6장. 회귀분석(Regression Analysis)

- 하나 이상의 변량으로부터 어떤 변량의 값을 예측한다는 의미를 내포.

- 회귀관계란 두 변량간의 함수적인 변동 관계를 나타내는 것.

- 종속변수(dependent variable) 또는 반응변수(response variable)
  : 두 변수 중에서 다른 변수의 변화에 따라 그 값이 변화하는 변수(y)

- 독립변수(independent variable) 또는 설명변수(explanatory variable)
  : 이에 대응되는 변수(x)

- 회귀식(regression equation)
  : 독립변수 x와 종속변수 y간의 관계를 수식으로 표현한 것.
  : 회귀모형(regression model).

1. 단순 회귀 모형(Simple Regression Model)

1) 모형

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

y_i : i-번째 관측의 종속변수 값

x_i : i-번째 관측의 독립변수 값

β_0 : 회귀 직선의 y절편

β_1 : 회귀 직선의 기울기

ϵ_i : 오차항 $\sim iid \ \ N(0, \sigma^2)$

$\Rightarrow$ 모형으로부터 주어진 x값에 대해 y의 기대값과 분산

(1) y의 기대값 : $E(y|x) = \beta_0 + \beta_1 x = \mu_{y|x}$ $\Leftrightarrow$ 모 회귀식

(2) y의 분산 : $Var(y) = \sigma^2$

⇒ 표본자료로부터 모회귀계수 β_0, β_1를 추정해야 한다.

⇒ 추정된 회귀식

$$\hat{y} = b_0 + b_1 x$$

여기서 b_0, b_1 은 각각 β_0, β_1의 추정치이며, $\hat{y}$ 는 $\mu_{y|x}$ 의 추정치.

※ 추정 방법 - 최소제곱법(Method of Least Square estimation)

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{s_{xy}}{s_{xx}} = r\frac{s_y}{s_x}$$

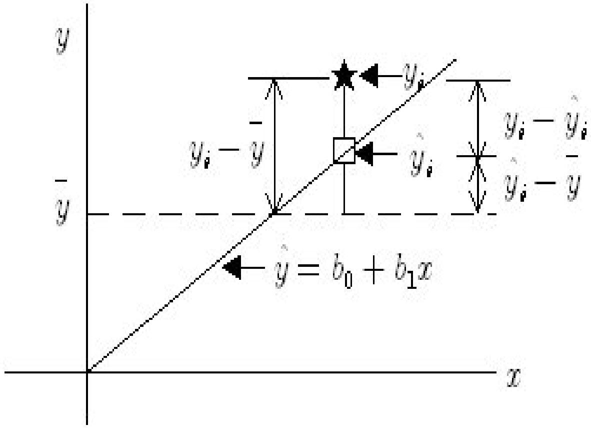
$$b_0 = \bar{y} - b_1 \bar{x}$$

2) 분산분석(Analysis of Variance)

- 관찰치들의 전체변동을 회귀선에 의한 변동과 회귀선으로 설명되지 않는 변동으로 분할.

- 관찰치 y_i의 변동을 나타내는 편차 $(y_i - \bar{y})$

$$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$



- 의미

(관찰치의 편차) = (회귀선에 의해 설명되는 편차) + (회귀선에 의해 설명되지 않는 편차)

$$\Rightarrow \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$\sum_{i=1}^{n} (y_i - \bar{y})^2$: 총변동 또는 총제곱합(total sum of square)이라고 하며 SST 로 표기.

$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$: 회귀선에 의해 설명되지 않는 변동으로 잔차제곱합(residual sum of square) 이라 하며 SSE 로 표기.

$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$: 회귀선에 의해 설명되는 변동으로 회귀제곱합(regression sum of square)이라 하며 SSR 로 표기.

○ 분산분석

귀무가설(H0) : $\beta_1 = 0$, 회귀선이 유의하지 않다.

대립가설(H1) : $\beta_1 \neq 0$, 회귀선이 유의하다.

요인	자유도	제곱합	평균제곱	F-통계량
회귀	1	SSR	MSR=SSR/1	F0=MSR/MSE
잔차	n-2	SSE	MSE=SSE/(n-2)	
총합	n-1	SST		

⇒ F-통계량 값은 MSR과 MSE간의 비율로서, 이 비율이 크다는 의미는 잔차들에 의해 설명되는 변동 보다 회귀선에 의해 설명되는 변동이 크다는 의미로서 회귀선이 의미가 있다는 것.

⇒ 따라서 회귀선의 가정이 옳다는 의미가 되며, x와 y간의 관계를 회귀선이 잘 설명한다는 의미가 된다. 결과적으로 귀무가설이 기각된다.

3) 결정계수(Coefficient of determination)

- 표본으로 추정된 회귀선이 관찰치를 얼마나 적절히 설명하는가를 나타내는 척도.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- 결정계수 R^2의 성질

(1) 추정된 회귀선에 모든 관찰치가 놓이면 SSE=0 이므로 $R^2 = 1$이 된다.

(2) 추정된 회귀선의 기울기가 $b_1 = 0$ 이면 결정계수 $R^2 = 0$ 이다.

(3) 결정계수 R^2와 상관계수 r 간에는 $R^2 = r^2$ 인 관계가 성립한다.

(4) 결정계수 R^2 값이 1에 가까운 값일수록 추정된 회귀선은 유의하다.
반대로 R^2 값이 0에 가까울수록 추정된 회귀선은 의미가 없게 된다.

(5) 결정계수 R^2 값은 자유도에 영향을 받기 때문에 절대적인 값은 아니다.

4) 개별 회귀 계수에 대한 검정

- 회귀 계수 β_0과 β_1의 유의성을 검정.
- 직선의 기울기와 절편에 대한 검정.

<절편 β_0에 대한 검정>

○ 가설

H0 : $\beta_0 = 0$, 절편은 유의하지 않다.

H1 : $\beta_0 \neq 0$, 절편은 유의하다.

○ 검정 통계량

$$T = \frac{\hat{\beta}_0 - \beta_0}{s_{\hat{\beta}_0}} \sim t(n-2)$$

○ 판정

$|T_0| > t(n-2, \alpha/2)$ 또는 유의확률 p-값 < 0.05 이면 귀무가설 기각.

⇒ 절편은 유의하다 즉, 모형에서 절편항은 유지되어야 한다.

<기울기 β_1에 대한 검정>

○ 가설

H0 : $\beta_1 = 0$, 회귀모형은 유의하지 않다.

H1 : $\beta_1 \neq 0$, 회귀모형은 유의하다.

○ 검정 통계량

$$T = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} \sim t(n-2)$$

○ 판정

$|T_0| > t(n-2, \alpha/2)$ 또는 유의확률 p-값 < 0.05 이면 귀무가설 기각.

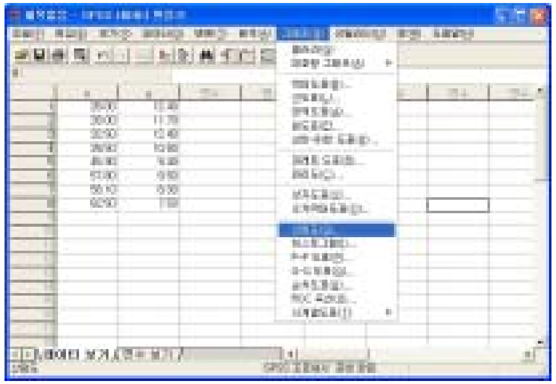
$\Rightarrow$ 회귀모형은 유의하다.

[예1] 지난 8일간의 데이터로서 y는 매일의 연료소비량이고 x는 평균온도(화씨)를 나타낸다.

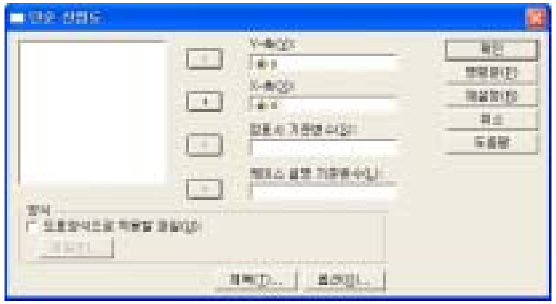
일	1	2	3	4	5	6	7	8
평균온도	28.0	28.0	32.5	39.5	45.9	57.8	58.1	62.5
연료소비량	12.4	11.7	12.4	10.8	9.4	9.5	8.0	7.5

① x와 y의 산점도를 작성하라.
② x와 y의 상관계수를 구하라.
③ x와 y의 모상관계수가 0인지를 유의수준 0.05하에서 검정하라.
④ x와 y의 회귀 모형을 추정하라.
⑤ 분산분석표를 작성하라.

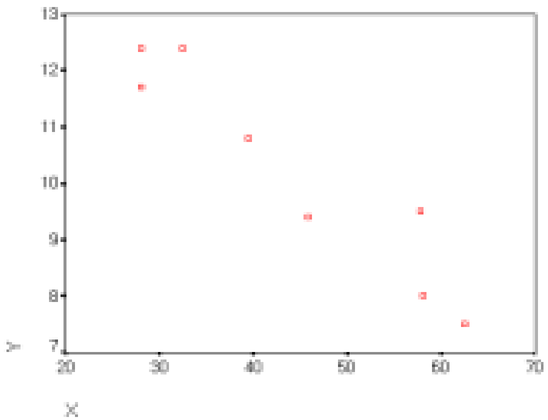
풀이]
① 산점도 작성 : 데이터입력 ->그래프 -> 산점도



: 단순 -> 정의 : x축, y축 값 설정



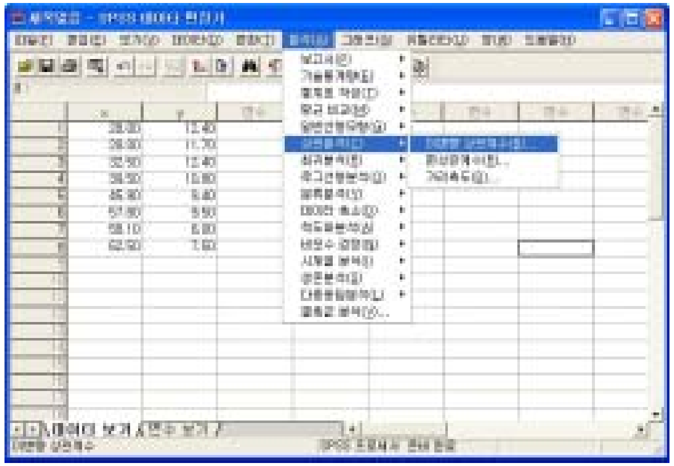
[산점도 출력 결과]



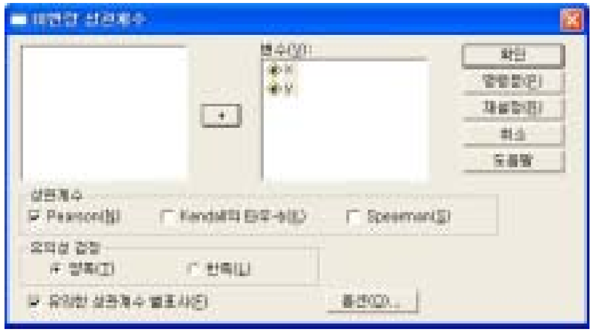
⇒ X와 Y의 산점도 결과로부터 두 변수 간에는 (-)음의 관계가 있음을 파악.

⇒ 회귀 모형 적용의 경우 기울기가 (-) 가 될 것으로 예측.

② 상관계수 : 분석 -> 상관분석 -> 이변량 상관계수



: 변수 -> X와 Y 선택



: 옵션 -> 평균과 표준편차

[상관분석 출력결과]

기술통계량

	평균	표준편차	N
X	44.0375	14.1401	8
Y	10.2125	1.9105	8

상관계수

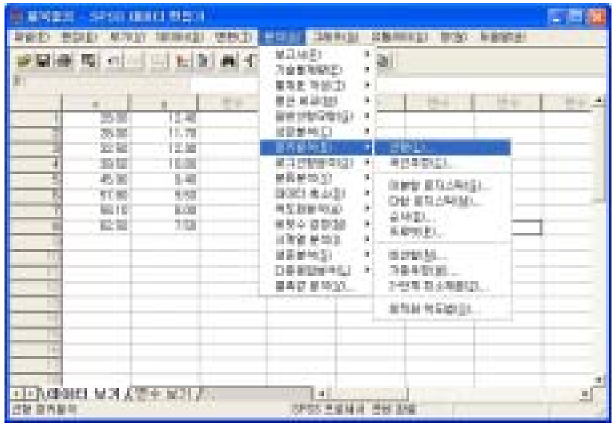
		X	Y
X	Pearson 상관계수	1.000	-.948
	유의확률 (양쪽)	.	.000
	N	8	8
Y	Pearson 상관계수	-.948	1.000
	유의확률 (양쪽)	.000	.
	N	8	8

⇒ X와 Y의 상관계수 : $r(x, y) = -0.948$

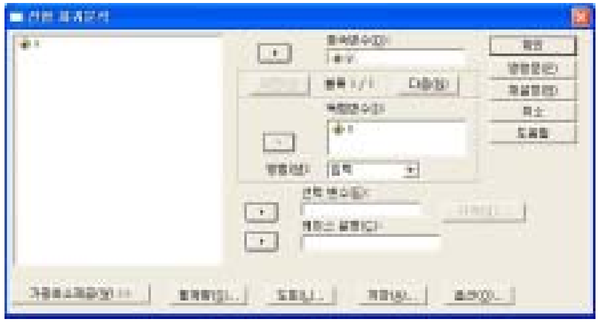
⇒ 유의확률 p-값 : $0.000 < 0.05$ 이므로 귀무가설 기각(모상관계수가 0 이라는 가설을 기각).
즉, X와 Y는 상관관계가 존재한다.

⇒ 상관계수의 출력결과 “산점도 출력” 결과에서 예측한 (-)음의 상관이 존재할 것이라는 사실이 옳음.

④ 회귀 모형의 추정 : 분석-> 회귀분석 -> 선형



: 종속변수와 독립변수의 선택



[모형 추정]

계수

		비표준화 계수		표준화 계수	t	유의확률
모형		B	표준오차	베타		
1	(상수)	15.856	.804		19.730	.000
	X	-.128	.017	-.948	-7.332	.000

⇒ 추정된 회귀식

$\hat{y} = 15.865 - 0.128x$ ⇔ 연료소비량의 추정치 = 15.895 - 0.128×온도

⇒ 개별 회귀계수의 검정

<β_0에 대한 가설 검정>

H0: $\beta_0 = 0$

H1: $\beta_0 \neq 0$

에 대해 유의확률 p-값 = 0.000 <0.05 이므로 귀무가설 기각, 즉 절편은 유의하다.

<β_1에 대한 가설 검정>

H0: $\beta_1 = 0$

H1: $\beta_1 \neq 0$

에 대해 유의확률 p-값 = 0.000 <0.05 이므로 귀무가설 기각, 즉 기울기는 유의하다.

⑤ 분산분석표

분산분석

모형		제곱합	자유도	평균제곱	F	유의확률
1	선형회귀분석	22.984	1	22.984	53.759	.000
	잔차	2.565	6	.428		
	합계	25.549	7			

<가설>

H0 : 회귀모형은 유의하지 않다. ($\beta_1 = 0$)

H1 : 회귀모형은 유의하다.($\beta_1 \neq 0$)

⇒ 유의확률 p-값=0.000<0.05 이므로 귀무가설 기각.

⇒ 결과적으로 적합된 회귀모형은 관찰치들을 잘 설명한다.

⇒ 분산분석 결과는 개별회귀 계수에 대한 검정 중 기울기에 대한 검정과 같은 결과를 얻게 된다.

※참고 - 결정계수

모형 요약

모형	R	R 제곱	수정된 R 제곱	추정값의 표준오차
1	.948	.900	.883	.6539

5) 회귀 진단 - 잔차분석

- 잔차에 대한 제 가정을 만족하는지 여부를 검토.
- 등분산성, 독립성, 정규성.
- 잔차도표를 이용하여 확인.

○ 정규성에 위배되는 경우 : 잔차를 이용한 콜모고로프-스미로노프 검정.
⇒ 변수변환을 수행한다.

○ 등분산성에 위배되는 경우 : 잔차의 산점도를 작성하여 확인.
⇒ 변수변환 또는 가중회귀 방법을 수행한다.

○ 독립성에 위배되는 경우 : Durbin-Waston 검정을 수행.
⇒ 시계열 분석을 고려한다.

[예2] 예제1]의 데이터를 이용하여 잔차에 대한 검토를 수행하라.

1) 독립성
2) 정규성
3) 등분산성

○ 독립성 : D-W 검정

분석 -> 회귀분석 -> 선형 -> 통계량 : Durbin-Waston 선택.

⇒ 가설

H0 : 독립이다(오차항의 자기상관이 없다).

H1 : 독립이 아니다(오차항은 양의 자기상관이 존재한다).

※ 더빈왓슨 d- 통계량

$$d = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}, \quad 0 \le d \le 4$$

⇒ 판정 기준

$d < d_L$: 귀무가설 기각.

$d > d_U$: 귀무가설 채택.

$d_L < d < d_U$: 판정 보류.

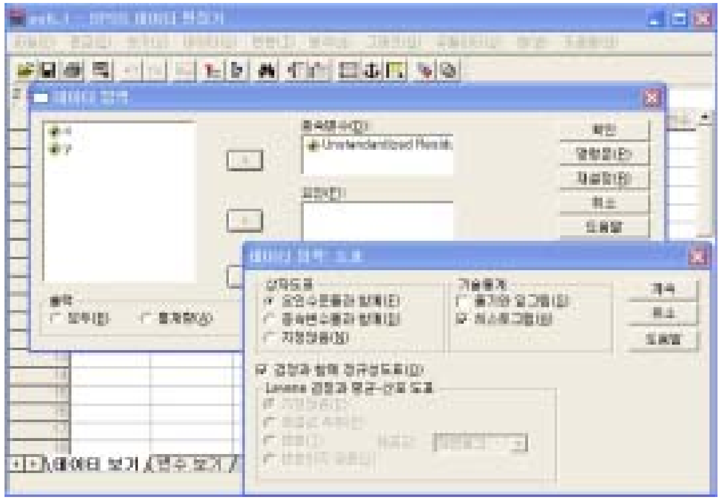
양의 상관	판정보류	독립	독립	판정보류	음의상관

0 d_L d_U 2 $4-d_U$ $4-d_L$ 4

모형	R	R 제곱	수정된 R 제곱	추정값의 표준오차	Durbin-Watson
1	.948	.900	.883	.6539	3.014

⇒ d=3.014 > 상한값 1.36 이므로 귀무가설 채택, 즉 오차의 자기상관은 존재하지 않는다.

○ 정규성 : 저장 -> 잔차 표준화 하지 않음.
분석-> 기술통계량-> 데이터 탐색 ->정규성 검정



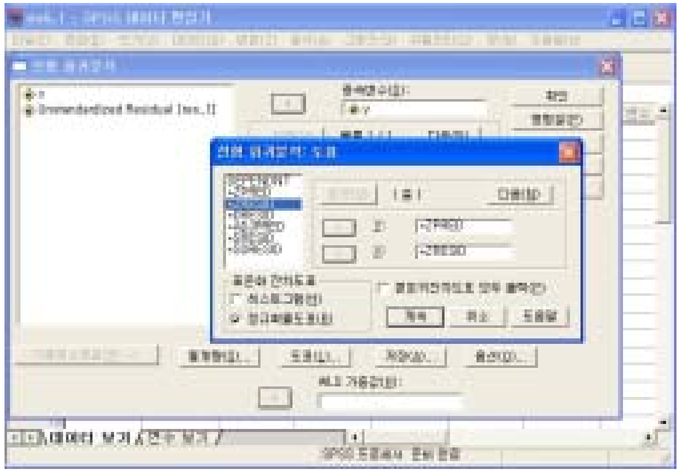
<정규성 검정>

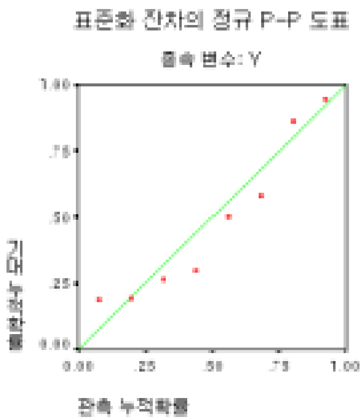
	Kolmogorov-Smirnov			Shapiro-Wilk		
	통계량	자유도	유의확률	통계량	자유도	유의확률
Unstandardized Residual	.217	8	.200	.880	8	.245

⇒ 정규성 검정 결과 : 유의확률 p-값=0.200 >0.05 이므로 귀무가설 채택, 즉 정규분포를 따른다.

<표준화 잔차를 이용한 정규성 검토>

○ 분석->회귀분석-> 선형 : 도표 단추 -> 정규 확률도표

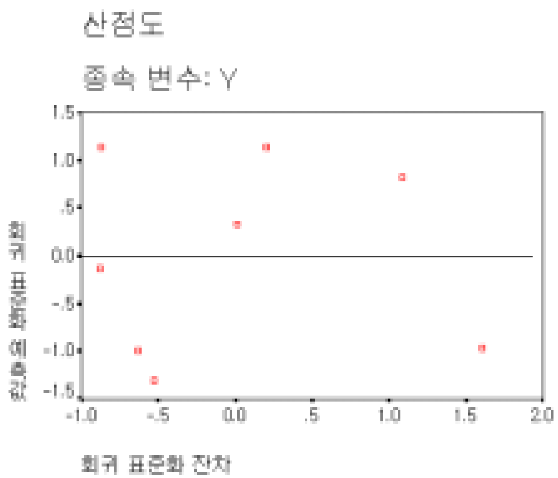




⇒ 직선 주위에 관찰치들이 나타나있는 것으로 보아 정규분포를 따른다고 볼 수 있다.

<등분산성>

○ 선형 : 도표 단추 ->Z_PRED, Z_RESID 선택



⇒ 표준화 잔차도표로부터 잔차값 0을 중심으로 관찰치들의 패턴이 없는 것으로 보임. 결과적으로 등분산성을 만족함.