**SAS 데이터 분석 입문 7장**

**2019020650 김형욱**

**\* 7장 예제문제**

**<예 7-1>**

**data** sasadv.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

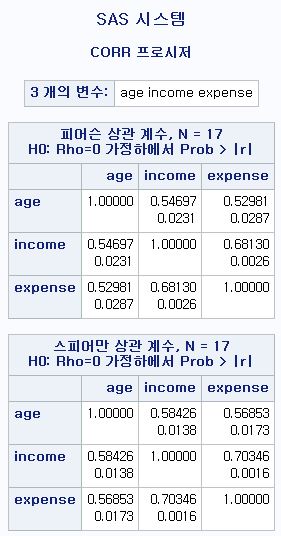
;

**run**;

**proc** **corr** data=sasadv.student pearson spearman nosimple;

var age income expense;

**run**;



해석 : 위 문제는 각 변수들 간의 피어슨 및 스피어만 상관계수와 모상관계수가 0이라는 가설을 검정하기 위한 문제이다. 검정 결과, 변수 age와 income 간의 피어슨 상관계수는 0.547이고 이에 대응되는 유의확률은 0.0231로 출력되어 있다. 또한 변소 income과 expense 간의 상관계수가 0.681로 가장 높으며, 전반적으로 모든 변수들 간의 상관계수가 비교적 높고 유의수준 5% 하에서 유의하다는 것을 볼 수 있다. 스피어만의 순위상관계수도 피어슨 상관계수와 거의 유사한 결과를 보여주고 있다.

**<예 7-2>**

**data** sasadv.satis;

input age age\_level satis1 satis2 @@;

cards;

28 2 0 70 23 2 0 55 26 2 5 65 27 2 5 65

25 2 10 60 26 2 20 65 29 2 25 70 31 3 25 75

32 3 25 80 34 3 40 85 31 3 40 75 33 3 50 80

39 3 55 95 36 3 60 90 30 3 65 75 36 3 65 90

32 3 80 80 39 3 85 95 31 3 90 75 32 3 95 80

;

**run**;

/\* 피어슨 상관계수 \*/

**proc** **corr** data=sasadv.satis;

var age satis1 satis2;

**run**;

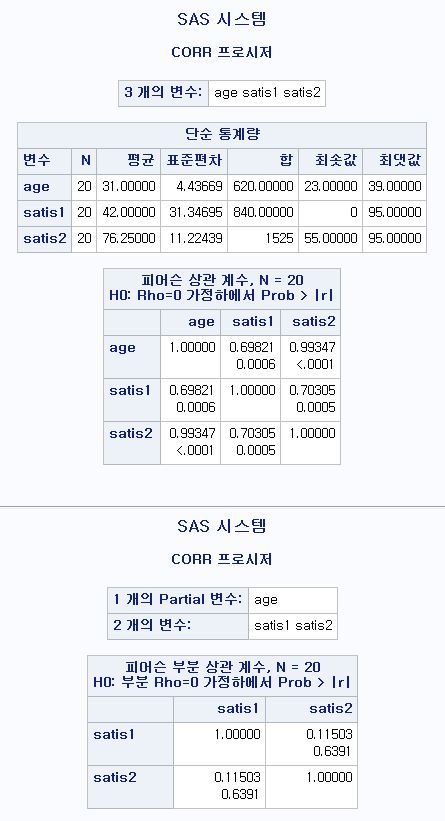
/\* 편상관계수 \*/

**proc** **corr** data=sasadv.satis nosimple;

var satis1 satis2;

partial age;

**run**;



해석 : 위 문제는 특정 제품에 대해서 20명의 고객으로부터 연령, 기능 만족도, 디자인 만족도를 조사하여 얻은 ‘제품 만족도’ 데이터에 대한 문제이다. 피어슨 상관계수 및 편상관계수를 각각 구해보았다. 먼저 피어슨 상관계수의 경우, 세 변수에 대해 모든 상관계수들이 0.6보다 큰 값을 가지므로 전반적으로 높은 상관이 있다는 것을 알 수 있고, 특히 연령(age)과 디자인 만족도(satis2) 간의 피어슨 상관계수는 0.993으로 거의 1에 가까움을 알 수 있다. 다음으로 변수 age(연령)를 통제했을 때 satis1(기능 만족도)과 satis2(디자인 만족도)의 편상관계수를 살펴보면, 편상관계수가 0.11503으로 피어슨 상관계수보다 작음을 알 수 있다. 따라서 통제변수인 age(연령)가 같은 값을 가지도록 고정된다면 satis1(기능 만족도)과 satis2(디자인 만족도)사이의 연관성이 크지 않음을 알 수 있다. 즉, 연령(age)이 클수록 이 제품에 대해 기능 만족도(satis1)와 디자인 만족도(satis2)가 모두 커지는 경향이 있다(변수 age가 두 변수 사이의 매개변수 역할)는 것을 알 수 있다.

**<예 7-3>**

**proc** **sort** data=sasadv.satis;

by age\_level;

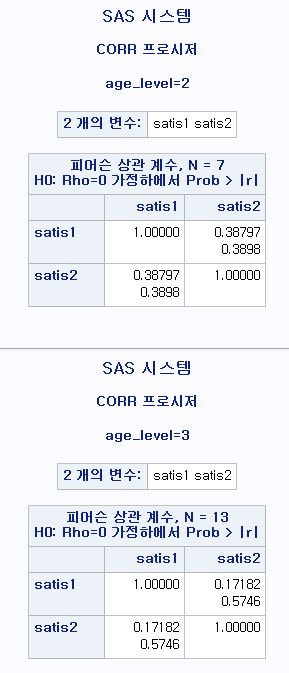
**run**;

**proc** **corr** data=sasadv.satis nosimple;

var satis1 satis2;

by age\_level;

**run**;



해석 : 다음은 연령대별(20대, 30대)로 satis1과 satis2의 피어슨 상관계수를 각각 계산한 것으로, 각 연령대에서 기능 만족도(satis1)와 디자인 만족도(satis2)의 상관계수가 매우 작고 유의확률도 각각 0.390, 0.575로 유의수준 5% 하에서 유의하지 않다는 것을 알 수 있다.

**<예 7-4>**

**data** sasadv.alpha;

infile 'F:\SAS 독학\예제로 배우는 SAS 데이터 분석 입문 데이터\alpha.txt';

input q01-q10;

**run**;

**proc** **corr** data=sasadv.alpha nocorr alpha;

var q01 q02 q03;

**run**;



해석 : 위 예제는 ‘상사의 업무수행능력’에 대한 신뢰도 분석이다. 계산된 크론바흐 알파계수의 값은 0.735(표준화된 신뢰도는 0.734)로 신뢰도가 비교적 높다는 것을 알 수 있다. 또한 ‘합계에 대한 상관계수’는 각 변수와 해당되는 구성개념(합점수)과의 상관계수를 의미하고, ‘변수를 제외했을 때의 크론바흐 알파계수’는 해당 변수를 제거한 뒤 남아 있는 나머지 두 문항만을 기초로 계산된 알파계수를 의미한다. 이 예에서는 신뢰도 분석의 결과가 전반적으로 만족스러운 수준이라고 할 수 있다.

**<예 7-5>**

**proc** **corr** data=sasadv.alpha nocorr alpha;

var q04 q05 q06 q07;

**run**;



해석 : 위 예제는 ‘상사와의 공적/개인적 긴밀함’에 대한 신뢰도 분석이다. 결과를 살펴보면, 크론바흐 알파계수가 -0.222로 특히 문항 ‘q07’이 합점수와 매우 큰 음의 상관을 가진다는 것을 알 수 있다. 그런데 10개의 문항들을 자세히 살펴보면, 문항 ‘q07’의 경우 다른 문항들과는 반대의 방향으로 질문을 하고 있다는 것을 알 수 있다. 따라서 문항 ‘q07’에 대해서는 관측값 5, 4, 3, 2, 1을 1, 2, 3, 4, 5로 바꾼 후 합점수를 구해야 한다.

**<예 7-6>**

**data** sasadv.alpha2;

set sasadv.alpha;

q07=**6**-q07;

**run**;

**proc** **corr** data=sasadv.alpha2 nocorr alpha;

var q04 q05 q06 q07;

**run**;



해석 : 위 예제는 문항 ‘q07’에 대해서 관측값 5, 4, 3, 2, 1을 1, 2, 3, 4, 5로 바꾼 후의 ‘상사와의 공적/개인적 긴밀함’에 대한 신뢰도 분석이다. 결과를 살펴보면, 다시 계산된 크론바흐 알파계수가 0.588로 비교적 만족스럽고 각 문항들과 합점수의 상관계수는 모두 양의 부호를 가진다는 것을 알 수 있다.

**<예 7-7>**

**proc** **corr** data=sasadv.alpha nocorr alpha;

var q08 q09 q10;

**run**;



해석 : 위 예제는 ‘업무추진의 독자성’에 대한 신뢰도 분석이다. 결과를 살펴보면, 크론바흐 알파계수가 0.251로 매우 낮기 때문에 이들 문항이 하나의 동질적인 개념을 측정한다는 것을 신뢰할 수 없으며, 세 문항의 합점수를 사용하는 것은 문제가 된다는 것을 알 수 있다.

**\* 7장 연습문제**

**<연습문제 7-2>**

**data** sasadv.ex7\_2;

input x y;

cards;

1 4

2 1

3 0

4 1

5 4

;

**run**;

/\* (가), (나), (다) \*/

**proc** **gplot** data=sasadv.ex7\_2;

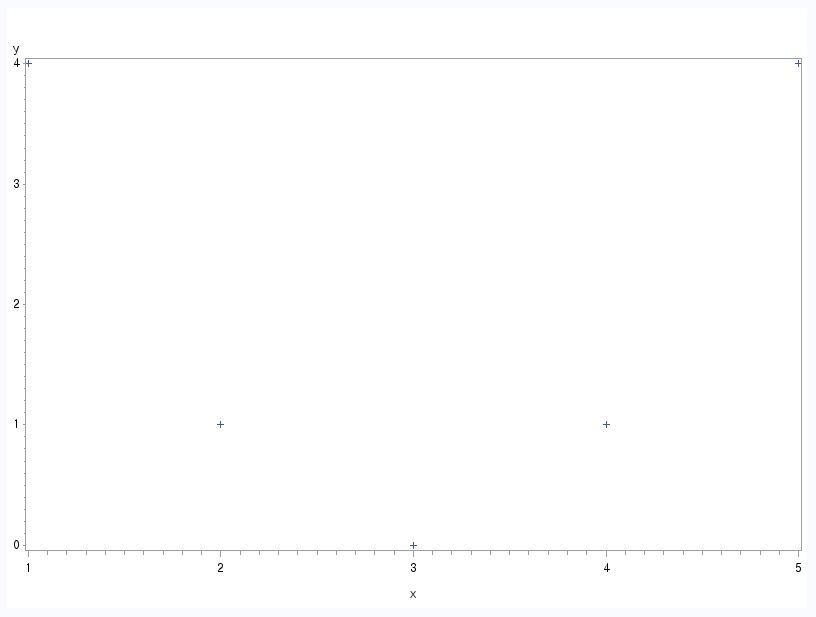
plot y\*x;

**run**;

**proc** **corr** data=sasadv.ex7\_2 pearson;

var x y;

**run**;



(가) 변수 x와 y는 이차곡선의 관계를 나타내므로, 피어슨 상관계수의 값은 ‘0’이라고 유추해낼 수 있고 실제로 계산된 값도 ‘0’으로 나왔다. 이 값이 의미하는 것은 두 변수 간의 연관성이 없다는 것이 아니라 선형적 관계가 없다는 것이며 비선형 관계(이차곡선 관계)를 나타낸다는 의미이다.

(나) 귀무가설 “모상관계수=0”에 대한 검정 결과, 표본상관계수가 0이 나왔으므로 귀무가설을 기각하지 못한다. 이는 두 변수 x와 y가 서로 비선형 관계(이차곡선 관계)를 나타내기 때문이다.

(다) 위의 데이터는 이변량 정규분포로부터 얻은 표본이라고 할 수 없다. 그 이유는 데이터의 산점도를 그려본 결과, 두 변수 x와 y는 서로 선형적인 관계가 아니기 때문이다.

/\* (라) \*/

**proc** **corr** data=sasadv.ex7\_2 spearman;

var x y;

**run**;



(라) 데이터의 정규성이 인정되지 않을 경우에는 비모수적 방법인 ‘스피어만 순위상관계수’를 사용해야 한다.

**<연습문제 7-4>**

**data** sasadv.ex7\_4;

input x y @@;

cards;

45 30 37 55 22 28

43 52 34 25 58 44

46 45 30 30 60 61

49 38 31 40 52 58

50 62 27 45 26 17

;

**run**;

/\* (가) \*/

**proc** **gplot** data=sasadv.ex7\_4;

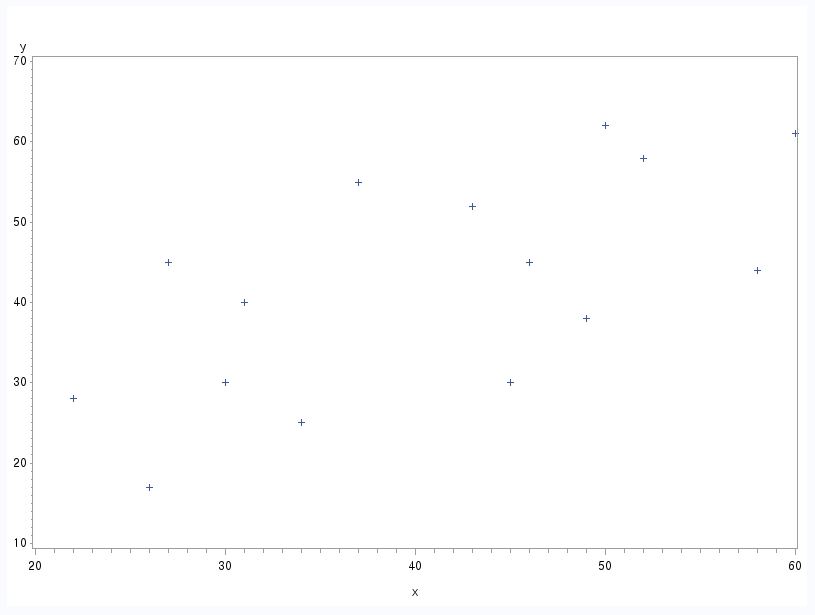
plot y\*x;

**run**;

**proc** **corr** data=sasadv.ex7\_4 pearson;

var x y;

**run**;





(가) 주어진 데이터의 산점도를 그려보면 변수 x와 y는 약한 선형적 관계를 보이는 것을 알 수 있고, 표본상관계수의 값은 0.64828로 나왔다.

/\* (나) \*/

**data** sasadv.n\_ex7\_4;

set sasadv.ex7\_4;

if x>=**27** & x<=**37** then delete;

**run**;

**proc** **gplot** data=sasadv.n\_ex7\_4;

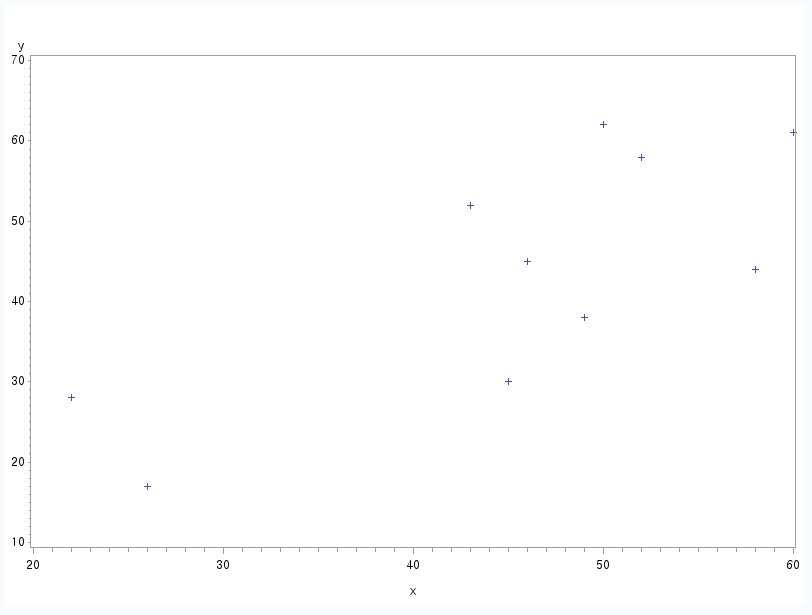
plot y\*x;

**run**;

**proc** **corr** data=sasadv.n\_ex7\_4 pearson;

var x y;

**run**;





(나) 주어진 데이터에서 27세에서 37세 사이의 5명을 제외한 나머지 10명의 데이터를 이용하여 산점도를 그려보면, 다음과 같이 조금 더 선형적 관계에 가까워 진 것을 확인할 수 있다. 또한 표본상관계수의 값도 0.75751로 조금 더 높아진 것을 확인할 수 있다.

(다) (가)에서 구한 표본상관계수의 값은 0.64828로 (나)에서 구한 표본상관계수의 값인 0.75751보다 낮은 값이다. 이는 원데이터에서 27세에서 37세 사이의 5명을 제외한 나머지 10명의 데이터를 이용하면 변수 x와 y의 선형적 관계가 더욱 뚜렷해짐을 의미한다.