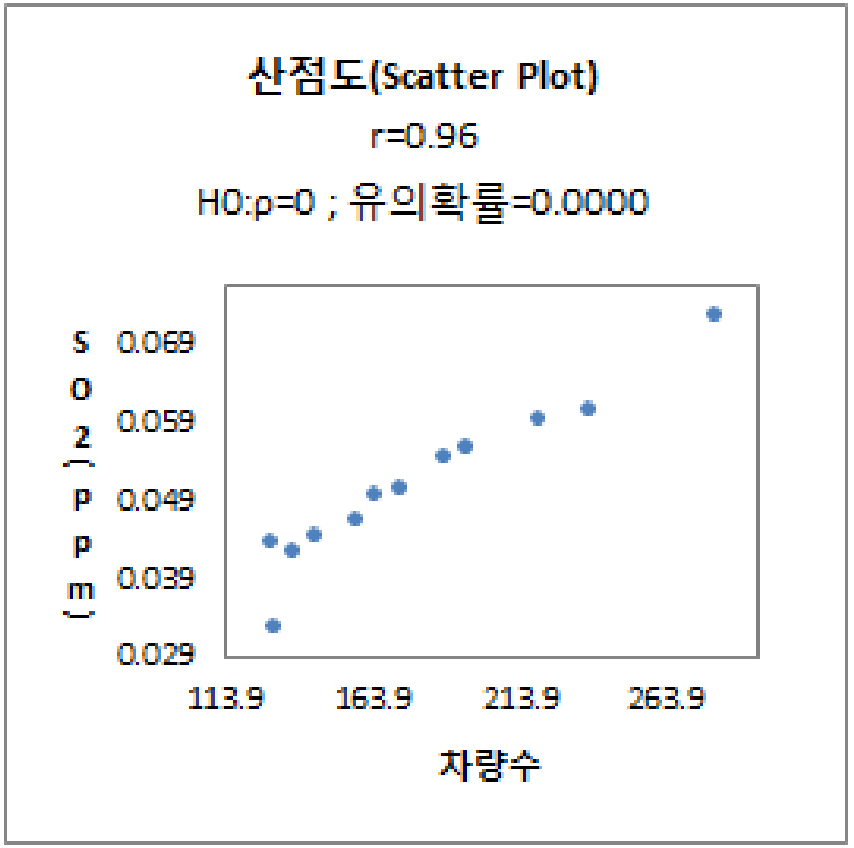
	A	B
1	차량수	SO_2(ppm)
2	130	0.033
3	129	0.044
4	137	0.043
5	144	0.045
6	158	0.047
7	165	0.05
8	173	0.051
9	188	0.055
10	195	0.056
11	220	0.06
12	237	0.061
13	280	0.073



예제-산점도그리기

- 산점도

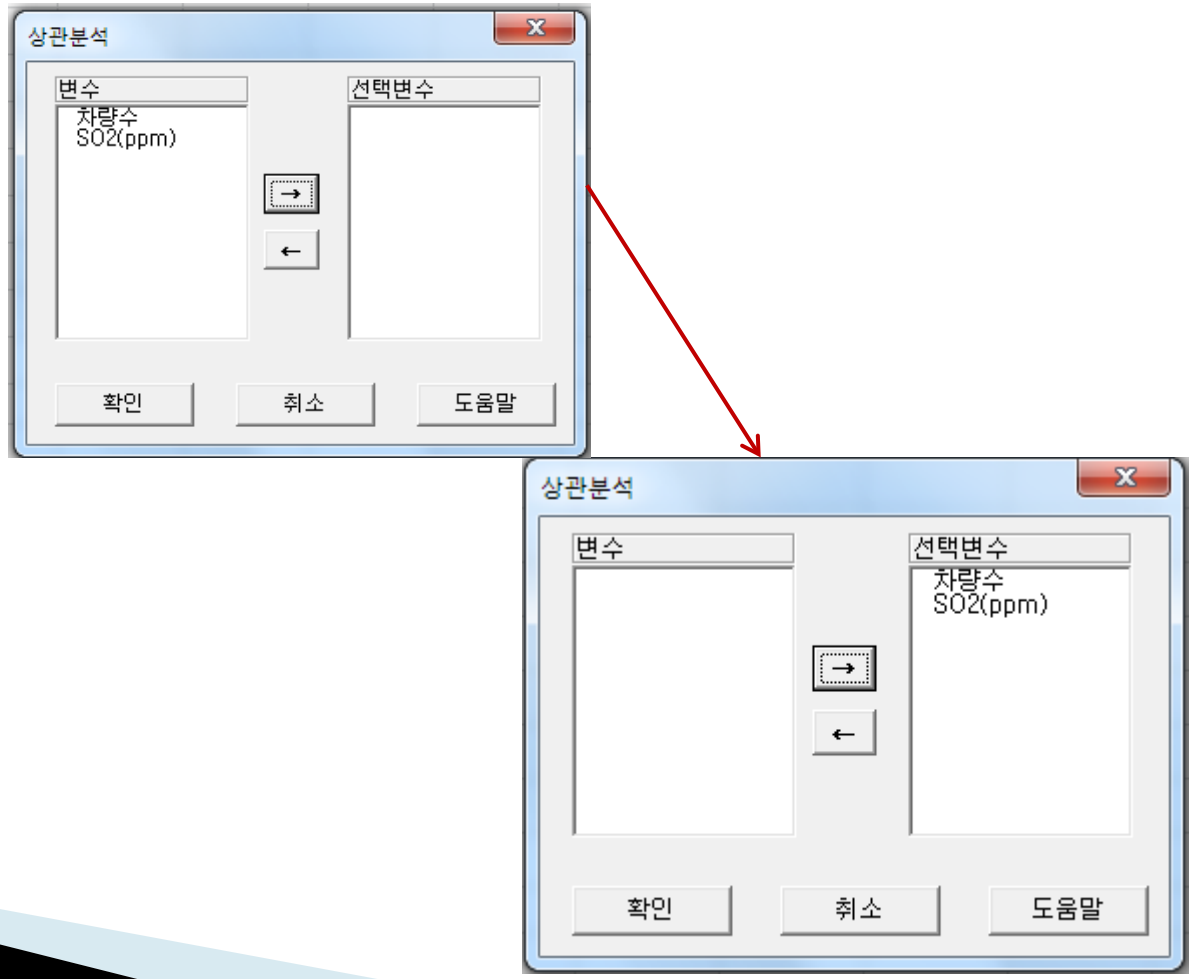
그래프출력



예제-상관분석

- [통계분석]-[상관분석]

차량수	대기오염
130	.033
129	.044
137	.043
144	.045
158	.047
165	.050
173	.051
188	.055
195	.056
220	.060
237	.061
280	.073



예제-상관분석

- 결과해석
  - 차량수와 대기오염의 Pearson 상관계수는 0.961로 1에 매우 가까운 강한 양의 선형상관을 가짐
  - 차량수가 많을수록 대기오염에 미치는 영향이 대단히 크다.
  - 유의한 상관계수 별표시
    - 유의확률 < 0.01 **
    - 유의확률 < 0.05 *

상관분석결과

상관분석

상관계수
(유의확률)

	차량수	SO2(ppm)
차량수	1	0.9613
(유의확률)	.	0
SO2(ppm)	0.9613	1
(유의확률)	0	.

회귀분석

- 단순선형 회귀모형(simple linear regression model)
  - 하나의 독립변수가 하나의 종속변수에 미치는 영향
  - 두 변수 사이의 관계를 설명할 수 있는 직선의 방정식을 표현

$$Y = a + bX$$

- 회귀분석의 가정
  - 종속변수와 독립변수가 연속형의 변수
  - 종속변수와 독립변수가 직선의 관계를 가지고 있음
  - 독립변수들 간의 상호작용이 없어야 함

회귀분석

- 분산분석: 회귀선의 유의성 검토
- 분산분석의 귀무가설
  - 회귀직선이 통계적으로 의미가 없다.
- 분산분석의 대립가설
  - 회귀직선이 통계적으로 의미가 있다.

- 계수 : 회귀계수의 유의성 검토
- 계수의 귀무가설
  - 해당 계수가 통계적으로 의미가 없다.
- 계수의 대립가설
  - 해당 계수가 통계적으로 의미가 있다.

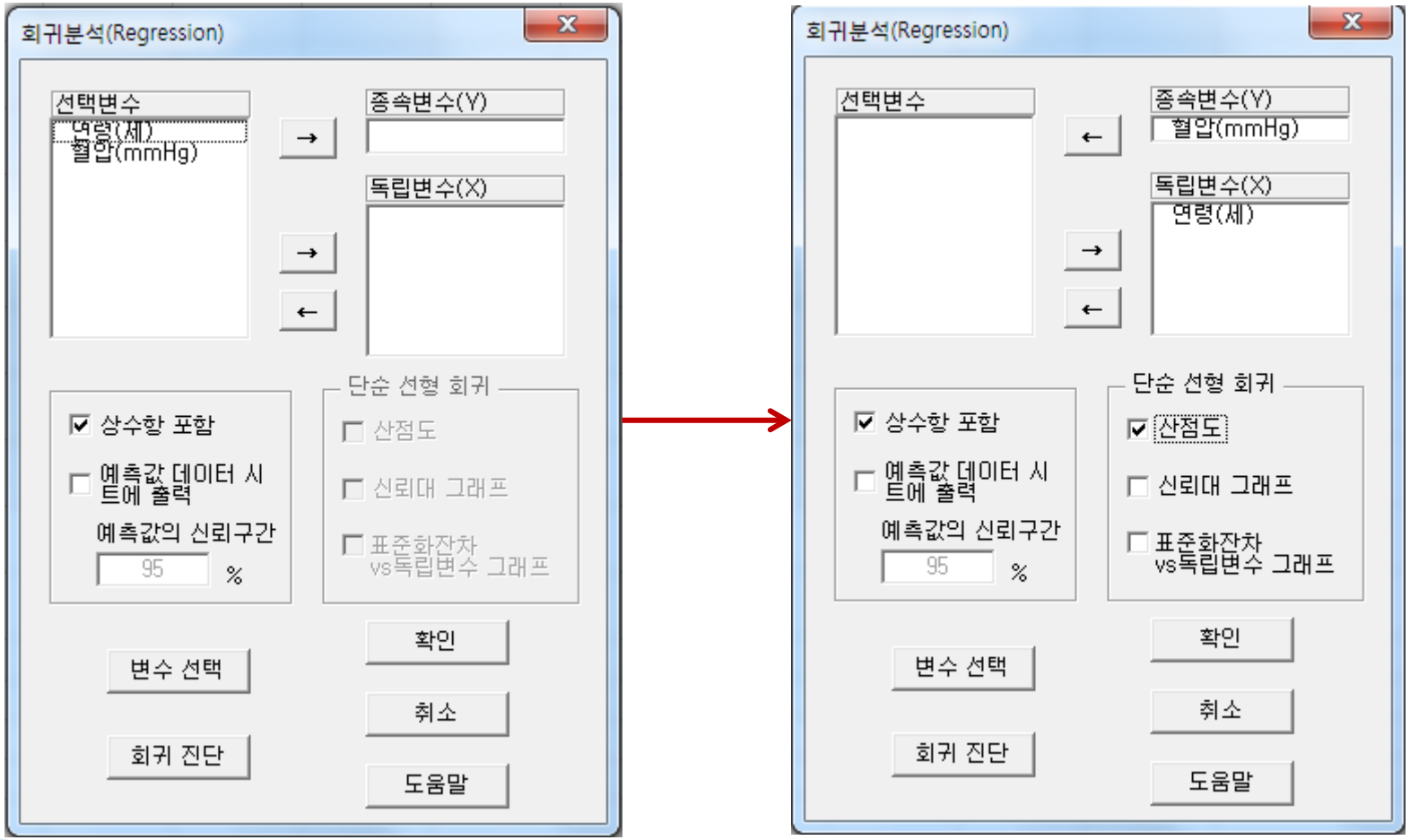
예제

연령(세)	혈압(mmHg)
30	113
35	125
40	128
45	128
50	135
55	137
60	138
65	148
70	150
75	151

- 10명의 여자의 연령과 최고 혈압과의 관계를 나타낸 것이다. 이 자료를 이용하여 회귀분석을 하라.

- [통계분석]-[회귀분석]-[회귀분석]
- 종속변수 - 혈압
- 독립변수 - 연령

예제


회귀분석(Regression)
선택변수
연령(세)
혈압(mmHg)
종속변수(Y)
독립변수(X)
상수항 포함
예측값 데이터 시트에 출력
예측값의 신뢰구간
95
%
단순 선형 회귀
산점도
신뢰대 그래프
표준화잔차 vs독립변수 그래프
변수 선택
회귀 진단
확인
취소
도움말
회귀분석(Regression)
선택변수
종속변수(Y)
혈압(mmHg)
독립변수(X)
연령(세)
상수항 포함
예측값 데이터 시트에 출력
예측값의 신뢰구간
95
%
단순 선형 회귀
산점도
신뢰대 그래프
표준화잔차 vs독립변수 그래프
변수 선택
회귀 진단
확인
취소
도움말

예제

Root MSE	2.9098
결정계수	0.9496
수정결정계수	0.9433

$$Y = a + bX$$

- 결과해석
  - 결정계수는 이 회귀모형이 전체 자료를 얼마나 잘 설명하는가를 알 수 있다.
  - 결정계수=0.9496으로 실제로 조사된 관측값을 현재 이 모형은 95%정도 설명하고 있다.

예제

- 분산분석
  - 회귀모형이 통계적으로 적합한지를 검정
  - F값이 150.752이고, 유의확률< 0.0001은 유의확률이 0.0001보다 작다는 의미이다.
  - 귀무가설 “회귀식은 통계적으로 의미가 없다”가 기각된다
  - 따라서 회귀식은 통계적으로 유의한 의미가 있다.

회귀분석결과

분산분석표

요인	제곱합	자유도	평균제곱	F 값	유의확률
회귀	1276,3667	1	1276,3667	150,752	< 0,0001
잔차	67,7333	8	8,4667		
계	1344,1000	9			

예제

- 계수
  - 회귀계수와 상수값을 표시
  - 계수의 귀무가설 : “해당계수가 통계적으로 유의하지 않다”이다.
  - (상수)는 회귀식의 a
  - 연령은 연령이 혈압에 미치는 영향력이며, 회귀식의 b
  - (상수)의 t값은 26.955이며 유의확률< 0.0001이다.
  - (상수)는 유의확률<0.0001으로 유의수준 0.0001에서 귀무가설을 기각하므로 유의하다.
  - 연령 또한 유의한 의미를 가진다.

$$Y = a + bX$$

모수 추정

변수명	추정값	표준오차	t-통계량	유의확률
절편	94,00000	3,48729	26,955	< 0,0001
연령(세)	0,78667	0,06407	12,278	< 0,0001

예제

- 회귀식 표현
  - a값을 의미하는 (상수)가 유의하다.
  - b값을 의미하는 (연령)이 유의하다.
  - 연령이 혈압에 영향을 미치는 식을 표현하고 있으므로, 독립변수 X는 (연령), 종속변수 Y는 (혈압)이다.
  - (상수)의 값은 94.000
  - (연령)의 값은 0.787

$$Y = a + bX$$

- 회귀식 : [혈압 = 94 + 0.787*연령]

모수 추정

변수명	추정값	표준오차	t-통계량	유의확률
절편	94.00000	3.48729	26.955	< 0.0001
연령(세)	0.78667	0.06407	12.278	< 0.0001

연습예제:상관분석(교재 p186)

- A 대학교 입학생 중 5명을 임의추출하여 이들의 언어영역과 수리영역에 대한 수능성적자료를 얻었다. 두 영역 성적과 관련한 상관분석을 하시오.

언어영역	수리영역
405	519
372	416
394	430
390	446
379	423

연습예제:회귀분석(교재 p190)

입시성적	평균학점
11	2.7
13	2.7
14.5	4.1
12	3.8
17	3.8
15	4.2
15.5	2.9
14	2.5
14	2.9
13	4.0

- 자녀의 수가 자녀와 함께 보내는 시간에 미치는 영향에 대해 다음과 같은 자료를 얻었다. 이를 이용하여 회귀분석을 하시오.
  - 선형회귀하는가?
  - 결정계수는?
  - 선형 회귀식은?
  - 자녀수가 6일 때 함께한 시간을 추정하면?

과제10-1

- 다음은 피검사자 12명의 분당 맥박수와 신장(cm)이다. 표본상관계수를 구하고 가설을 검정하고, 선형그래프가 포함된 산점도를 그려라.

맥박수	69	68	70	71	73	71	74	69	70	73	69	70
신장	172	165	175	168	173	174	177	170	168	170	171	173

과제10-2

인구 (만명)	CO농도 (ppm)
23	6.3
28	6.5
34	6.7
39	6.7
50	7.0
80	7.7
100	8.0
150	8.9
200	10.0

- 도시 지역의 인구 및 CO농도는 다음과 같다. 인구가 농도에 미치는 영향력에 관한 회귀모형을 적합하세요.
- 과제10 한글파일을 완성해서 제출하세요~