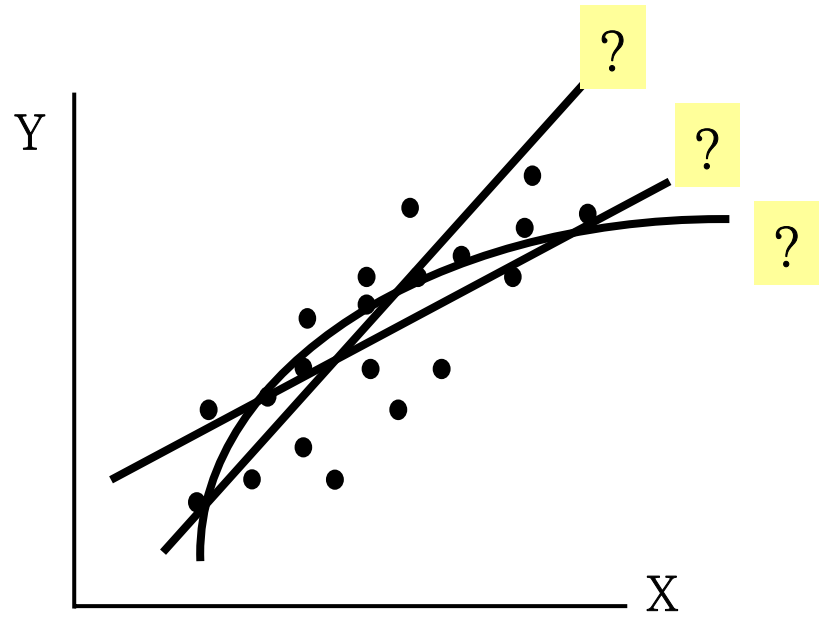
회귀 분석
(Regression)

회귀(Regression) 분석

- 변수들간의 관련성을 규명하기 위하여 어떤 수학적 모델을 가정하고,
- 이 모델을 측정된 추정된 모형을 사용하여 필요한 예측을 하거나 관심 있는 통계적 추론을 함
- 영국의 우생학자 Francis Galton(1822~1911)과 통계학자 Pearson 이 처음 사용

$$Y = f\,[\,X_1, X_2, X_3, X_4, X_5, \ldots\ldots]$$

프로세스의 특징, 현상 → Modeling → 수리적 관계식



회귀(Regression) 분석의 유형

$$Y = f(X)$$

종속변수 반응변수 — 독립변수 설명변수

❑ **단순 선형 회귀(Simple linear regression)**

독립변수가 한 개이고 종속변수가 한 개인 경우 함수관계가 직선방정식의 관계인 것을 가정

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

❑ **중 선형 회귀(Multiple linear regression)**

독립변수가 두 개 이상 이고 종속변수가 한 개인 경우 변수관계가 선형방정식의 관계인 것을 가정

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_2 + \beta_3 x_3 + \cdots + \varepsilon_i$$

❑ **비 선형 회귀(Non-linear regression)**

독립변수와 종속변수의 관계를 곡선(비선형)으로 가정하여 분석

❑ X와 y의 관련성 및 독립변수 x가 정해질 때 종속변수 y값을 추정할 수 있음.

$$y = f(x_1, x_2, \ldots, x_n)$$

단순 선형 회귀 (Simple Linear Regression)

- 단순 선형 회귀에서는 계수는 1차이며 회귀식은 다음과 같이 나타난다.

$$y = b_0 + b_1 x + e$$

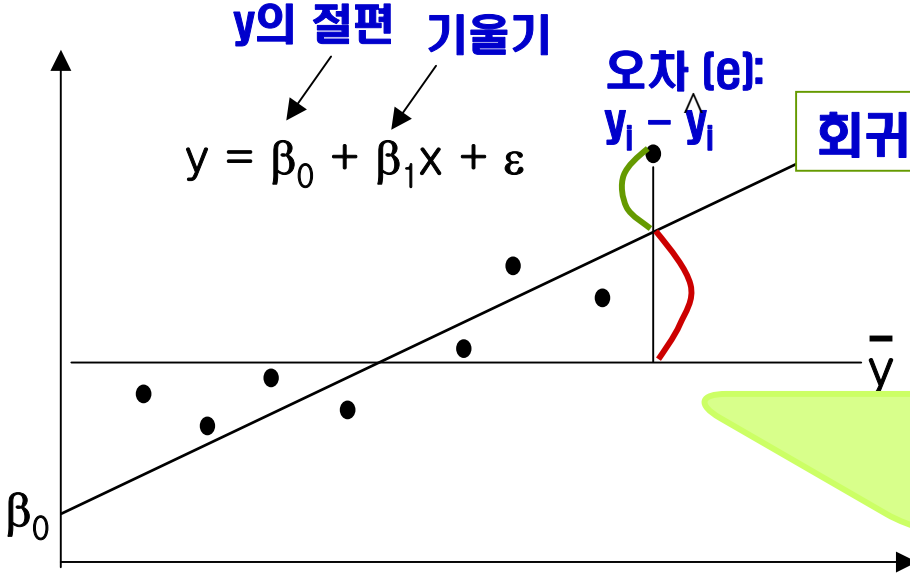
y: Response, 종속 변수

x: 독립 변수, 회귀자 (regressor)

e: 오차, 잡음

1. 가장 좋은 b_0 와 b_1를 어떻게 얻는가?
2. 회귀식이 예측하는 데 사용되기가 적합한 것인가?

회귀식의 구조



$SST = \Sigma (y_i - \bar{y})^2$ $\bar{y}$: 전체 데이터의 평균

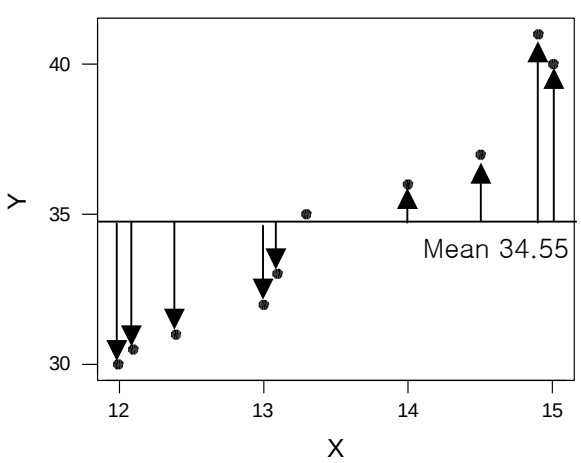
$SSR = \Sigma (\hat{y}_i - \bar{y})^2$ y_i: 각각의 데이터

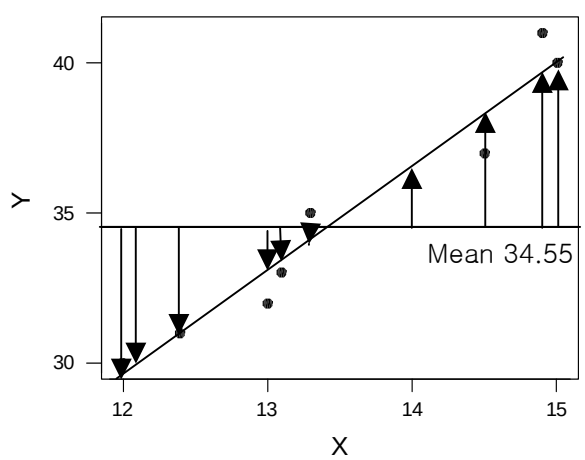
$SSE = \Sigma (y_i - \hat{y}_i)^2$ $\hat{y}_i$: 회귀식을 통한 추정 값

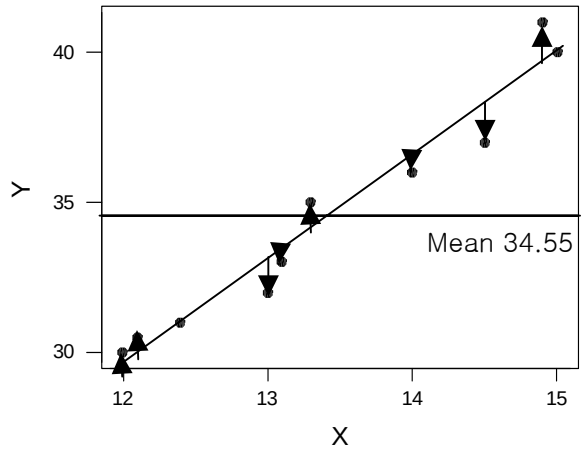
- 회귀선은 각 점에서 직선거리의 수직거리(오차)의 자승 값을 최소화 시키는 선으로 error의 변동 합을 최소화 시킨다. – 최소자승법(least square regression)
- 오차(e)는 회귀선과 관찰된 점의 수직 거리 임.
- 회귀선의 예측력은 오차(error)에 의존되며 , R^2 값으로 나타낸다.

R^2 의 의미-결정계수

전체 변동 SST (Total Sum of Square) = **회귀식으로 설명 되어지는 변동 SSR** + **오차에 의한 변동 SSE**







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

R^2 은 총 변동 중 ***회귀 모형에 의해 설명되는 부분의 비로***

$0 \le R^2 \le 1$ (항상 양의 값)

이 값이 클수록 x는 중요한 인자임.

회귀계수의 추정

모델의 가정

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

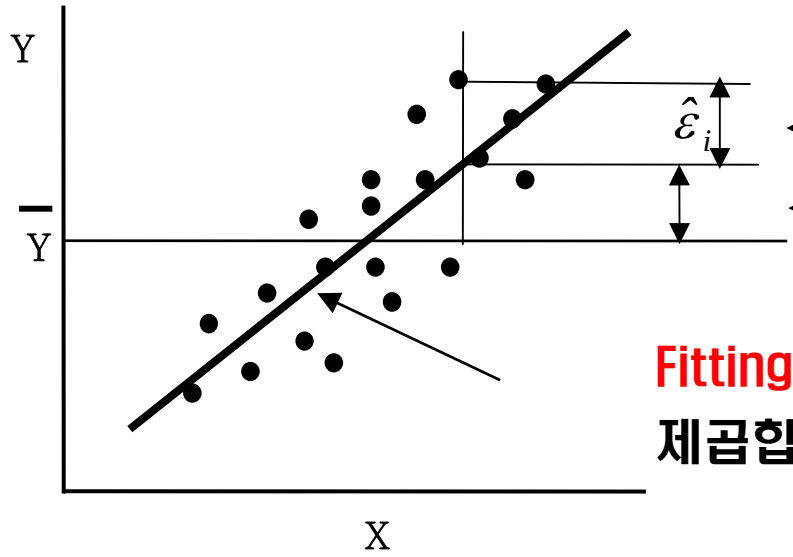
β_0 : 절편(미지의 모수)

β_1 : 기울기(미지의 모수)

ε_i : 오차(잔차)항 $\sim N(0, \sigma^2)$

회귀 계수의 추정방법

➢ **오차제곱의 합을 최소화하는 최소제곱법으로 추정**



$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

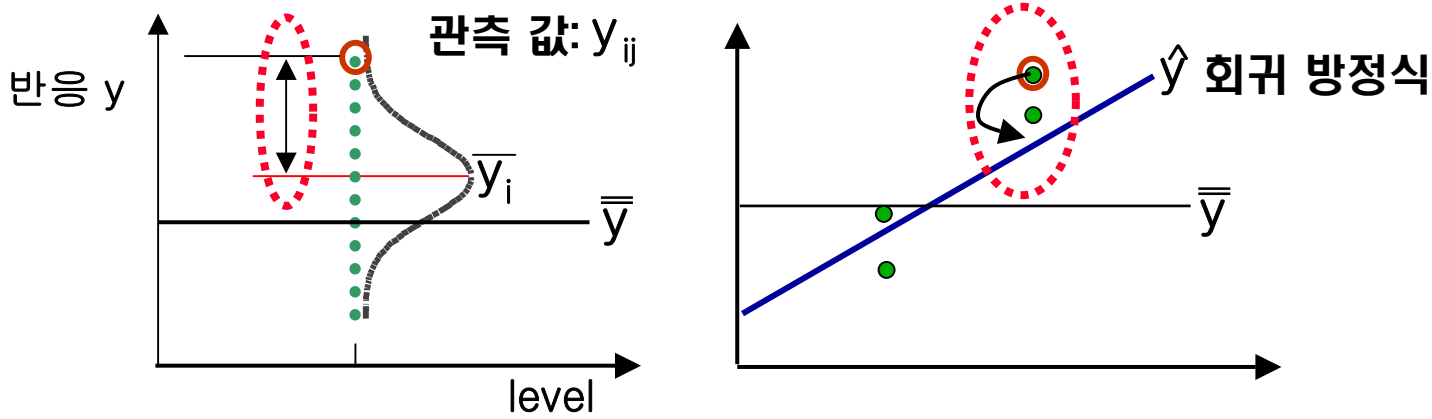
오차편차(설명될 수 없는 편차)

회귀편차(모델로 설명될 수 있는 편차)

Fitting한다는 것은 최소 제곱법에 의해 오차의 제곱합을 최소화하는 직선을 선택하는 것이다.

잔차 분석 (Residual Analysis)

▪ 잔차란?



- 관측 값과 인자 수준의 평균과의 차이
- ***관측 값과 적합 값 (fitted value)과의 차이로***

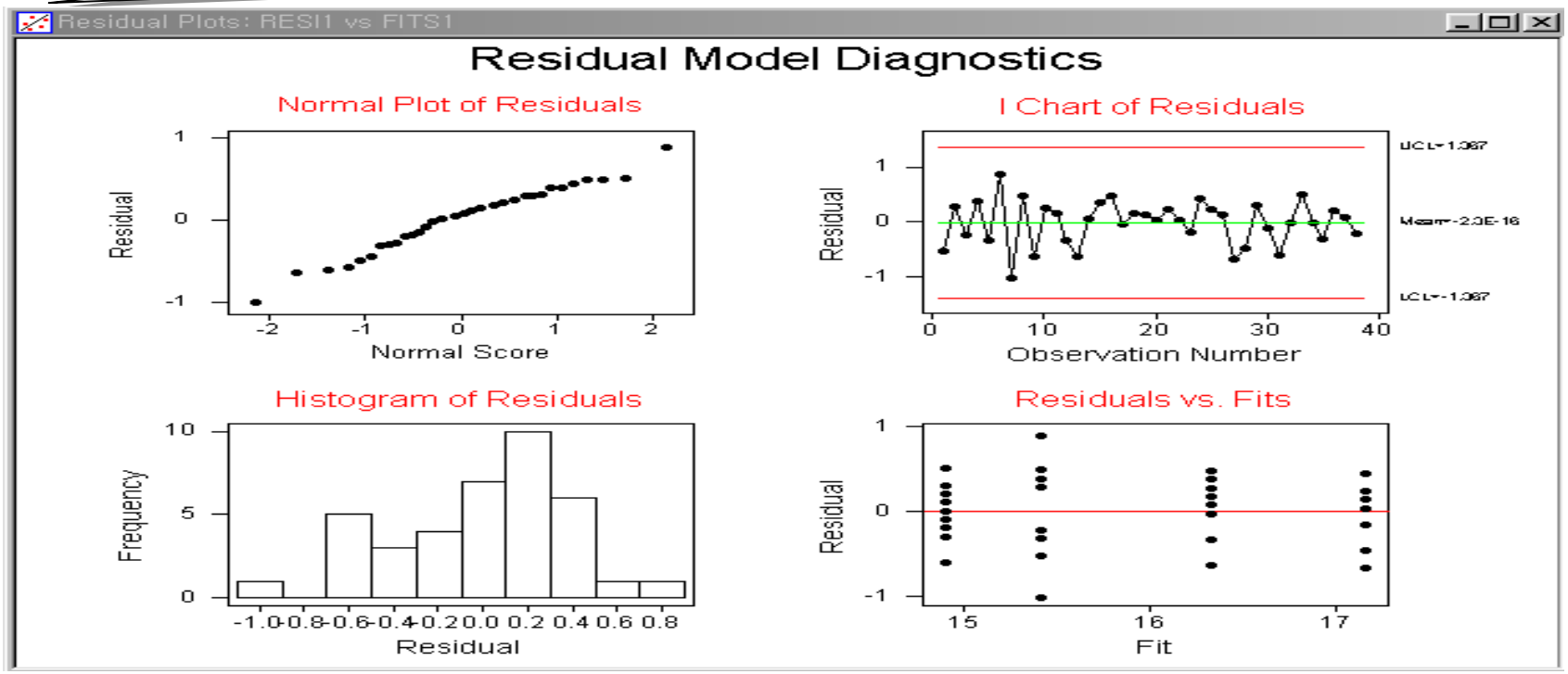
등급으로는 설명되지 않는 변동이며, 평균이 "0"인 정규분포를 이룬다.

▪ **반응치 y에 영향을 미치는 주요한 요인인지, 분석한 결과에 대하여 반드시 잔차 분석을 행하시오.**

잔차 분석 (Residual Analysis)

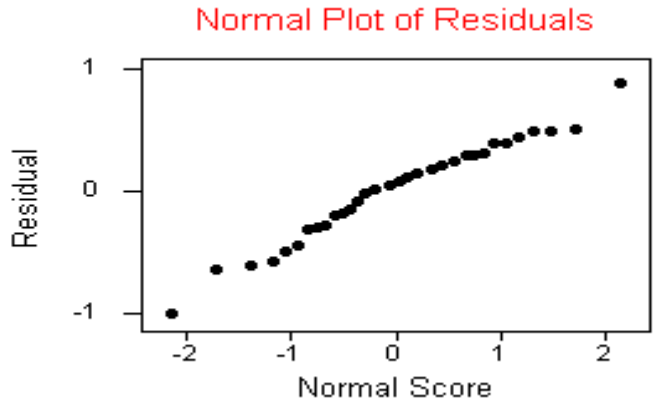
❑ 분석 결과의 타당성 검토 :

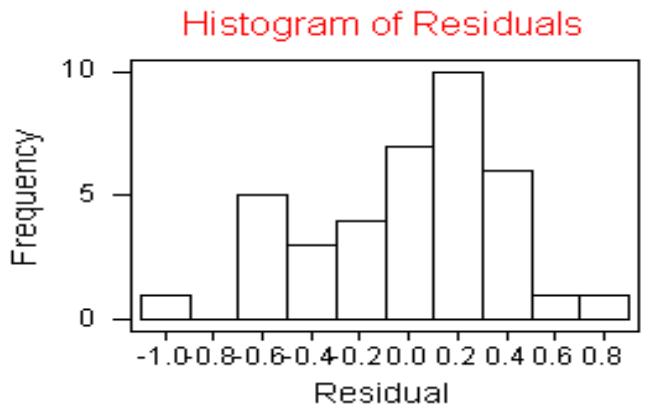
- 잔차는 정규성을 띄고 있는가?
- 잔차의 개별 데이터들은 서로 독립적인가? 어떤 경향을 갖고 있는지?
- 등 분산의 검토 – "0"를 중심으로 경향 없이 무작위로 분포되어 있는가?



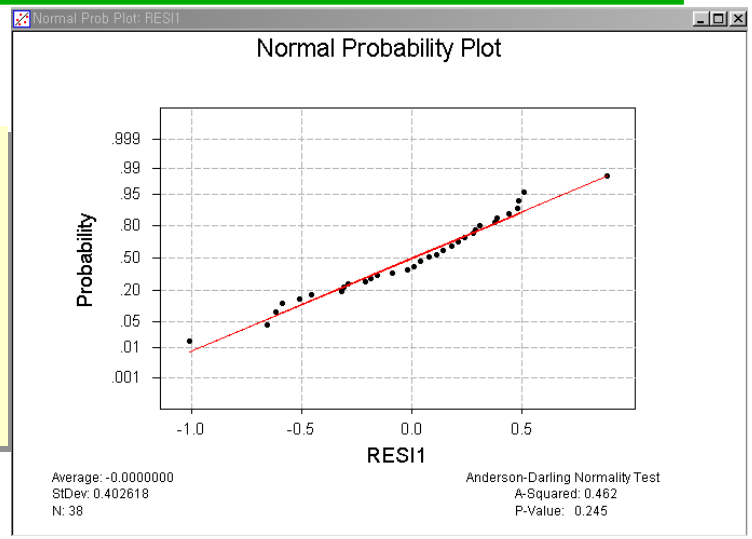
잔차의 정밀 분석

- ***정규성 (Normal Plot & Histogram)***



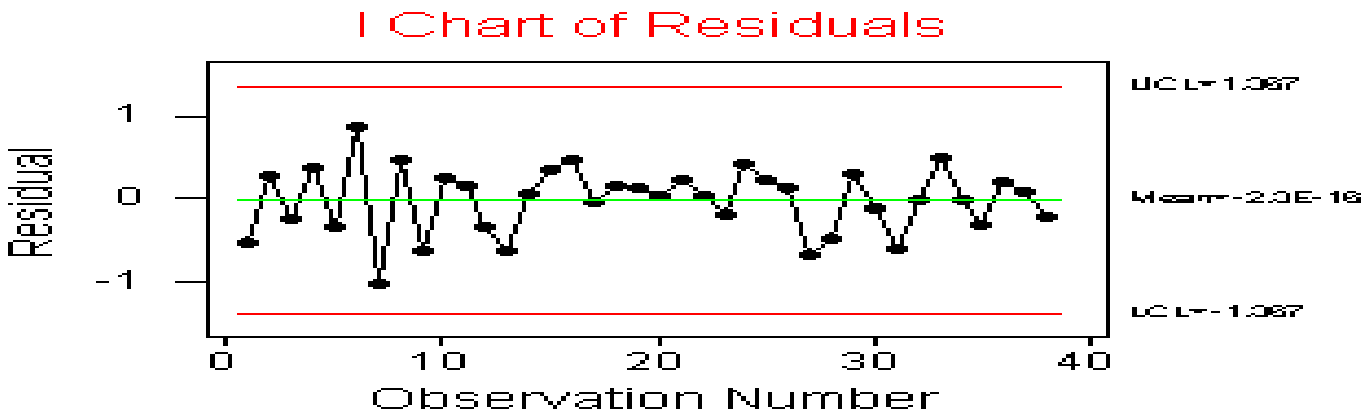


- 잔차가 정규분포에 근사하고 있는지에 대한 검토
- 히스토그램은 종 모양의 곡선인가?
  (데이터가 30개 이하이면 무시해도 좋음)
- 필요 시 잔차에 대한 정규성 검토



잔차의 정밀 분석

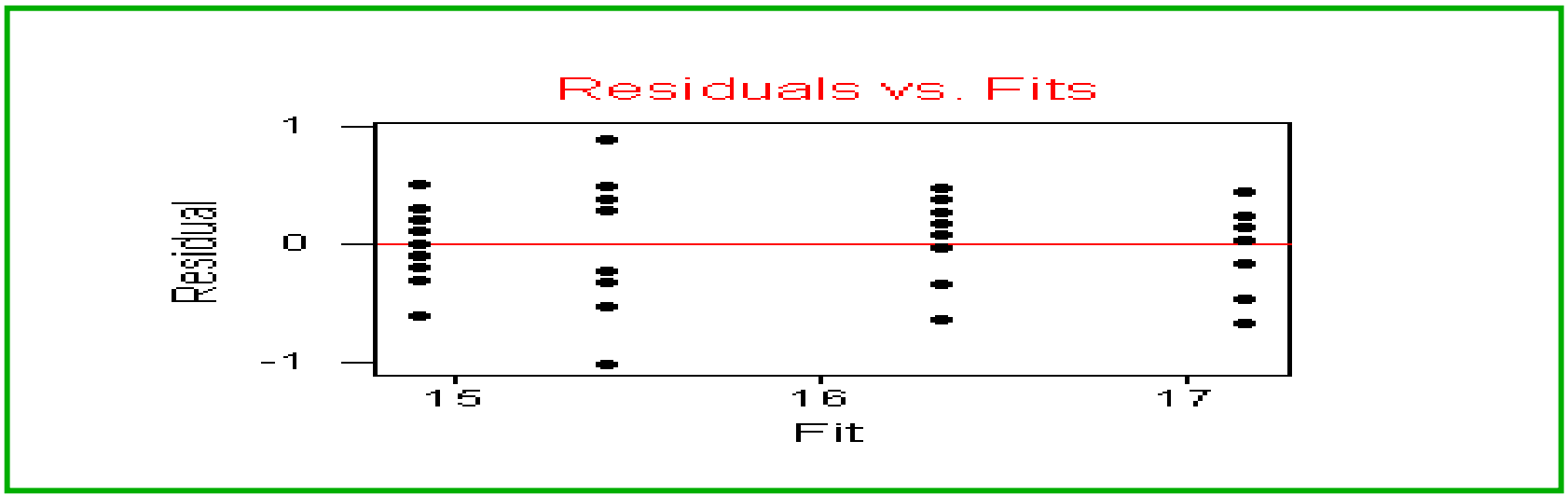
▪ *시계열 잔차 분석*



- 수평축은 관측된 값들의 잔차의 시간적 흐름을 나타낸다.
  - 첫번째 결과 값 1, 두 번째 결과 값 2, 처리 수준에 관계없이 시간 흐름 순으로 측정
  - 수직 축은 측정된 잔차 값을 표시
- 잔차들은 시간의 흐름에 따라 독립성을 갖고 있는지? 변화의 경향을 갖고 있는지?
- 잔차는 관리 한계선 내에 존재하여야 한다.

잔차의 정밀 분석

- ***잔차 대 적합 값 평균***



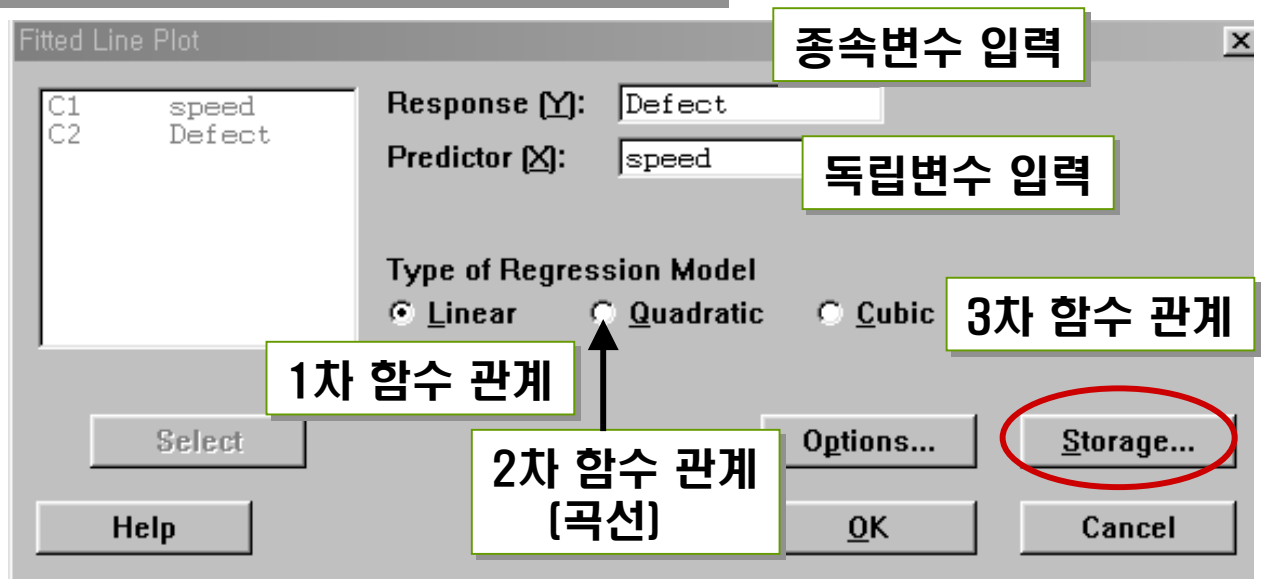
- 수평 축 적합 값(fit)은 각 수준에서의 평균값을 나타냄.
- 수직 축은 관측 값과 각 수준에서의 평균과의 차이를 나타냄.
- 분산이 같다는 가정 하에서 편차가 심하면 표본이 다른 변동을 갖고 있는지 또는 데이터 변환이나 다른 해결책이 필요할 수 있다.
- "0" 근처에서 랜덤하게 타점 될 수록 적절한 모형으로 판단한다.

단순회귀분석 – Minitab

❑ 앞의 상관분석의 생산성 향상 예제의 데이터를 활용하여 회귀분석을 실시 하시오.
– 컨베어 속도와 결함율 간의 함수 관계는? (종속변수와 독립변수)

C1	C2
speed	Defect
11	14
8	10
13	14
7	10
10	11
6	7
9	10
11	13
12	15

STAT > Regression > Fitted line plots



단순회귀분석 – Minitab

잔차 분석을 위하여

Graph

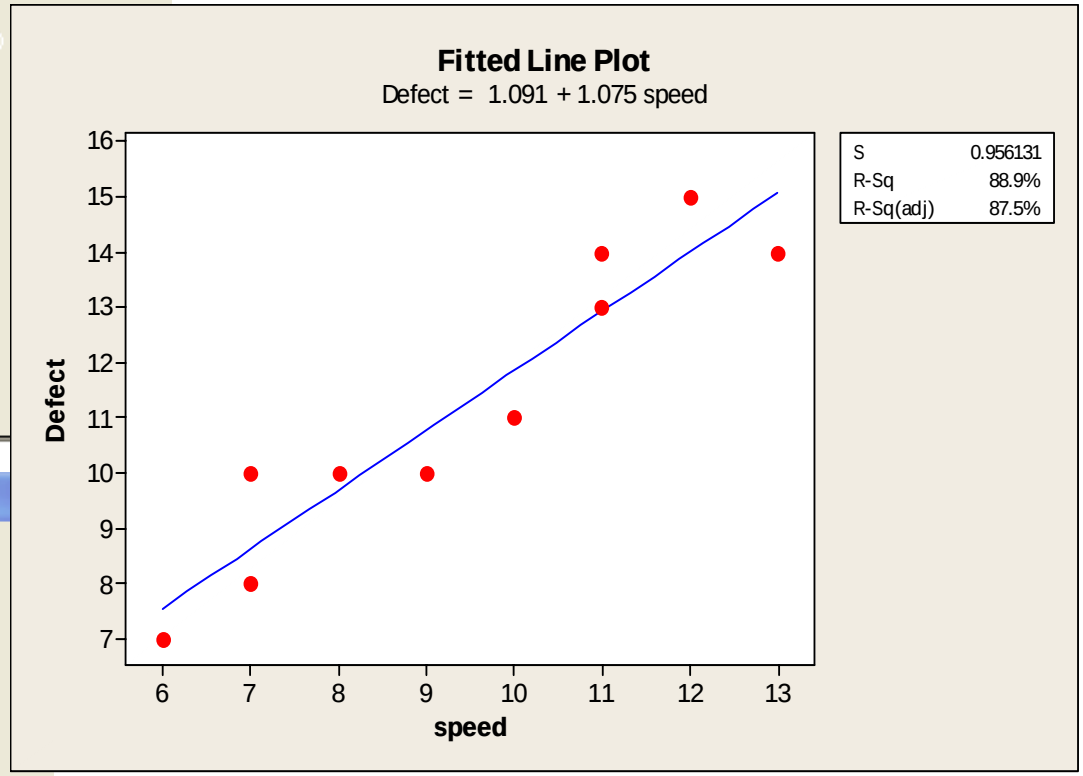
Residuals for Plots:
Regular Standardized

Residual Plots
Individual plots
Histogram of residuals
Normal plot of residuals
Residuals versus fits
Residuals versus order
Four in one

Residuals versus the variables:

Fitted Line Plot – Storage
Residuals
Standardized residuals
Deleted t residuals
Fits
Coefficients
Residuals in original units
Fits in original units

Help OK Cancel



세션 창 정보

Regression Analysis: Defect versus speed

The regression equation is
Defect = 1.09127 + 1.07540 speed → **회귀식**

S = 0.956131 R-Sq = 88.9 % R-Sq(adj) = 87.5 %

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	58.2865	58.2865	63.7578	0.000
Error	8	7.3135	0.9142		
Total	9	65.6000			

- *P-value가 0.05보다 작으므로 컨베어 속도를 높이면 불량율이 증가함.*

단순회귀분석 – Minitab

회귀식
오차제곱 (ε_i^2)의 합을 최소화하는 방법으로 추정 (Least Square Method)

Regression Analysis: Defect versus speed

The regression equation is

Defect = 1.09 + 1.08 speed

Predictor	Coef	SE Coef	T	P
Constant	1.091	1.302	0.84	0.426
speed	1.0754	0.1347	7.98	0.000

S = 0.956131 R-Sq = 88.9% R-Sq(adj) = 87.5%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	58.287	58.287	63.76	0.000
Residual Error	8	7.313	**0.914**		
Total	9	65.600	ε_i 의 산포	$\hat{\sigma}^2$	

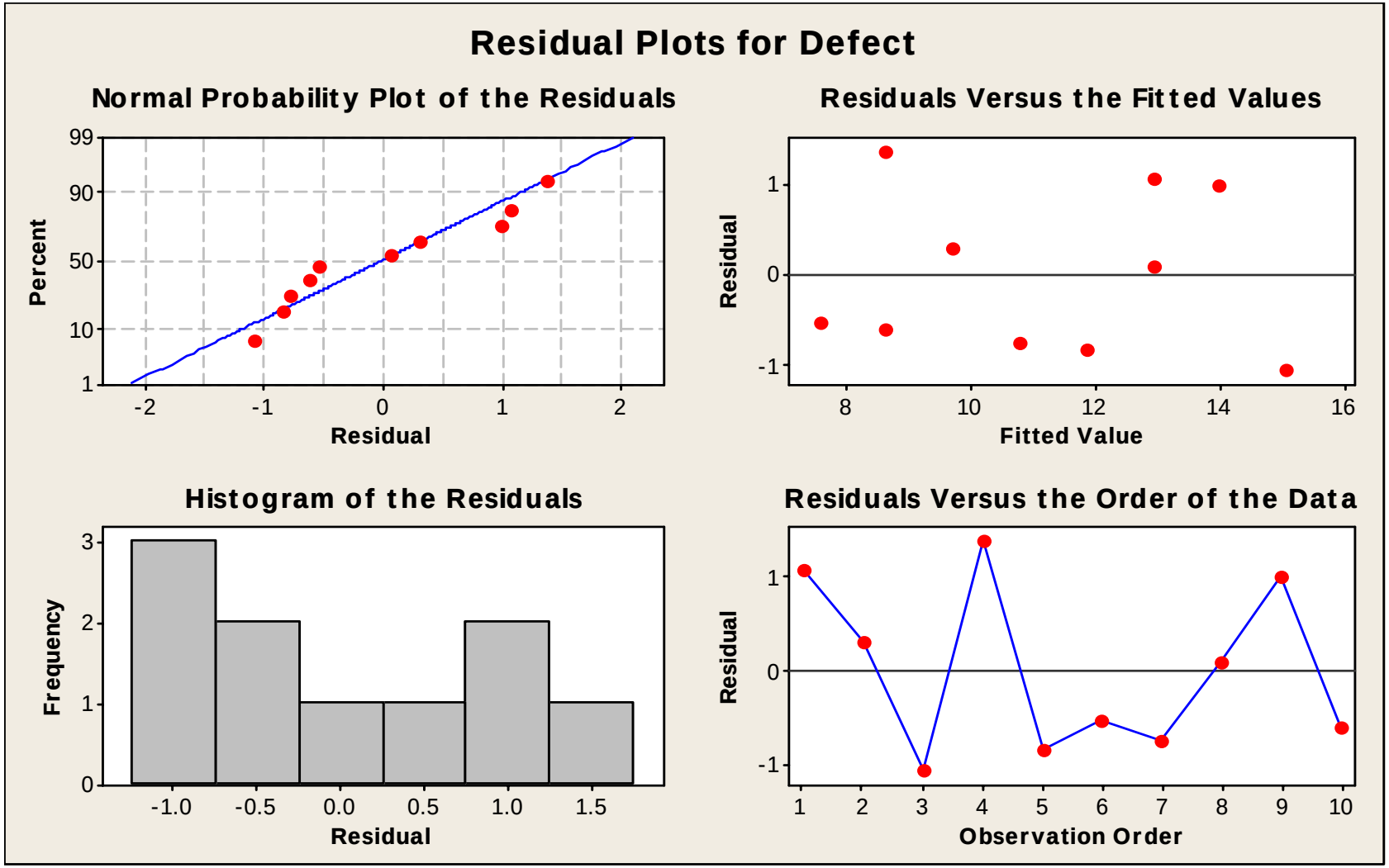
회귀계수가 0이 아닐 확률
회귀계수의 추정값은 통계적으로 유의한 것으로 나타남

Root Mean Square Error
회귀 모형의 적합도 를 측정하는 기준.
오차가 작을수록 모델의 적합도가 더 높다.

총변동 중에서 회귀방정식에 의해서 설명되는 변동

상관관계 없을 확률
회귀모형은 통계적으로 유의한 것으로 나타남

단순회귀분석 – Minitab



단순회귀분석 – Minitab

- **s (Root Mean Square Error) : 오차 제곱합의 제곱근.**

$$= \sqrt{\frac{SSE}{n-k-1}}$$

  - 회귀선과 각각 관측 값과의 표준편차
  - 이 값이 작을수록 모델의 적합도는 높다.

- **R-Sq** : 산출된 모델식의 설명의 정도를 나타내는 변동의 비율로 이 값이 *크면 클수록 정교한 모델식이라 할 수 있다.*

$$= 1 - \frac{SSE}{SST} = \frac{SSR}{SST}$$

  고려하는 입력변수의 수가 증가하면 복잡하고 재현성이 부족한 모델식이 도출될 수 있다.

- **R-Sq(adj)** : R-Sq(adj) 값은 모델에 포함되는 변수들의 개수를 반영한 값으로 *변수들의 수가 증가할수록 R-Sq 값은 커지는 반면, R-Sq(adj) 값은 감소하게 된다.*

$$= 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$$

(k는 regression식에서 독립변수의 개수)

이 값들의 차이가 10% 이상 차이가난다면 fitting에 유의하지 않은 변수가 포함된 것이 아닌지 회귀식을 의심해 볼 필요성이 있다.

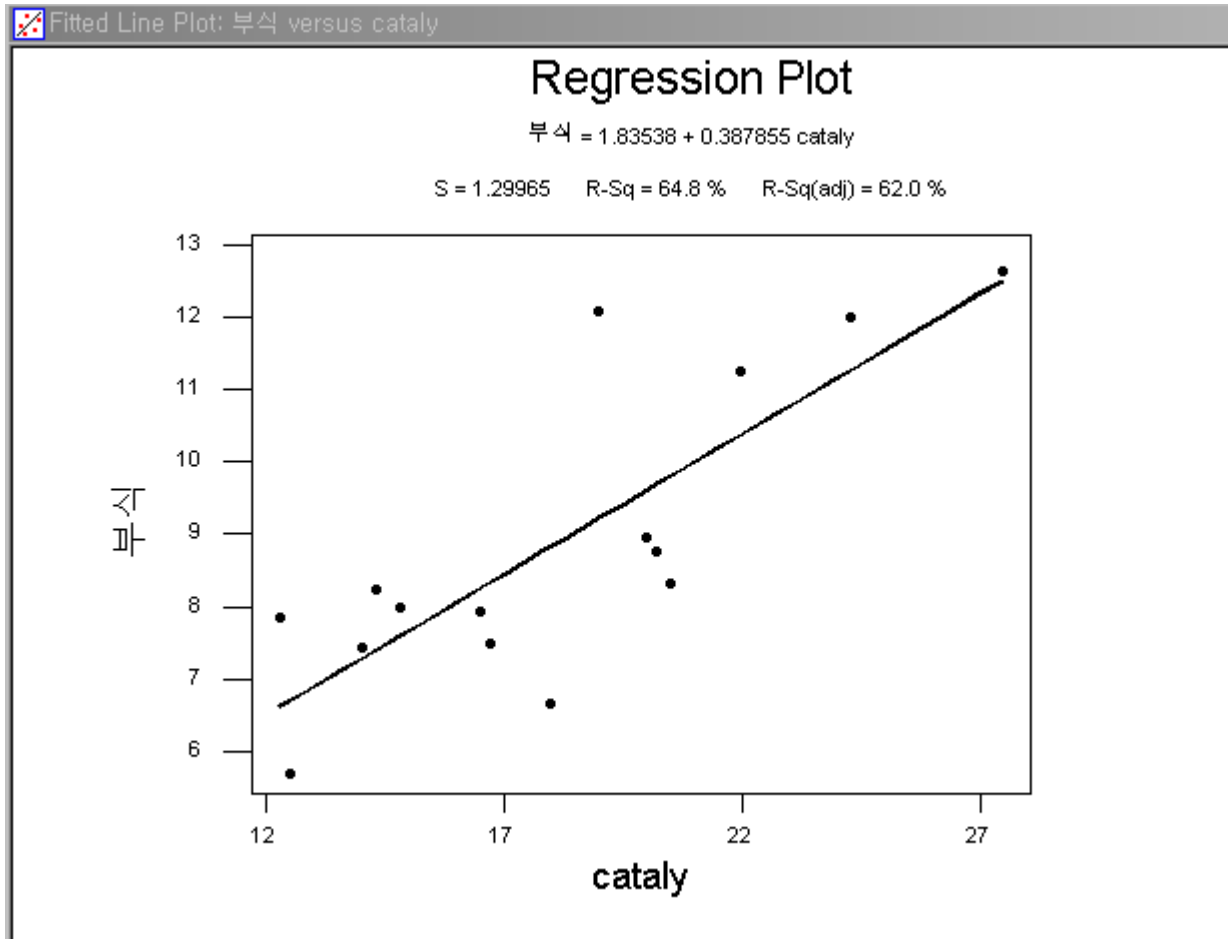
단순회귀 분석-실습

- 연구개발 부서에서 신제품을 개발하기 위하여 프로젝트팀을 결성하고, Survival 프로젝트를 추진하였다. 프로젝트를 진행하면서 황산의 농도가 부식에 영향을 미치는지 확인하기 위하여 다음과 같은 데이터를 확보하였다.

  이를 이용하여 결과를 해석하시오.

황산농도	부식감량
20	8.95
14.8	7.99
20.5	8.31
12.5	5.69
18	6.66
14.3	8.25
27.5	12.63
16.5	7.93
24.3	11.99
20.2	8.76
22	11.26
19	12.08
12.3	7.85
14	7.43
16.7	7.48

단순회귀 분석-Minitab



단순회귀 분석-Minitab

Regression Analysis: 부식 versus cataly

The regression equation is

부식 = 1.84 + 0.388 cataly

Predictor	Coef	SE Coef	T	P
Constant	1.835	1.481	1.24	0.237
cataly	0.38786	0.07936	4.89	0.000

S = 1.300 R-Sq = 64.8% R-Sq(adj) = 62.0%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	40.344	40.344	23.89	0.000
Residual Error	13	21.958	1.689		
Total	14	62.302			

Unusual Observations

Obs	cataly	부식	Fit	SE Fit	Residual	St Resid
12	19.0	12.080	9.205	0.342	2.875	2.29R

R denotes an observation with a large standardized residual

중 회귀 (Multiple Regression) 분석

❑ *모델방정식* $y = b_0 + b_1x_1 + b_2x_2 + \cdot\ \cdot\ \cdot + b_kx_k + e$

❑ 국내 수도권 중심 대리점의 매출액과 관련하여 *Marketing 비용 및 매장 규모가 매출액에* 영향을 주는 잠재인자로 파악하였으며, 관련 대리점의 마케팅 비용 및 매장 규모, 매출액에 대한 현황을 오른쪽 sheet와 같이 확보하였다.

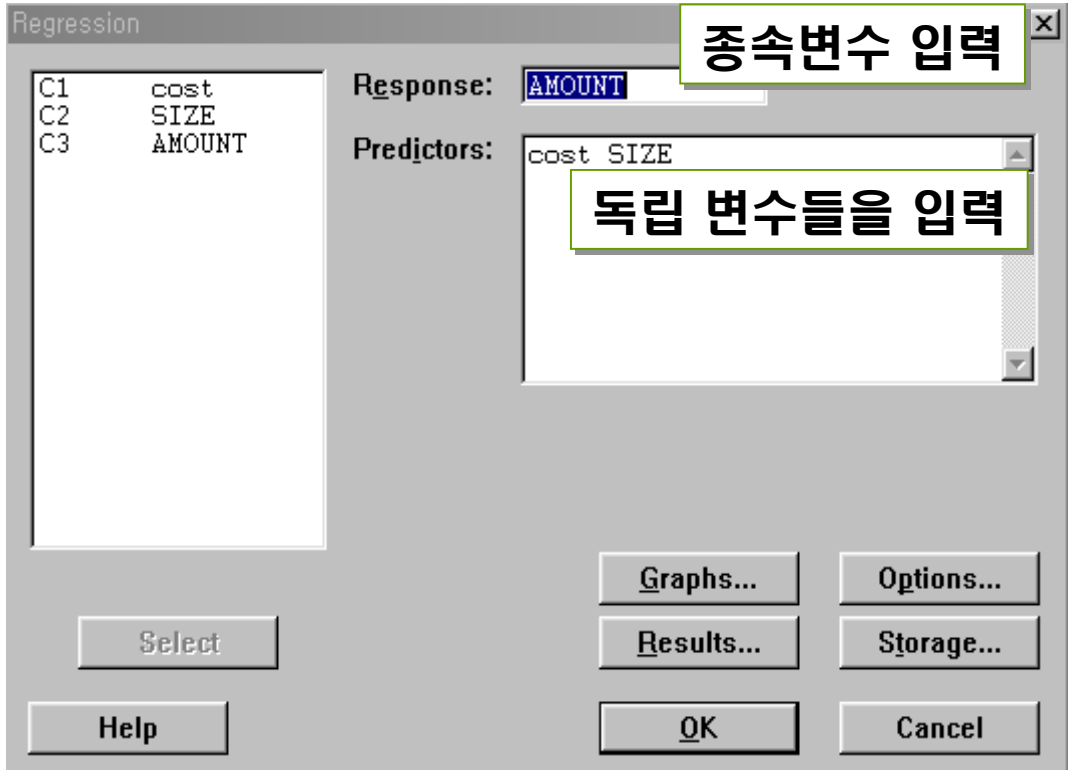
이 데이터를 활용하여 결론을 해석하시오.

cost	SIZE	AMOUNT
4	4	9
8	10	20
9	8	22
8	5	15
8	10	17
12	15	30
6	8	18
10	13	25
6	5	10
9	12	20

중 회귀분석 – Minitab

STAT > Regression > Regression

C1	C2	C3
cost	SIZE	AMOUNT
4	4	9
8	10	20
9	8	22
8	5	15
8	10	17
12	15	30
6	8	18
10	13	25
6	5	10



중 회귀 분석-Minitab

Regression Analysis: AMOUNT versus cost, SIZE

The regression equation is
AMOUNT = - 0.65 + 1.55 cost + 0.760 SIZE

회귀방정식

Predictor	Coef	SE Coef	T	P
Constant	-0.651	2.908	-0.22	0.829
cost	1.5515	0.6462	2.40	0.047
SIZE	0.7599	0.3968	1.91	0.097

S = 2.278 R-Sq = 90.1% R-Sq(adj) = 87.3%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	332.07	166.04	32.00	0.000
Residual Error	7	36.33	5.19		
Total	9	368.40			

중 회귀 분석-Minitab

❑ 자동차 타이어에서 발생되는 열은 타이어에 걸리는 하중, 속도, 두께, 온도 측정시간 5가지 변수에 의하여 영향을 받을 것으로 알려져 실내 주행실험을 실시한 결과 아래의 데이터를 확보하였다.
실험 결과를 분석하시오.

wei	speed	thick	temp	time	y
60	60	3.65	26	3	81
60	60	3.6	26	4	79
60	80	3.7	27	4	95
60	80	3.63	27	4	96
60	100	3.65	29	2	103
60	100	3.6	29	3	104
80	60	3.65	28	3	107
80	60	3.63	28	4	105
80	80	3.66	29	3	115
80	80	3.66	29	4	116
80	100	3.7	28	4	130
80	100	3.56	28	4	131
100	60	3.53	28	5	130
100	60	3.68	25	5	132
100	80	3.53	28	3	140
100	80	3.53	28	4	139
100	100	3.71	28	2	158
100	100	3.56	27	3	156

중 회귀 분석-Minitab

STAT > Regression > Regression

Regression Analysis: y versus wei, speed, thick, temp, time

The regression equation is

y = - 58.7 + 1.25 wei + 0.649 speed + 13.7 thick - 0.940 temp + 0.138 time

Predictor	Coef	SE Coef	T	P	VIF
Constant	-58.72	35.09	-1.67	0.120	
wei	1.25291	0.02853	43.92	0.000	1.1
speed	0.64917	0.03302	19.66	0.000	1.5
thick	13.707	8.054	1.70	0.115	1.2
temp	-0.9400	0.4712	-1.99	0.069	1.4
time	0.1384	0.6366	0.22	0.832	1.5

S = 1.845 R-Sq = 99.6% R-Sq(adj) = 99.4%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	5	9205.4	1841.1	540.66	0.000
Residual Error	12	40.9	3.4		
Total	17	9246.3			

기업의 매출은 사원 수와 사원들의 제 경비의 사용 금액과 관계를 파악하기 위하여 1개월 간 데이터를 수집하였다.
분석 결과를 해석하고, 발표하시오.

- 매출에 영향을 주는 인자는 무엇입니까?

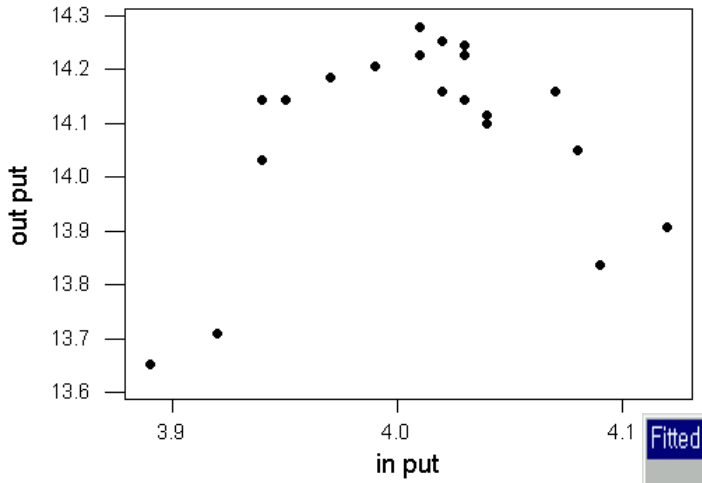
곡선회귀 (Curvilinear Regression Analysis) 분석 – Minitab

- 음질의 떨림 향상 TFT에서는 전압(in put voltage)이 떨림에 큰 영향을 주는 인자로 분석되었고, 오른쪽 sheet와 같이 in put과 Out put 간의 관계를 Data화 하였다.

  주어진 Data를 활용하여 전압과 떨림 간의 관계를 규명하시오.

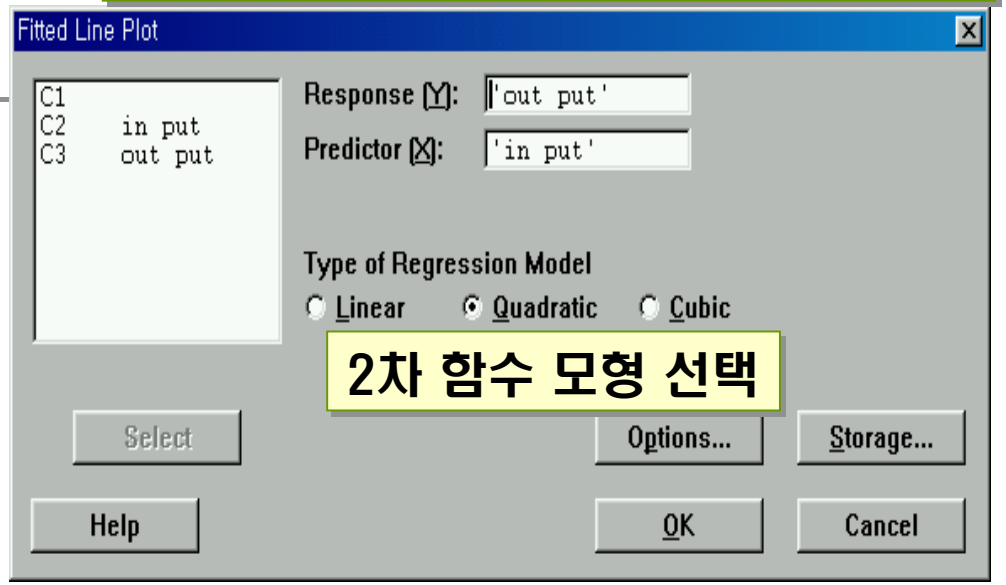
	in put	out put
1	3.97	14.1845
2	3.92	13.7083
3	4.02	14.2539
4	4.03	14.2459
5	4.01	14.2795
6	3.89	13.6520
7	4.03	14.2280
8	4.08	14.0491
9	4.07	14.1589
10	3.99	14.2075
11	3.94	14.0309
12	3.95	14.1451
13	3.94	14.1451
14	4.12	13.9080
15	4.01	14.2280
16	4.03	14.1451
17	4.04	14.1000
18	4.02	14.1589
19	4.04	14.1157
20	4.09	13.8379

곡선회귀 분석 – Minitab



➢ plotting하여 1차 함수의 모형인지, 2차 함수의 모형인지 확인 한다.

Stat ▶Regression ▶Fitted Line Plot

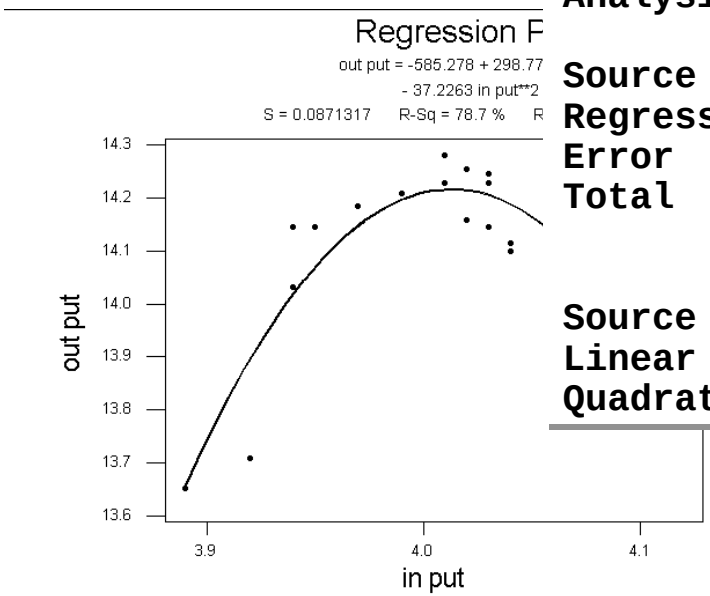


예제 풀이

회귀방정식

Polynomial Regression Analysis: out put versus in put

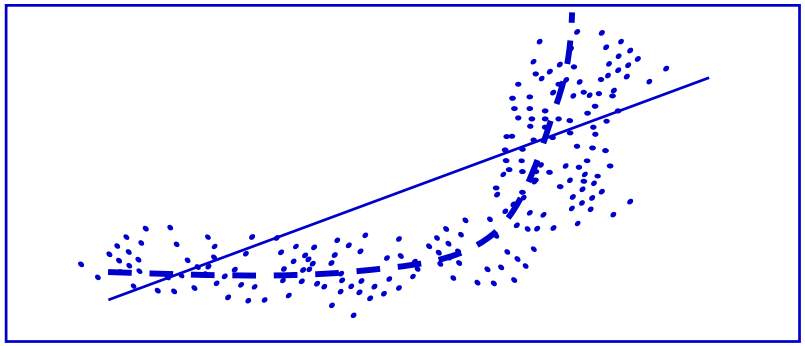
```
The regression equation is
out put = -585.278 + 298.777 in put - 37.2263 in put**2

S = 0.0871317      R-Sq = 78.7 %      R-Sq(adj) = 76.2 %

Analysis of Variance

Source              DF         SS          MS          F      P
Regression           2   0.477570    0.238785    31.4524  0.000
Error               17   0.129063    0.007592
Total               19   0.606633


Source       DF     Seq SS          F      P
Linear        1   0.035563     1.1209  0.304
Quadratic     1   0.442007    58.2206  0.000
```



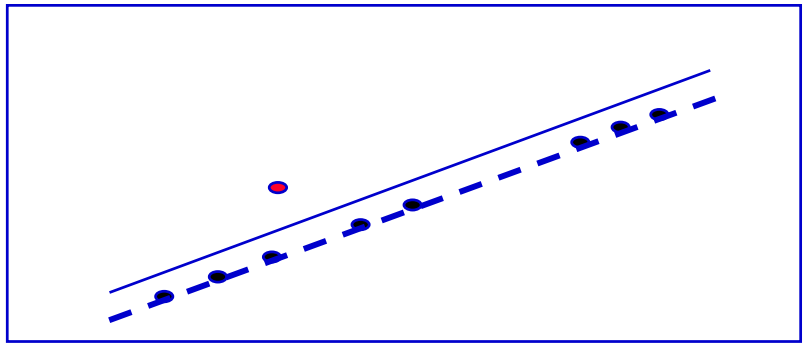
효과적인 회귀분석

- R^2 값 — 높을 수록 좋다
- 회귀 적합의 ANOVA — $p < 0.05$
- 효과적 데이터 적합(Valid Data Fit) — Residual이 정규적이다

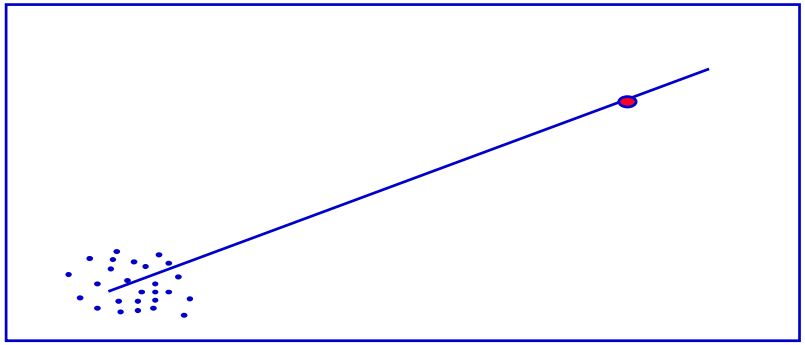
범하기 쉬운 오류



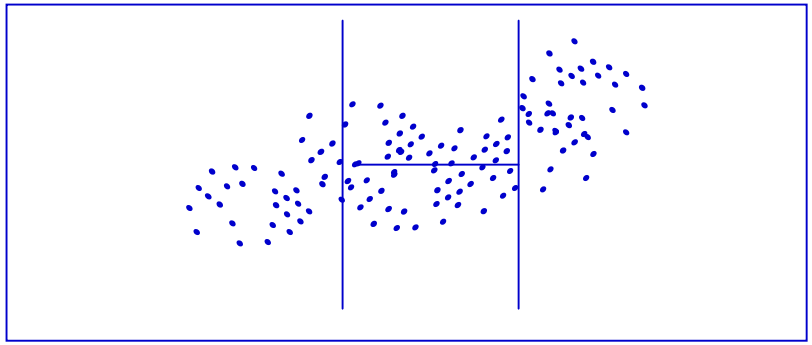
비선형 데이터



이상점 (Outliers)



모여 있는 데이터



데이터 범위가 좁을 경우

데이터 수집은 인자의 낮은 한계 값에서 높은 한계 값까지 넓은 범위의 데이터를 사용하라.

요약

- 프로세스를 모니터링하거나 실험에 의해서 데이터가 수집된 경우 보편적 분석 방법이 회귀 분석이다.
- 회귀 분석은 둘 이상의 변수사이의 관계를 만들어 분석하는 기법이다.
- 회귀 분석은 가장 널리 사용되기는 하나 또한 남용되기도 하는 통계 기법이다.
- ANOVA 와 회귀 분석은 밀접하게 연관되어 있다.