제 11장 상관과 회귀분석

설정된 생산조건에서 품질을 추정하는 경우 품질과 생산조건을 함수관계로 나타낼 수 있다. (x, z)을 생산조건 y를 품질특성이라 하면 함수의 일반적인 표현은

$$y = f(x)$$
$$y = f(x, z)$$

만일 이들이 직선관계에 있다면

$$y = ax + b$$
$$y = ax + bz + c$$
$$y = ax + bz + cxz + d$$

여기서 이 수식이 어느 공장의 생산조건에 따른 품질특성을 나타내는 것이라면 x나 z은 생산조건, y는 그에 따른 품질특성의 관계로 해석할 수 있다. 이때 수학적으로 x나 z를 독립변수, y를 종속변수라 하고, 독립변수가 하나일 때 이러한 인과관계를 찾아내는 것을 단순회귀분석(simple regression analysis), 둘 이상일 때 중회귀분석 (multiple regression analysis)이라 한다.

단순회귀분석(simple regression analysis): 종속, 독립변수가 각각 하나인 관계식
중회귀분석(multiple regression analysis): 독립변수가 하나 이상인 관계식

11.1 상관분석

x가 증가하면 y가 따라서 증가하거나 감소할 때 이 둘은 상관이 있다고 하고, 두 확률변수 간 의상호관련성 정도에 관한 통계적 분석방법을 상관분석(correlation analysis)이라 하며, 관련성 정도의 측도를 상관계수(coefficient of correlation)라 한다.

회귀분석에서 독립변수는 확정변수(조절되는 확률변수가 아닌 주어진 값)인 반면, 상관분석에서 두 변수는 모두 확률변수이다. 일반적으로 두 개의 확률변수 x, y간에는 상관계수만 구하는 것이 아니고 회귀직선에 의하여 x와 y의 정량적인 관계도 함께 구한다. 이처럼 상관과 회귀문제에 대한 분석방법은 매우 유사하지만 엄밀하게 말하면 이론적 배경이 다른 것이어서 주의 깊게 사용하여야 한다.

11.1.1 산점도

두 확률변수 X, Y가 서로 대응관계에 있는 측정치 (x, y)의 상관관계를 알고자 할 때 크기 n의 확률표본을 취한 후 대응 측정치간에 어떠한 관계가 있는지 산점도를 그려보는 것이 좋다.

산점도(scatter diagram): n개의 짝지어진 (x, y) 표본을 그래프 용지에 점으로 나타내어 x, y 사이의 관계를 그림으로 보는 것.

산점도를 그릴 때 유의할 점은 이상점이 있으면 원인을 조사하여 제거하고, 원인이 분명치 않으면 그대로 둔 상태에서 해석한다.

<산점도를 그린 후 검토 사항>

(1) 점들이 산재해 있는 모양에 대해 x와 y가 양 또는 음의 관계인지를 파악한다.

(2) x와 y가 직선관계에 있는지를 파악한다. 이때 직선이 아니면 상관계수는 의미가 없다.

(3) 이상점이 발견되면 원인을 찾아 개선하고 수정한다.

(4) 점들이 뚜렷하게 주기 및 경향 등을 파악한다.

11.1.2 상관계수(Correlation coefficient)

두 변수간에 상관관계를 측정하는 측도가 상관계수이다. 모상관계수는 ρ로, 표본상관계수는 γ로 표시한다.두 확률변수 x, y가 정규분포를 따른다고 할 때

모상관계수: $\rho = \dfrac{Cov(x,y)}{\sqrt{Var(x)Var(y)}} = \dfrac{\sigma_{xy}}{\sigma_x \sigma_y}$

여기서 $Cov(x,y) = \dfrac{1}{n}\sum (x_i - \bar{x})(y_i - \bar{y}) = \dfrac{1}{n}\sum x_i y_i - \bar{x}\bar{y} = E(xy) - E(x)E(y)$

상관계수의 범위: $-1 \le \rho \le 1$

현실적으로 모상관계수를 구하는 것은 불가능하므로 모집단으로부터 표본을 추출하여 다음과 같이 표본에 대한 상관계수를 구한다.

표본상관계수: $r = \dfrac{S_{xy}/(n-1)}{\sqrt{[S_{xx}/(n-1)][S_{yy}/(n-1)]}} = \dfrac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$

여기서 표본상관 계수를 구하기 위하여 변동(분산)은 다음과 같이 구할 수 있다.

변동의 계산:

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

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2$$

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}(\sum_{i=1}^{n} y_i)^2$$

상관계수의 성질

(1) 범위: $-1 \le r \le 1$

(2) $r = -1$: 역상관(음의 상관관계)

(3) $r = 0$: 무상관

(4) $r = 1$: 정상관(양의 상관관계)

(5) x, y의 상관관계가 r일 때 $ax+b$, $cx+d$의 상관계수는 $ac > 0$이면 $r > 0$이고 $ac < 0$이면 $r < 0$이다.

SPSS 통계처리문제(Pearson 상관계수)

[보기 11_1]다음과 같은 자료에서두 변수 x와 y에 대한 상관계수와 공분산을 구하여라.

표 [11-1] 상관계수용 표

x	1	2	3	4	5	6	7
y	6	7	9	13	11	15	16

(풀이) 다음의 계산용 표에 의해 필요한 계산을 하도록 하자.

표 [11-2] 상관계수 계산용 표

id	x	y	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})(y_i-\bar{y})$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$
1	1	6	-3	-5	15	9	25
2	2	7	-2	-4	8	4	16
3	3	9	-1	-2	2	1	4
4	4	13	0	2	0	0	4
5	5	11	1	0	0	1	0
6	6	15	2	4	8	4	16
7	7	16	3	5	15	9	25
합계	28	77	0	0	48	28	90
평균	4.0	7.0					

$$S_{xy}=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=48$$

$$S_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2=28$$

$$S_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2=90$$

공분산: $Cov(x,y)=V_{xy}=\frac{1}{n-1}S_{xy}=\frac{1}{6}(48)=8$

상관계수: $\gamma=\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}=\frac{48}{\sqrt{(28)(90)}}=0.9562$

SPSS 통계처리[11_1_상관계수.sav]

분석>상관분석>이변량상관계수

보조창이 뜨면 변수 [x]와 [y]를 **변수**로 옮기고 **상관계수**에서 Pearson을 선택

옵션을 눌러 평균과 표준편차와 교차곱 편차와 공분산을 선택

계속>확인

상관계수 결과

기술통계량

	평균	표준편차	N
x	4.00	2.160	7
y	11.00	3.873	7

상관계수

		x	y
x	Pearson 상관계수	1	.956**
	유의확률 (양쪽)		.001
	제곱합 및 교차곱	28.000	48.000
	공분산	4.667	8.000
	N	7	7
y	Pearson 상관계수	.956**	1
	유의확률 (양쪽)	.001	
	제곱합 및 교차곱	48.000	90.000
	공분산	8.000	15.000
	N	7	7

**. 상관계수는 0.01 수준(양쪽)에서 유의합니다.

11. 1.3 순위상관계수

변량이 정규분포를 따르지 않고 순위로 평가할 수 있을 때 비모수적 방법에 의하여 두 변량간의 상관관계를 추정하는 경우 Spearman 순위상관계수(rank correlation coefficient)가 있다. 이 방법은 두 변수의 측정치를 각각 크기의 순서에 따라 번호를 매길 수 있기만 하면 되며, 두 변량에 각각 x_i, y_i를 정하고 순위 사이의 상관관계를 계산한다. 이때 Spearman 순위상관계수 r_s 는 다음과 같이 구한다.

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

SPSS 통계처리문제(Spearman 순위상관계수)

[보기 11_2] 영어와 수학의 두 과목에서 10명이 성적을 다음과 같은 순위로 받았다. 영어와 수학 성적순위에 대한 순위상관계수를 구하여라.

표 [11-3] 순위상관계수 계산용 표

영어성적 순위(x)	1	2	3	4	5	6	7	8	9	10
수학성적 순위(y)	4	3	5	2	6	1	10	7	8	9

(풀이) 두 순위의 차: −3, −1, −2, 2, −1, 5, −3, 1, 1, 1

$$r_s = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} d_i^2 = 1 - \frac{6}{10(10^2-1)} [(-3)^2 + (-1)^2 + \cdots + (1)^2] = 0.6606$$

SPSS 통계처리[11_2_순위상관계수.sav]

분석>상관분석>이변량상관계수

보조창이 뜨면 변수 [x]와 [y]를 **변수**로 옮기고 **상관계수**에서 Spearman을 선택

계속>확인

비모수상관

상관계수

			영어순위	수학순위
Spearman의 rho	영어순위	상관계수	1.000	.661*
		유의확률(양측)	.	.038
		N	10	10
	수학순위	상관계수	.661*	1.000
		유의확률(양측)	.038	.
		N	10	10

*. 상관 유의수준이 0.05입니다(양측).

11.1.4 상관계수의 추정과 검정

x, y 가 이변량 정규분포라 하면 확률밀도함수는

$$f(x,y)=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}e^{-\frac{1}{2}Q_{xy}}$$

$$Q_{xy}=\frac{1}{1-\rho^2}[(\frac{x-\mu_x}{\sigma_x})^2-2\rho\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}+(\frac{y-\mu_y}{\sigma_y})^2]$$

이에 대한 표본상관계수의 분포는 다음과 같다.

(1) $\rho=0$ 일 때

표본 상관계수 r 은 다음과 같이 자유도 $\phi=n-2$ 의 t 분포를 따르는 함수에 적용된다.

$$t=\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

이것을 이용하여 귀무가설 H_0 : $\rho=0$ 의 검정을 행할 수 있다.

(2) $\rho\neq 0$ 일 때

표본이 충분히 크면 정규분포를 하므로 r 을 다음과 같이 변환하여 사용한다.

$$z=\frac{1}{2}\ln\frac{1+\rho}{1-\rho}=\tanh^{-1}\rho$$

z 의 기대치: $E(z)=\frac{1}{2}\ln\frac{1+\rho}{1-\rho}=\tanh^{-1}\rho$

표준편차: $D(z)=\frac{1}{\sqrt{n-3}}$

<모상관계수 ρ에 대한 검정>

(1) 가설 H_0: $\rho=0$, H_1: $\rho\neq 0$

(2) 표본크기 n에서 x,y를 측정하여 r을 계산

(3) 검정통계량: $T=\dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$

(4) t 표에서 $t(n-2,\dfrac{\alpha}{2})$를 구함.

(5) 판정: $|T|\geq t(\phi,\dfrac{\alpha}{2})$, (여기서 $\phi=n-2$)이면 H_0 기각하고 H_1 채택. 즉 x,y는 유의수준 α에서 서로 상관이 있다고 말할 수 있다.

<모상관계수 ρ의 구간추정>

(1) 모상관계수의 점추정은 표본상관계수 γ를 이용한다.

(2) 모상관계수의 신뢰한계(신뢰구간): $\dfrac{1}{2}\ln\dfrac{1+\gamma}{1-\gamma}-\dfrac{z_{\alpha/2}}{\sqrt{n-3}}\leq\rho\leq\dfrac{1}{2}\ln\dfrac{1+\gamma}{1-\gamma}+\dfrac{z_{\alpha/2}}{\sqrt{n-3}}$

SPSS 통계처리문제(상관계수 검정)

[보기 11_3] 다음은 남녀 한 쌍씩 임의로 8조를 추출하여 얻은 통계학 성적이다.

(a) 공분산을 구하여라.

(b) 통계학 성적에 관한 남녀간에 상관유무를 유의수준 $\alpha=0.05$에서 검정하라.

표 [11-4] 남녀 각 쌍의 통계학 성적

남자(x)	79	78	83	86	85	87	90	94
여자(y)	84	70	88	75	81	75	88	90

(풀이) 검정하고자 하는 가설은 다음과 같다.

귀무(영)가설 H_o: $\rho=0$ (남녀 사이의 통계학 성적은 상관관계가 없다).

대립(연구)가설 H_1: $\rho\neq 0$ (남녀 사이의 통계학 성적은 상관관계가 있다).

이 가설을 검정하기 위하여 다음의 계산용 표 [13-4]를 만들고 변동을 계산한다.

$$S_{xy}=\sum_{i=1}^{8}(x_i-\bar{x})(y_i-\bar{y})=142.25$$

$$S_{xx}=\sum_{i=1}^{8}(x_i-\bar{x})^2=199.5$$

$$S_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2=379.875$$

(a) 공분산: $Cov(x,y)=V_{xy}=\dfrac{1}{n-1}S_{xy}=\dfrac{1}{7}(142.25)=20.32$

(b) 상관계수: $\gamma = \dfrac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \dfrac{142.25}{\sqrt{(199.5)(379.875)}} = 0.5167$

표 [11-5] 변동들을 계산하기 위한 표

	x(남자)	y(여자)	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
	79	84	-6.25	2.625	-16.40625	39.0625	6.890625
	78	70	-7.25	-11.375	82.46875	52.5625	129.390625
	83	88	-2.25	6.625	-14.90625	5.0625	43.890625
	86	75	0.75	-6.375	-4.78125	0.5625	40.640625
	85	81	-0.25	-0.375	0.09375	0.0625	0.140625
	87	75	1.75	-6.375	-11.15625	3.0625	40.640625
	90	88	4.75	6.625	31.46875	22.5625	43.890625
	94	90	8.75	8.625	75.46875	76.5625	74.390625
합계	682	651	0	0	142.25	199.5	379.875
평균	85.250	81.375					

검정통계량: $T = \dfrac{\gamma\sqrt{n-2}}{\sqrt{1-\gamma^2}} = (0.5167)\sqrt{\dfrac{8-2}{1-(0.5167)^2}} = 1.4783$

※ t-분포:**http://www.statdistributions.com/t/**

(1) $t(\phi, \dfrac{\alpha}{2}) = t(6, 0.025)$ 값 구하기. 여기서 $\phi = n-2 = 6$, $\alpha = 0.05$.

(a) [p-value] box에 0.05 입력.

(b) [d.f.]box에 6 입력.

(c) [two tails]를 선택.

[t-value] box에서 $t(6, 0.025) = 2.447$ 를 얻을 것이다.

(2) 검정통계량 $T = 1.4783$ 유의확률 구하기

(a) [d.f.]box에 6 입력.

(b) [two tails]를 선택.

(c) [t-value] box에 1.478 입력.

[p-value] box에서 $P(T = 1.478) = 0.190$ 을 얻을 것이다.

검정결과: $T = 1.478 < t(7, 0.025) = 2.447$ 이므로 영가설(귀무가설)이 채택된다.즉, $\rho = 0$ 이므로 남녀의 통계학 성적은 상관관계가 없다(유의하지 않다)고 결론을 내릴 수 있다. 확률로 보면 검정통계량의 유의확률 0.190 은 유의수준 $\alpha = 0.05$보다 크기 때문에 H_o가 채택된다.

SPSS 통계처리[11_3_통계성적.sav]

분석>상관분석>이변량상관계수

보조창이 뜨면 변수 [x]와 [y]를 **변수**로 옮기고 **상관계수**에서 Pearson을 선택

옵션을 눌러 평균과 표준편차와 교차곱 편차와 공분산을 선택

계속>확인

상관계수 결과

기술통계량

	평균	표준편차	N
남자성적	85.25	5.339	8
여자성적	81.38	7.367	8

상관계수

		남자성적	여자성적
남자성적	Pearson 상관계수	1	.517
	유의확률 (양쪽)		.190
	제곱합 및 교차곱	199.500	142.250
	공분산	28.500	20.321
	N	8	8
여자성적	Pearson 상관계수	.517	1
	유의확률 (양쪽)	.190	
	제곱합 및 교차곱	142.250	379.875
	공분산	20.321	54.268
	N	8	8

SPSS 통계처리문제(산포도 및 상관계수 검정)

[보기 11_4]다음의 표는 표본도시별 교통사고 건수, 차량대수 및 교통경찰관 수이다. 산점도 및 사고건수와 차량대수, 사고건수와 경찰관수의 상관계수와 검정을 하여라.

표 [11-6] 도시별 교통사고 관련획득 자료

표본도시	사고 건수	차량대수	경찰관 수
1	1	4	20
2	4	10	6
3	5	15	2
4	4	12	8
5	3	8	9
6	4	16	8
7	2	5	12
8	1	7	15
9	4	9	10
10	2	10	10

표 [11-7] 변동들을 계산하기 위한 표

도시번호	x	y	z	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$	$(z_i-\bar{z})^2$	교차곱1	교차곱2	교차곱3
1	1	4	20	4	31.36	100	11.2	−20	−56
2	4	10	6	1	0.16	16	0.4	−4	−1.6
3	5	15	2	4	29.16	64	10.8	−16	−43.2
4	4	12	8	1	5.76	4	2.4	−2	−4.8
5	3	8	9	0	2.56	1	0	0	1.6
6	4	16	8	1	40.96	4	6.4	−2	−12.8
7	2	5	12	1	21.16	4	4.6	−2	−9.2
8	1	7	15	4	6.76	25	5.2	−10	−13
9	4	9	10	1	0.36	0	−0.6	0	0
10	2	10	10	1	0.16	0	−0.4	0	0
합계	30	96	100	18	138.4	218	40	−56	−139
평균	3.0	9.6	10.0						

※ 여기서 교차곱1: $(x_i-\bar{x})\,(y_i-\bar{y})$, 교차곱2: $(x_i-\bar{x})\,(z_i-\bar{z})$, 교차곱3: $(y_i-\bar{y})\,(z_i-\bar{z})$

(풀이) 검정하고자 하는 가설은 다음과 같다.

귀무(영)가설 H_o: $\rho_{xy}=0$ (교통사고 건수와 차량 수는 상관관계가 없다).

$\rho_{xz}=0$ (교통사고 건수와 경찰관 수는 상관관계가 없다).

$\rho_{yz}=0$ (차량 수와 경찰관 수는 상관관계가 없다).

이들의 반대가 대립가설 H_1이다.

가설을 검정하기 위하여 다음의 계산을 한다.

$$S_{xy}=\sum_{i=1}^{10}(x_i-\bar{x})(y_i-\bar{y})=40,\ S_{xz}=\sum_{i=1}^{10}(x_i-\bar{x})(z_i-\bar{z})=-56,\ S_{yz}=\sum_{i=1}^{10}(y_i-\bar{y})(z_i-\bar{z})=-139$$

$$S_{xx}=\sum_{i=1}^{10}(x_i-\bar{x})^2=18,\quad S_{yy}=\sum_{i=1}^{10}(y_i-\bar{y})^2=138.4,\quad S_{zz}=\sum_{i=1}^{10}(z_i-\bar{z})^2=218$$

(a) 공분산

$$Cov(x,y)=V_{xy}=\frac{1}{n-1}S_{xy}=\frac{1}{9}(40)=4.44$$

$$V_{xz}=\frac{1}{n-1}S_{xz}=\frac{1}{9}(-56)=-6.22$$

$$V_{yz}=\frac{1}{n-1}S_{yz}=\frac{1}{9}(-139)=-15.44$$

(b) 상관계수

$$\gamma_{xy}=\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}=\frac{40}{\sqrt{(18)(138.4)}}=0.8014$$

$$\gamma_{xz}=\frac{S_{xz}}{\sqrt{S_{xx}S_{zz}}}=\frac{-56}{\sqrt{(18)(218)}}=-0.8940$$

$$\gamma_{yz}=\frac{S_{yz}}{\sqrt{S_{yy}S_{zz}}}=\frac{-139}{\sqrt{(138.4)(218)}}=-0.8002$$

(c) 검정통계량

$$T_{xy}=\frac{\gamma_{xy}\sqrt{n-2}}{\sqrt{1-\gamma_{xy}^2}}=(0.8014)\sqrt{\frac{10-2}{1-(0.8014)^2}}=3.790$$

$$T_{xz}=\frac{\gamma_{xz}\sqrt{n-2}}{\sqrt{1-\gamma_{xz}^2}}=(-0.8940)\sqrt{\frac{10-2}{1-(-0.8940)^2}}=-5.643$$

$$T_{yz}=\frac{\gamma_{yz}\sqrt{n-2}}{\sqrt{1-\gamma_{yz}^2}}=(-0.8002)\sqrt{\frac{10-2}{1-(-0.8002)^2}}=-3.774$$

※ t-분포:**http://www.statdistributions.com/t/**

(1) $t(\phi,\frac{\alpha}{2})=t(8,0.025)$ 값 구하기. 여기서 $\phi=n-2=8$, $\alpha=0.05$.

(a) [p-value] box에 0.05 입력.

(b) [d.f.]box에 6 입력.

(c) [two tails]를 선택.

[t-value] box에서 $t(8,0.025)=2.306$를 얻을 것이다.

(2) 검정통계량 T_{xy}, T_{xz}, T_{yz} 유의확률 구하기

(a) [d.f.]box에 8 입력.

(b) [two tails]를 선택.

(c) [t-value] box에 3.790, 5.643, 3.774를 차례로 입력.

[p-value] box에서 다음을각각 얻을 것이다.

$P(T_{xy}=3.790)=0.005$

$P(T_{xz}=5.643)=0.00$

$P(T_{yz}=3.774)=0.005$

검정결과: T_{xy}(또는 T_{xz} 또는 T_{yz})$<t(8,0.025)=2.306$이므로 모든 영가설(귀무가설)이 기각되고대립가설 H_1를이 채택된다. 즉 교통사고 건수와 차량 수, 교통사고 건수와 경찰관 수 그리고 차량 수와 경찰관 수는 모두 상관관계가 있다. 확률로 보면 세 검정통계량의 유의확률은 모두 유의수준 $\alpha=0.05$보다 작기 때문에 H_1이 채택되므로 상관관계가 모두 유의하다.

SPSS 통계처리[11_4_교통사고.sav]

(a) 산점도

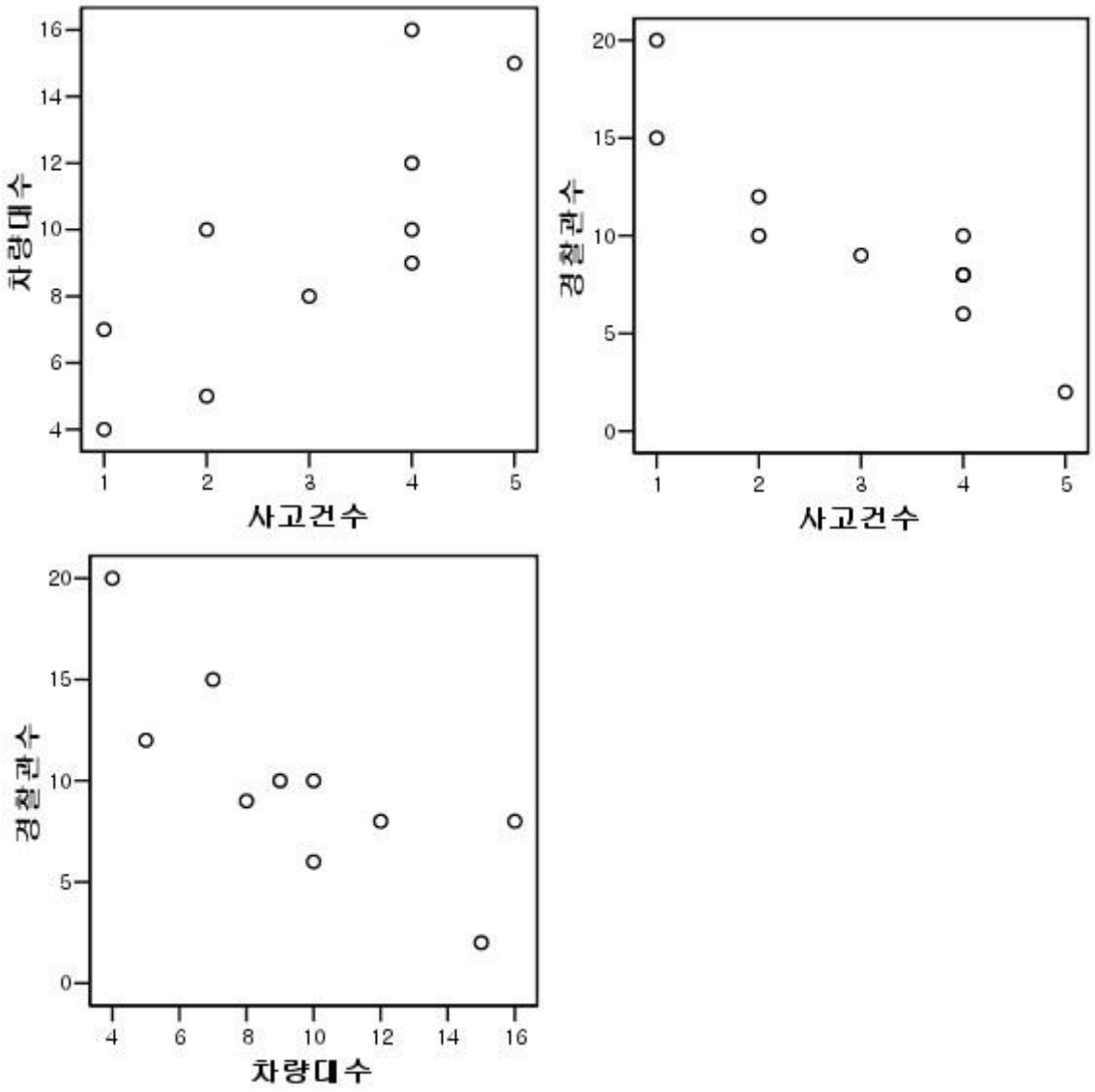
그래프>산점도>단순선택후정의를 누름. 그리고 아래와 같은 방법으로 세 개의 산포도를 획득.

(1) 사고건수[x]를 **X축**, 차량대수[y]를 **Y축**으로 각각 이동. **확인**.

(2) 사고건수[x]를 **X축**, 경찰관수[z]를 **Y축**으로 각각 이동. **확인**.

(3) 차량대수[y]를 **X축**, 경찰관수[z]를 **Y축**으로 각각 이동. **확인**.

얻은 산포도들은 다음과 같다.



(b) 상관계수

분석>상관분석>이변량상관계수

보조창이 뜨면 변수 [x], [y], [z]를 **변수**로 옮기고 **상관계수**에서 Pearson을 선택

옵션을 눌러 평균과 표준편차와 교차곱 편차와 공분산을 선택

계속>확인

상관계수 결과

기술통계량

	평균	표준편차	N
사고건수	3.00	1.414	10
차량대수	9.60	3.921	10
경찰관수	10.00	4.922	10

상관계수

		사고건수	차량대수	경찰관수
사고건수	Pearson 상관계수	1	.801(**)	-.894(**)
	유의확률 (양쪽)		.005	.000
	제곱합및교차곱	18.000	40.000	-56.000
	공분산	2.000	4.444	-6.222
	N	10	10	10
차량대수	Pearson 상관계수	.801(**)	1	-.800(**)
	유의확률 (양쪽)	.005		.005
	제곱합및교차곱	40.000	138.400	-139.000
	공분산	4.444	15.378	-15.444
	N	10	10	10
경찰관수	Pearson 상관계수	-.894(**)	-.800(**)	1
	유의확률 (양쪽)	.000	.005	
	제곱합및교차곱	-56.000	-139.000	218.000
	공분산	-6.222	-15.444	24.222
	N	10	10	10

** 상관계수는 0.01 수준(양쪽)에서유의합니다.

검정: SPSS 검정결과는 이론에서 계산한 것과 동일하므로 분석은 생략한다.

[보기 11-5] 임의로 추출한 표본크기 $n=10$ 에서 구한 표본상관계수는 $\gamma=0.89$ 라고 한다. 모상관계수 ρ 의 95% 신뢰구간을 구하여라.

(풀이) 신뢰구간: $\frac{1}{2}\ln\frac{1+\gamma}{1-\gamma}-\frac{z_{\alpha/2}}{\sqrt{n-3}}\le\rho\le\frac{1}{2}\ln\frac{1+\gamma}{1-\gamma}+\frac{z_{\alpha/2}}{\sqrt{n-3}}$

$\frac{1}{2}\ln\frac{1+\gamma}{1-\gamma}=\frac{1}{2}\ln(\frac{1+0.89}{1-0.89})=1.422$, $\frac{z_{\alpha/2}}{\sqrt{n-3}}=\frac{1.96}{\sqrt{7}}=0.7408$

$1.422-0.7408\le\rho\le1.422+0.7408 \rightarrow 0.593\le\rho\le0.974$

11.2 단순회귀 분석

단순회귀 분석은 오직 한 개의 독립변수를 고려하여 두 변수간의 상호관련성을 하나의 함수식으로 나타낸 것이다.

11.2.1 단순회귀 모형

함수식: $y = ax + b$

여기서 a 는 직선의 기울기, b 는 절편이다.

자료는 직선으로부터 오차가 존재한다. 따라서 각 점은

$y_i = ax_i + b + \varepsilon_i$

여기서 ε_i 는 실험오차로서로 독립, 평균은 0, 분산은 σ_e^2 인 정규분포 $N(0, \sigma_e^2)$ 을 따른다.

오차의 특성

(1) 오차 ε_i 는 서로 독립

(2) ε_i 의 평균은 0

(3) ε_i 는 모든 x 에 대해 동일한 분산 σ_e^2 을 갖음.

(4) ε_i 는 정규분포

11.2.2 최소제곱법

ε_i 의 변동을 최소로 하는 a 와 b 를 찾는 것이 목적이다.

$\varepsilon_i = y_i - (ax_i + b)$, 여기서 $i = 1, 2, \cdots, n$

$$S_E = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [y_i - (ax_i - b)])^2$$

오차 제곱합을 최소로하는 α 와 β 구하기

$$\frac{\partial S_E}{\partial \alpha} = -2\sum_i (y_i - ax_i - b) = 0$$

$$\frac{\partial S_E}{\partial \beta} = -2\sum_i x_i (y_i - ax_i - b) = 0$$

이들 식을 다시 쓰면 아래의 정규방정식(normal equations)이 된다.

$$a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i$$

$$a\sum_{i=1}^{n} x_i + nb = \sum_{i=1}^{n} y_i$$

a 와 b 에 대하여 풀면

$$a = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{b} = \bar{y} - a\bar{x}$$

따라서 구한 모회귀선의 추정식은 다음과 같다.

$$y_i = \bar{y} + a(x_i - \bar{x})$$

종합: 최소제곱 추정 값

(1) 회귀선: $y_i = \bar{y} + a(x_i - \bar{x})$

(2) 모수 a (직선의 기울기): $a = \frac{S_{xy}}{S_{xx}}$

(3) 모수 b (절편): $\hat{b} = \bar{y} - a\bar{x}$

11.2.3 최소제곱 추정량의 성질

i) a 의 추정과 검정

(1) 기대값: $E(a) = a$

(2) 분산: $Var(a) = \frac{\sigma^2}{S_{xx}}$

(3) a 의 분포: $a \sim N(a, \frac{\sigma^2}{S_{xx}})$

(4) 잔차평균제곱: $MSE = V_E = \frac{S_E}{\phi}$

여기서 ($\phi = n-2$), $S_E = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2$

$E(V_E) = \sigma^2$

(5) 분포: $\frac{a-a}{\sqrt{V_E / S_{xx}}} \sim t(\phi, \alpha/2)$

(6) $(1-\alpha)$의 신뢰구간: $a - t(\phi, \frac{\alpha}{2})\sqrt{\frac{V_E}{S_{xx}}} \le a \le a + t(\phi, \frac{\alpha}{2})\sqrt{\frac{V_E}{S_{xx}}}$

(7) 검정통계량: $T = \frac{a - a_o}{\sqrt{V_E / S_{xx}}}$

검정방법:

귀무가설	대립가설	유의수준 α 인 기각역
H_o: $a = a_o$	H_1: $a \neq a_o$	$T \ge t(\phi, \frac{\alpha}{2})$

(ii) b 의 추정과 검정

(1) 기대값: $E(\hat{b}) = b$

(2) 분산: $Var(\hat{b}) = [\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}]\sigma^2$

(3) $1-\alpha$ 의 신뢰구간: $\hat{b} - t(\phi, \frac{\alpha}{2})\sqrt{[\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}]V_E} \le b \le \hat{b} + t(\phi, \frac{\alpha}{2})\sqrt{[\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}]V_E}$

(3) 검정통계량: $T = (\hat{b} - b_o) / \sqrt{[\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}]V_E}$

검정방법:

귀무가설	대립가설	유의수준 α 인 기각역
H_o: $b = b_o$	H_1: $b \neq b_o$	$T \ge t(\phi, \frac{\alpha}{2})$ (여기서 $\phi = n-2$)

11.2.3 단순회귀 분산분석

잔차(오차)제곱합: $S_E = \sum_{i=1}^{n}[y_i - (ax_i + b)]^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 - a^2\sum_{i=1}^{n}(x_i - \bar{x})^2$

$$S_E = S_{yy} - a^2 S_{xx} = S_{yy} - \frac{(S_{xy})^2}{S_{xx}}$$

y 의 총제곱합: $S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$

회귀에 의한 제곱합: $S_R = \frac{S_{xy}^2}{S_{xx}}$ $S_R = \frac{S_{xy}^2}{S_{xx}}$

[총 제곱합]=[회귀에 의한 제곱합]+[잔차에 의한 제곱합]: $S_{yy} = S_R + S_E$

S_{yy} 의 자유도 ϕ_T: $\phi_T = n-1$

S_R 의 자유도 ϕ_R: $\phi_R = 1$

S_E 의 자유도 ϕ_E: $\phi_E = \phi_T - \phi_R = n-2$

H_o: $a = a_o$ 의 검정은 $F_t \ge F(\phi_R, \phi_E; \alpha)$ 면 H_o 기각.

표 [11-8] 단순회귀의 분산분석표

요인	제곱합	자유도	평균제곱	F_t	$F(\alpha)$
회귀 잔차	$S_R = \frac{S_{xy}^2}{S_{xx}}$ $S_e = S_{yy} - S_R$	$\phi_R = 1$ $\phi_E = n-2$	$V_R = \frac{S_R}{\phi_R}$ $V_E = \frac{S_E}{\phi_E}$	$\frac{V_R}{V_E}$	$F(\phi_R, \phi_E; \alpha)$
합계	S_{yy}	$\phi_T = n-1$			

결정계수: $r^2 = \frac{S_R}{S_{yy}} = 1 - \frac{S_E}{S_{yy}}$

상관계수: $r = \pm\sqrt{r^2} = \pm\sqrt{1 - \frac{S_E}{S_{yy}}}$

SPSS 통계처리문제

[보기 11_5]어떤 화학제품의 첨가물(x)과 수율(y)에 대한 데이터가 다음과 같다.

(a) 최소제곱법을 사용하여 회귀직선을 추정하라.(b) 분산분석표를 작성하고 유의수준 $\alpha = 0.05$에서 회귀식의 유의성 검정과 결정계수 r^2을 구하여라. (c) 기울기 a와 절편 b의 95% 신뢰구간과 검정통계량을 구하여라.

표 [11-9] 첨가물과 수율

x(첨가물)	1.7	2.3	2.8	3.5	4.2	4.9
y(수율)	33	35	50	45	66	63

(풀이) 계산을 용이하게 하기 위해 다음의 표를 작성한다.

표 [11-10] 회귀선 계산용 표

id	x	y	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})$	$(y - \bar{y})^2$	$(x_i - \bar{x})\,(y_i - \bar{y})$
1	1.7	33	-1.53333	2.351101	-15.6667	245.4455	24.02222
2	2.3	35	-0.93333	0.871105	-13.6667	186.7787	12.75554
3	2.8	50	-0.43333	0.187775	1.3333	1.777689	-0.57776
4	3.5	45	0.26667	0.071113	-3.6667	13.44469	-0.9778
5	4.2	66	0.96667	0.934451	17.3333	300.4433	16.75558
6	4.9	63	1.66667	2.777789	14.3333	205.4435	23.88888
합계	19.4	292	0	7.1933	0	953.3333	75.86667
평균	3.233	48.667					

$$S_{xy} = \sum_{i=1}^{6}(x_i - \bar{x})(y_i - \bar{y}) = 75.867$$

$$S_{xx} = \sum_{i=1}^{6}(x_i - \bar{x})^2 = 7.193$$

$$S_{yy} = S_T = \sum_{i=1}^{6}(y - \bar{y})^2 = 953.333$$

$$S_R = \frac{S_{xy}^2}{S_{xx}} = \frac{(75.867)^2}{7.193} = 800.155$$

$$S_E = S_T - S_R = 953.333 - 800.155 = 153.178$$

(a)직선의 기울기: $a=\frac{S_{xy}}{S_{xx}}=\frac{75.867}{7.193}=10.5468$

절편: $\hat{b}=\bar{y}-a\bar{x}=48.667-(10.5468)(3.233)=14.5687$

회귀식(직선식): $y=10.547x+14.565$

(b) 분산분석표 및 검정 및 결정계수 r^2

$S_T=S_{yy}$의 자유도 ϕ_T: $\phi_T=n-1=6-1=5$

S_R의 자유도 ϕ_R: $\phi_R=1$

S_E의 자유도 ϕ_E: $\phi_E=\phi_T-\phi_R=n-2=6-2=4$

$V_R=\frac{S_R}{\phi_R}=\frac{800.155}{1}=800.155$

$V_E(MSE)=\frac{S_E}{\phi_E}=\frac{153.178}{4}=38.294$

회귀식의 검정통계량: $F_t=\frac{V_R}{V_E}=\frac{800.155}{38.294}=20.895$

F 분포 값: $F(\phi_R,\phi_E;\alpha)=F(1,4;0.05)=7.71$

표 [11-11] 단순회귀 분산분석표

요인	제곱합	자유도	평균제곱	F_t	$F(\alpha)$
회귀 잔차	$S_R=800.155$ $S_E=153.178$	$\phi_R=1$ $\phi_e=4$	$V_R=800.155$ $V_E=38.294$	20.895	$F(1,5;0.05)=7.699$
합계	S_{yy}	$\phi_T=n-1$			

귀무(영)가설 H_o: 기울기 $a=0$

검정: $F_t=20.985>F(1,4;0.05)=7.71$이므로 H_o는 기각된다. 즉 기울기 a는 $a=0$가 아니다.

유의확률: 자유도 $\phi_R=1$, $\phi_E=4$에서 $P(F_t=20.895)=0.01$

※ F **-분포:http://www.statdistributions.com/f/**

(1) $F(\phi_1,\phi_2;\alpha)=F(1,4;0.05)$의 값

(a) [p-value] box에 0.05 입력.

(b)[numerator d.f.]box에 1 입력.

(c) [denominator d.f.]box에 4 입력.

(d)[right tail]을 선택.

[F-value] box에서 7.699를 얻을 것이다.

(2) $F_t=20.895$의 확률(유의확률). 계산

(a) [F-value] box에 20.895 입력.

(b)[numerator d.f.]box에 1 입력.

(c) [denominator d.f.]box에 4 입력.

(d)[right tail]을 선택.

[p-value] box에서 0.01을 얻을 것이다.

결정계수: $r^2 = \frac{S_R}{S_{yy}} = \frac{800.1552}{953.333} = 0.8393$

상관계수: $r = \pm\sqrt{0.8393} = 0.9164$

신뢰구간:

(1) 기울기 a의 신뢰구간및 검정통계량

공식: $a - t(\phi_E, \frac{\alpha}{2})\sqrt{\frac{V_E}{S_{xx}}} \le a \le a + t(\phi_E, \frac{\alpha}{2})\sqrt{\frac{V_E}{S_{xx}}}$

$t(\phi_E, \frac{\alpha}{2})\sqrt{\frac{V_E}{S_{xx}}} = t(4, 0.025)\sqrt{\frac{38.294}{7.193}} = (2.777)(2.3073) = 6.4075$

기울기 a의 신뢰구간: $10.547 - 6.407 \le a \le 10.547 + 6.407 \rightarrow 4.140 \le a \le 16.954$

검정통계량: $T = \frac{a - a_o}{\sqrt{V_E / S_{xx}}} = \frac{10.547 - 0}{\sqrt{38.294 / 7.193}} = 4.571$

유의확률: $P(T = 4.571) = 0.010$

(2) 절편 b의 신뢰구간및 검정통계량

공식: $\hat{b} - t(\phi, \frac{\alpha}{2})\sqrt{[\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}]V_E} \le b \le \hat{b} + t(\phi, \frac{\alpha}{2})\sqrt{[\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}]V_E}$

$t(\phi, \frac{\alpha}{2})\sqrt{[\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}]V_E} = (2.777)\sqrt{[\frac{1}{6} + \frac{(3.233)^2}{7.193}](38.294)} = (2.777)(7.876) = 21.8710$

절편 b의 신뢰구간: $14.565 - 21.871 \le b \le 14.565 + 21.871 \rightarrow -7.306 \le b \le 36.436$

검정통계량: $T = (\hat{b} - b_o) / \sqrt{[\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}]V_E} = \frac{14.565 - 0}{7.876} = 1.8493$

유의확률: $P(T = 1.8493) = 0.138$

※ t-분포:**http://www.statdistributions.com/t/**

(1) $t(\phi, \frac{\alpha}{2}) = t(4, 0.025)$ 값.

(a) [p-value] box에 0.05 입력.

(b) [d.f.]box에 4 입력.

(c) [two tails]를 선택.

[t-value] box에서 2.777을 얻을 것이다.

(2) $T=4.571$과 $T=1.849$의 확률(유의확률)

(a) [d.f.]box에 4 입력.

(b) [two tails]를 선택.

(c) [t-value] box에 4.571과 1.849를 각각 입력.

[p-value] box에서 0.01과 0.138을 각각 얻을 것이다.

SPSS 통계처리[11_5_회귀분석.sav]

분석>회귀분석>선형

변수 [y]를 **종속변수**로 이동

변수 [x]를 **독립변수**로 이동

통계량을 눌러 **회귀계수**에서 추정값, 신뢰구간, 모형적합을 선택

계속>확인

선형회귀분석 결과

진입/제거된 변수[b]

모형	진입된 변수	제거된 변수	방법
1	x[a]	.	입력

a. 요청된 모든 변수가 입력되었습니다.

b. 종속변수: y

모형 요약

모형	R	R 제곱	수정된 R 제곱	추정값의 표준오차
1	.916[a]	.839	.799	6.188

a. 예측값: (상수), x

분산분석[b]

모형		제곱합	자유도	평균제곱	F	유의확률
1	선형회귀분석	800.151	1	800.151	20.894	.010[a]
	잔차	153.183	4	38.296		
	합계	953.333	5			

a. 예측값: (상수), x

b. 종속변수: y

계수[a]

모형		비표준화 계수		표준화 계수	t	유의확률	B에 대한 95% 신뢰구간	
		B	표준오차	베타			하한값	상한값
1	(상수)	14.565	7.877		1.849	.138	-7.303	36.434
	x	10.547	2.307	.916	4.571	.010	4.141	16.953

a. 종속변수: y

※ 회귀분석에 앞서 공부한 상관분석을 통해 상관계수를 분석하여 회귀분석 결과와 비교하자.

분석>상관분석>이변량상관계수

보조창이 뜨면 변수 [x]와 [y]를 **변수**로 옮기고 **상관계수**에서 Pearson을 선택

옵션을 눌러 평균과 표준편차와 교차곱 편차와 공분산을 선택

계속>확인

상관계수 결과

기술통계량

	평균	표준편차	N
x	3.23	1.199	6
y	48.67	13.808	6

상관계수

		x	y
x	Pearson 상관계수	1	.916*
	유의확률 (양쪽)		.010
	제곱합 및 교차곱	7.193	75.867
	공분산	1.439	15.173
	N	6	6
y	Pearson 상관계수	.916*	1
	유의확률 (양쪽)	.010	
	제곱합 및 교차곱	75.867	953.333
	공분산	15.173	190.667
	N	6	6

*. 상관계수는 0.05 수준(양쪽)에서 유의합니다.

보는 바와 같이 회귀분석과 상관분석은 많은 공통점을 가지고 있다. 그 이유는 회귀분석의 비례상수 b 와 상관분석 r 은 밀접한 관계에 있기 때문이다.

연습문제

1.회귀식 $y_i = bx_i + c + \varepsilon_i$ 에서 ε_i 는 정규분포 $N(0,\sigma^2)$ 을 따르고 독립이다. x_i 는 측정치라 할 때 다음의 자료를 이요하여 물음에 답하라.

$$n = 200 \qquad \sum_i x_i = 12.00 \qquad \sum_i y_i = 20.00$$

$$\sum_i x_i^2 = 11.22 \quad \sum_i y_i^2 = 86.00 \qquad \sum_i x_i y_i = 22.20$$

(1) 직선에 접합시킬 때 최소제곱추정값 $\hat{b}$ 와 $\hat{c}$ 를 각각 구하여라.

(2) b 에 대한 신뢰구간을 구하여라.

(3) 회귀직선의 유의여부를 알기 위하여 분산분석을 유의수준 $\alpha = 0.05$ 에서 검정하라.

2. 단순회귀모형이 x 와 y 간의 관계를 설명하는 데 적절하다고 판단될 때, 다음의 자료를 이용하여 물음에 답하라.

실험번호	1	2	3	4	5	6	7	8	9	10
촉진제항(x)	1	1	2	3	4	4	5	6	6	7
반응량(y)	2.1	2.5	3.1	3.0	3.8	3.2	4.3	3.9	4.4	4.8

(1) 단순회귀모형이 x와 y간을 설명하는 데 적절하다고 판단될 때, 단순회귀식을 최소제곱법에 의하여 구하여라.

(2) 분산분석을 작성하고 유의수준 $\alpha = 0.05$에서 F-검정을 하라. 또 결정계수 r^2을 구하고 r^2으로부터 상관계수 r을 구하여라.

(3) $x = 7$에서 $E(y)$의 95% 신뢰구간을 구하고, 가설 H_o: $\eta = 5.2$, H_1: $\eta \neq 5.2$를 검정하라.

3. 다음은 어느 자동차 수리공장에서 종업원 5명을 임의로 뽑아 경력과 일주일 동안 수리한 자동차 수를 조사한 결과이다. 다음의 물음에 답하라.

경력(x)	5	7	6	10	3
수리대수(y)	6	13	9	20	7

(1) 최소자승법에 의해서 회귀직선($y_i = bx_i + c + \varepsilon_i$을 추정하라.

(2) 오차항의 표준편차 $S_{y|x}$를 구하여라.

(3) $x = 12$일 때 y의 평균($\mu_{y|x}$)에 대한 95% 신뢰구간을 구하여라.

(4) $x = 3$일 때 y의 실제값(y_a)에 대한 95% 신뢰구간을 구하여라.

(5) 유의수준 5%에서 H_o: $b = 0$, H_1: $b > 0$에 대해 검정하라.

(6) 분산분석을 작성하고 회귀직선의 유의 여부를 검정하여라.

(7) 결정계수를 구하고 계산된 계수의 의미를 설명하여라.

(8) x와 y의 상관계수를 구하여라.

(9) 유의수준 5%에서 양의 상관관계가 있는가를 검정하라.

4. 다음은 어느 중고차 대리점에서 중고차의 사용년수(x)에 따른 가격(y)의 자료이다. 물음에 답하여라.

경력(x)	5	7	6	10	3
수리대수(y)	6	13	9	20	7

(1) S_{xx}, S_{xy}, S_{yy}를 구하여라.

(2) 최소자승법에 의해서 회귀직선($y_i = bx_i + c + \varepsilon_i$을 추정하라.

(3) 사용년수가 2년인 중고차의 평균가격은 얼마인가?

(4) $MSE(V_E)$를 구하여라.

(5) $x = 2$일 때 평균가격에 대한 95% 신뢰구간을 구하여라.

(6) $x = 2$일 때 y의 어떤 값(y_o)에 대한 95% 신뢰구간을 구하여라.

(7) 분산분석을 작성하고 회귀직선의 유의 여부를 $\alpha = 0.05$에서 검정하여라.

(8) 결정계수를 구하고 계산된 계수의 의미를 설명하여라.

(9) 유의수준 5%에서 $b > 100$이지 검정하라.

(10) b의 95% 신뢰구간을 구하여라.

(11) x, y를 확률변수로 간주하여 두 변수의 상관계수를 구하여라.

(12) 상관계수가 0보다 작은지 유의수준 5%에서 검정하라.

5. $y = 2x + 5$를 10개의 data를 만들고 역으로 상관관계를 추적하자. 이 식은 완전한 비례식이므로 이것에 의해 만든 다음의 9개 data는 그 상관계수가 $r = 1$이어야만 한다. 따라서 이것을 수식과 SPSS로 증명해 보자.

x	1	2	3	4	5	6	7	8	9
y	7	9	11	13	15	17	19	21	23

(풀이) $\sum x_i^2 = (1)^2 + (2)^2 + \cdots + (9)^2 = 285$,

$$\sum y_i^2 = (7)^2 + (9)^2 + \cdots + (23)^2 = 2265$$

$$S_{xy} = \sum x_i y_i - \frac{1}{n}(\sum x_i)(\sum y_i) = 795 - \frac{(45)(135)}{9} = 120$$

$$S_{xx} = \sum x_i^2 - \frac{1}{n}(\sum x_i)^2 = 285 - \frac{(45)^2}{9} = 60$$

$$S_{yy} = \sum y_i^2 - \frac{1}{n}(\sum y_i)^2 = 2265 - \frac{(135)^2}{9} = 240$$

$$Cov(x, y) = V_{xy} = \frac{S_{xy}}{n-1} = \frac{120}{8} = 15$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{120}{\sqrt{(60)(240)}} = 1$$

계산결과 x와 y는 $r = 1$이므로 정확히 정관계 즉 양수의 비례관계에 있다. 9개의Data는 비례식 $y = 2x + 5$에서 만들었으므로 $r = 1$ 당연한 결과이다.

검정

(1) H_0 : $\rho = 0$, H_1 : $\rho \neq 0$,

(2) $T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = (1)\sqrt{\frac{9-2}{1-(1)^2}} = \infty$

(3) $t(7 : 0.025) = 2.365$

(4) $T = \infty > t(7 : 0.025) = 2.365$이므로 영가설(귀무가설: $r = 0$라는 가설)이 기각된다.

$T = \infty$의 확률은 0이므로 대립가설 즉 x와 y는 비례관계에 있다. 즉, $r = 1$이므로 정확한 상관관계가 있다고 결론을 내릴 수 있다.