선형회귀분석(Linear Regression Analysis)

eg) 약의 복용량(x)과 약효지속시간(y)의 관계조사
공정온도(x)와 제품의 강도(y)관계조사
TOEFL점수(x)와 영어과목성적(y)과의 관계

x : 독립변수(independent variable) or 설명변수(explanatory variable)

y : 종속변수(dependent variable) or 반응변수(response variable)

단순선형회귀분석(Simple Linear Regression Analysis)

1.단순선형회귀모형(Simple Linear Regression Model)

2변수의 x와 y의 관찰값이 관계가 다음과 같은 경우:

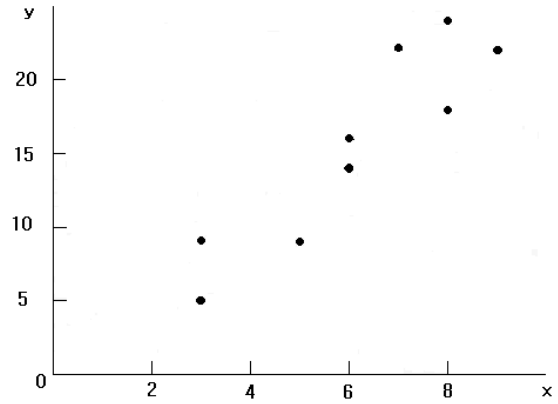
$y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1,2,,\cdots,n \ , \ e_i \sim iid \ N(0, \sigma^2)$

β_0 : y와 x의 직선의 절편(intercept)을 나타내는 모수(parameter)

β_1 : y와 x의 직선의 기울기(slope)를 나타내는 모수(parameter)

y_i : i번째 y의 값 , x_i : i번째 x의 값

eg1) 약의 복용량(x)과 약효지속시간(y)의 관계조사



2. 최소제곱법(Least Square Method)

주어진 데이터 $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$를 이용하여

β_0, β_1 의 추정값인 $\hat{\beta}_0, \hat{\beta}_1$ 를 구하기

$\hat{\beta}_0$,$\hat{\beta}_1$은 $Q = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$ 를 최소화시키는 b_0, b_1의 값

$\rightarrow \quad \hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}}$, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

where $S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2, \quad S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$

적합된 회귀직선(Fitted regression line) : $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

eg) 약효지속시간.sav

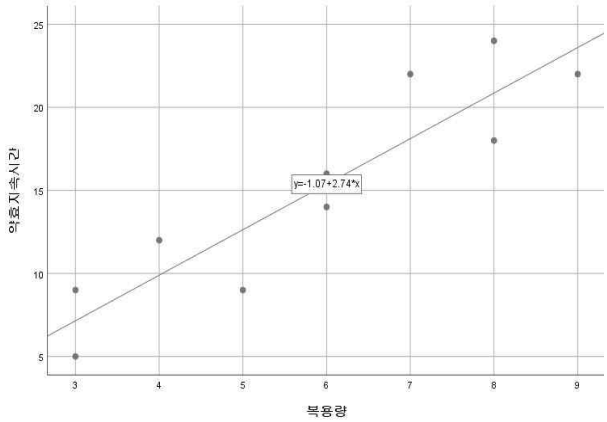
분석-회귀분석-선형-통계량에서 추정값, 신뢰구간 체크

적합된 회귀직선(Fitted regression line) : $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ = -1.07 + 2.74x

(x: 약복용량, y: 약효지속시간)

그래프-레거시 대화상자 -산점도/점도표

(산점도에 double click - 그림위  click - 선형(L))



분석-회귀분석-선형, 종속변수=y, 독립변수=x

모형 요약

모형	R	R 제곱	수정된 R 제곱	추정값의 표준 오차
1	.910[a]	.828	.807	2.821

a. 예측자: (상수), 복용량

R^2=0.828, R_a^2=0.807

결정계수(R^2 coefficient of determination): x변수를 사용함으로써 줄어드는 전체의 변동의 비율 - x변수가 설명할수 있는 y변수의 변동의 비율 - 주어진 모형이 전체변동(y)를 얼마나 잘 설명하는지 나타내는 값 [0,1]

$$R^2 = \frac{SSR}{SSTO}$$

수정된 결정계수(R_a^2, Adjusted R^2) : 독립변수가 많아지면 R^2은 증가만 함.

$$R_a^2 = 1 - \frac{(n-1)}{(n-p)} \frac{SSE}{SSTO}$$

x변수와 y변수의 상관계수(correlation coefficient)= $\pm\sqrt{R^2}$ (부호= $\hat{\beta}$의 부호) [- or +]

계수[a]

모형		비표준화 계수 B	비표준화 계수 표준화 오류	표준화 계수 베타	t	유의확률
1	(상수)	-1.071	2.751		-.389	.707
	복용량	2.741	.441	.910	6.214	.000

a. 종속변수: 약효지속시간

적합된 선형회귀식 : $\hat{y} = -1.071 + 2.741x$

$H_0 : \beta_1 = 0$ ~ x와 y사이에 선형관계가 없다.(x가 y에 영향을 끼친다)

p값(유의확률) < α이면, 유의수준 α100%로 $H_0 : \beta = 0$을 기각.

p값(유의확률)=0.000 < 0.05이므로, 유의수준 5%로 $H_0 : \beta_1 = 0$을 기각.

결론: 약복용량이 약효지속시간에 영향을 끼친다고 할 수 있다.

중선형회귀분석(Multiple Linear Regression Analysis)

eg)

체중(x1), 나이(x2)와 혈압(y)과의 관계조사

공정온도(x1), 공정압력(x2)과 제품의 강도(y)와의 관계조사

x : 독립변수(independent variable) or 설명변수(explanatory variable)가 2개 이상

y : 종속변수(dependent variable) or 반응변수(response variable)

1.중선형회귀모형(Multiple Linear Regression Model):

x1, x2 와 y의 관찰값이 관계가 선형 such that

$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i,\ i = 1,2,,\cdots,n,\quad e_i \sim iid\ N(0, \sigma^2)$

y_i : i번째 y의 값

x_{1i} : i번째 x1의 값 x_{2i}: i번째 x2의 값

2. 최소제곱법(Least Square Method)

주어진 자료를 이용하여 $\beta_0, \beta_1, \beta_2$ 의 추정값인 $\hat{\beta_0}, \hat{\beta_1}, \hat{\beta_2}$ 를 구하기

$\hat{\beta_0}, \hat{\beta_1}, \hat{\beta_2}$은 $Q = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i})^2$ 를 최소화시키는 b_0, b_1, b_2의 값

적합된 선형회귀식(Fitted regression line) : $\hat{y} = \hat{\beta_0} + \hat{\beta_1} x_1 + \hat{\beta_1} x_2$

eg) 혈압_중선형.sav

모형 요약

모형	R	R 제곱	수정된 R 제곱	추정값의 표준 오차
1	.882[a]	.777	.713	4.385

a. 예측자: (상수), 나이, 몸무게

계수[a]

모형		비표준화 계수 B	비표준화 계수 표준화 오류	표준화 계수 베타	t	유의확률
1	(상수)	-15.298	29.132		-.525	.616
	몸무게	1.728	.357	.960	4.841	.002
	나이	.332	.111	.594	2.997	.020

a. 종속변수: 혈압

범주형자료분석(χ^2-검정)

관측도수로 이루어진 범주형 자료를 이용하는 분석

(1) **적합도검정**(Goodness of fit test) : 각 개체의 비 혹은 비율이 $p_1^0 : \cdots : p_k^0$ 인지 검정

H_o : 각 개체의 비가 각각 5:2:3 이다.

$\equiv H_o : p_1 = 0.5, p_2 = 0.2, p_3 = 0.3$

H_o 기각 ~ 각 개체의 비가 각각 5:2:3이 아니다.

(2) **동질성검정**(Test of Homogeneity : 여러 모집단에서 각 범주의 비율이 같은지 검정

eg) H_o : 두 식이요법 받은 환자집단에서 건강상태(양호-보통-불량)의 비율이 같다.

$\equiv H_o$: 두 식이요법 받은 환자집단에서 건강상태는 차이가 없다(Homogeneous).

H_o 기각 ~ 두 식이요법 받은 환자집단에서 건강상태의 비율이 다르다.

(3) **독립성검정**(Test of Independence))

1개의 개체로부터 2가지변수에 대해 조사 할 때 두 변수의 관계가 서로 독립인지 검정

eg) H_o : 커피선호도와 흡연유무는 서로 독립이다.

H_o 기각 ~ 커피선호도와 흡연유무는 연관성이 있다.

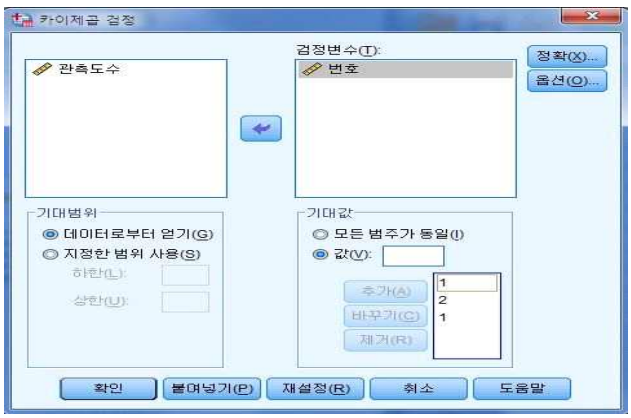
예제1) 적합도검정

자료(잡종.sav): 잡종 530개 관찰하여 잡종1,2,3의 비가 1:2:1 인지 자료를 이용하여 검정

H_0 : 잡종1,2,3의 비가 1:2:1이다.

데이터-가중케이스-가중케이스 지정

분석-레거시대화상자-카이제곱검정,



번호

	관측빈도	기대빈도	잔차
1.00	120	132.5	-12.5
2.00	280	265.0	15.0
3.00	130	132.5	-2.5
전체	530		

검정 통계량

	번호
카이제곱	2.075[a]
자유도	2
근사 유의확률	.354

유의확률=0.354 $\not<$ α=0.05 이므로 유의수준 5%로 "H_0 : 잡종1,2,3의 비가 1:2:1이다."를 기각못함. 결론: 잡종1,2,3의 비율이 1:2:1 이라 할 수 있다.

eg2) 음주여부와 커피선호도 조사(음주_커피.sav)

H_0: 음주여부와 커피선호도는 서로 독립이다.

데이터-가중케이스-가중케이스 지정

분석-기술통계량 -교차분석- 통계량(S)- 카이제곱 선택

음주여부 * 커피선호 교차표

빈도

		커피선호		
		안좋아함	좋아함	전체
음주여부	안좋아함	82	98	180
	좋아함	45	65	110
전체		127	163	290

카이제곱 검정

	값	자유도	근사 유의확률 (양측검정)	정확 유의확률 (양측검정)	정확 유의확률 (단측검정)
Pearson 카이제곱	.599[a]	1	.439		
연속성 수정[b]	.425	1	.514		
우도비	.600	1	.438		
Fisher의 정확검정				.466	.258
유효 케이스 수	290				

유의확률=0.439 $\not<$ α=0.05 이므로 유의수준 5%로

"H_0: 음주여부와 흡연유무는 서로 독립이다."를 기각못함.

~ 음주여부와 커피선호도는 관련없다고 할 수 있다.

** 셀(E)에서 기대빈도 √ → 기대빈도: 귀무가설이 맞다는 가정하에서 각셀의 기대도수

= 음주여부와 흡연유무는 서로 독립이다라는 가정하에서 각셀의 기대도수