5강

중회귀모형 (2)

정보통계학과 김성수교수

학습목차

1 표준화된 중회귀분석

변수 표준화

중회귀모형 $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \epsilon_i$

에서 종속변수와 독립변수를 다음과 같이 변수변환

$$Y_i^* = \frac{Y_i - \overline{Y}}{\sqrt{S_{YY}}}, \quad Z_{ij} = \frac{X_{ij} - \overline{X}_j}{\sqrt{S_{jj}}}$$

$$, \; S_{YY} = \sum (Y_i - \overline{Y})^2 \;, \quad S_{jj} = \sum_i (X_{ij} - \overline{X}_j)^2$$

$\overline{X}_j = j$ 번째 독립변수의 평균값

$$\Rightarrow \quad \sum Y_i^* = 0, \sum (Y_i^*)^2 = 1$$

$$\sum_i Z_{ii} = 0, \sum_i (Z_{ii})^2 = 1, \; (j = 1, 2, \cdots, k)$$

$\Rightarrow$ Y_i^*, Z_{ij} 를 표준화된 변수(standardized variables)라 함.

변수 표준화

✓ 표준화된 회귀모형

표준화된 중회귀모형

$$Y_i^* = \alpha_1 Z_{i1} + \alpha_2 Z_{i2} + \cdots + \alpha_k Z_{ik} + \epsilon_i'$$

참고 : 절편항 α_0의 추정값은 항상 0이 됨.

※ 표준화된 중회귀모형에서 추정된 회귀계수 a_i 의 절대값이 크면 클수록 설명변수 X_i가 반응변수 Y_i 에 주는 영향이 크게 됨.

R 활용 : 표준화 회귀모형

```
> install.packages("lm.beta")
> library(lm.beta)
> market2.lm = lm(Y ~ X1+X2, data=market2)
> market2.beta = lm.beta(market2.lm)
> print(market2.beta)
---
Standardized Coefficients::
(Intercept)          X1          X2
  0.0000000   0.7015566   0.3376137
> summary(market2.beta)
---
Coefficients:
            Estimate Standardized Std. Error t value Pr(>|t|)
(Intercept)  0.85041      0.00000    0.84624   1.005 0.334770
X1           1.55811      0.70156    0.14793  10.532 2.04e-07 ***
X2           0.42736      0.33761    0.08431   5.069 0.000276 ***
---
Residual standard error: 0.9318 on 12 degrees of freedom
Multiple R-squared:  0.9799,    Adjusted R-squared:  0.9765
F-statistic: 292.5 on 2 and 12 DF,  p-value: 6.597e-11
```

적합된 표준화 회귀모형

$$\hat{Y}^* = 0.7016Z_1 + 0.3376Z_2$$

※ 여기서 X1의 표준화계수가 X2의 표준화계수보다 크므로 상대적으로 X1의 영향이 더 큼을 알 수 있음.

2 추정과 검정

추정된 회귀계수의 분산

회귀계수벡터 $\beta' = (\beta_0, \beta_1, \beta_2, \cdots, \beta_k)$의 β의 추정량 $\hat{\beta}$

$$\begin{aligned} E(\hat{\beta}) &= E[(X'X)^{-1}X'Y] \\ &= (X'X)^{-1}X'E(Y) \\ &= (X'X)^{-1}X'X\beta \\ &= \beta \end{aligned}$$

이므로 $\hat{\beta}$는 β의 불편추정량. $\hat{\beta}$의 분산-공분산 행렬은

$$\begin{aligned} Var(\hat{\beta}) &= Var[(X'X)^{-1}X'Y] \\ &= (X'X)^{-1}X'[Var(Y)]X(X'X)^{-1} \\ &= (X'X)^{-1}X'(I\sigma^2)X(X'X)^{-1} \\ &= (X'X)^{-1}X'X(X'X)^{-1}(I\sigma^2) \\ &= (X'X)^{-1}\sigma^2 \end{aligned}$$

구간추정

독립변수들의 값 $(x_1, x_2, \cdots, x_k)$**에서** $E(Y)$**의 구간추정**

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

$$= (1, x_1, x_2, \cdots, x_k)\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_2 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{pmatrix}$$

$$= \boldsymbol{x}'\hat{\boldsymbol{\beta}}$$

$$\Rightarrow \quad Var(\hat{Y}) = Var(\boldsymbol{x}'\hat{\boldsymbol{\beta}}) = \boldsymbol{x}'\,Var(\hat{\boldsymbol{\beta}})\boldsymbol{x}$$
$$= \boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}\sigma^2$$

$\Rightarrow$ $E(Y)$**의** $100(1-\alpha)\%$ **신뢰구간**

$$\hat{Y} \pm t(n-k-1;\alpha/2)\sqrt{\boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}\cdot \; MSE}$$

회귀계수 가설검정

회귀계수 β_i 의 가설 및 검정통계량

가설

$$H_0 : \beta_i = \beta_{i0}$$

$$H_1 : \beta_i \neq \beta_{i0}$$

검정통계량

$$t_0 = \frac{b_i - \beta_{i0}}{\sqrt{c_{ii} \cdot MSE}}$$

, c_{ii} : $Var(\hat{\beta}) = (X'X)^{-1}\sigma^2$ 의 대각선 값

R 활용 예 : 신뢰구간

마켓데이터에 대하여 $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ 을 적합시켰을 때

(1) $x_1 = 10, x_2 = 10$에서 $E(y)$를 95% 신뢰구간으로 추정하고,

(2) $H_0 : \beta_1 = 0$, $H_0 : \beta_2 = 0$ 에 대하여 유의수준 $\alpha = 0.05$ 로 가설검정 하여보자.

```
> # 1 : 95% ( 99% 신뢰구간 추가 )
> pred.x = data.frame(X1=10, X2=10)
> pc = predict(market2.lm, int="c", newdata=pred.x)
> pc
       fit      lwr      upr
1 20.70503 19.95796 21.45209
> pc99 = predict(market2.lm, int="c", level=0.99, newdata=pred.x)
> pc99
       fit      lwr      upr
1 20.70503 19.65769 21.75236
```

R 활용 예 : 회귀계수 검정

```
> summary(market2.lm)
---
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.85041    0.84624   1.005 0.334770
X1           1.55811    0.14793  10.532 2.04e-07 ***
X2           0.42736    0.08431   5.069 0.000276 ***
---
```

Coefficients: 결과에서 $\hat{\beta}_1 = 1.55811$ 이고, 표준오차는 0.14793

$$t-\text{값} = \frac{1.55811}{0.14793} = 10.532$$

⇒ 유의확률 $p-\text{값} = 2.04 \times 10^{-7}$ 이 되므로 $H_0 : \beta_1 = 0$ 에 대한 귀무가설을 기각.

⇒ $H_0 : \beta_2 = 0$ 도 $p-\text{값} = 0.000276$ 이므로 귀무가설을 기각.

일반적 모형비교

✓ 두 모형을 비교하는 일반적인 방법

(예) 단순회귀모형 : $H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$ 의 경우

완전모형(Full Model) : $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

축소모형(Reduced Model) : $Y_i = \beta_0 + \epsilon_i$

, 완전모형 : 데이터에 잘 적합되리라고 고려되는 모형

, 축소모형 : 귀무가설 $H_0 : \beta_1 = 0$ 의 가정하에서의 모형

※ 데이터를 적합하는 데에 있어서 완전모형과 축소모형 간에 유의한 차이가 있느냐 없느냐를 고려하여 두 모형을 비교

일반적 모형비교

,

- 두 모형의 비교는 잔차제곱합의 차이를 이용함.
- 완전모형의 경우 잔차제곱합

$$SSE(F) = \sum [Y_i - (b_0 + b_1 X_i)]^2 = \sum (Y_i - \hat{Y}_i)^2 = SSE$$

- 축소모형 $Y_i = \beta_0 + \epsilon_i$ 에서 잔차제곱합

$$SSE(R) = \sum (Y_i - b_0)^2 = \sum (Y_i - \overline{Y})^2 = SST$$

- 두 모형을 비교하기 위한 검정통계량

$$F_0 = \frac{[SSE(R) - SSE(F)]/(df_R - df_F)}{SSE(F)/df_F}$$

$\Rightarrow$

$$F_0 = \frac{(SST - SSE)/[(n-1)-(n-2)]}{SSE/(n-2)} = \frac{SSR/1}{SSE/(n-2)} = \frac{MSR}{MSE}$$

3 변수추가

추가제곱합

✓ 중회귀모형을 적합하는데 있어서 어떤 특정한 변수를 회귀모형에 포함시키는 것이 바람직한가를 결정하고 싶은 경우

- 이 변수를 포함시키지 않고 구한 회귀제곱합에서 이 변수를 포함시키고 구한 회귀제곱합(regression sum of squares)이 추가적으로 어느 정도 커졌는가를 검토. 이와 같은 경우에 추가적으로 증가된 제곱합을 추가제곱합(extra sum of squares)이라고 함.

➢ 추가제곱합은 새로운 변수가 모형에 추가될 때의 회귀제곱합의 증가분을 나타내는 것으로서 이 값이 작을수록 회귀에 대한 기여도가 떨어진다는 것을 의미.

R 활용 : 추가제곱합

```
> health = read.table("c:/data/reg/health.txt", header=T)
> head(health,3)
  ID  X1 X2  X3 X4   Y
1  1 217 67 260 91 481
2  2 141 52 190 66 292
3  3 152 58 203 68 338
> h1.lm = lm(Y ~ X1, data=health)
> h2.lm = lm(Y ~ X1+X4, data=health)
> anova(h1.lm, h2.lm)
Analysis of Variance Table

Model 1: Y ~ X1
Model 2: Y ~ X1 + X4
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1     28 50795
2     27 24049  1     26746 30.027 8.419e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

< 헬스클럽 자료 >

번호	X_1	X_2	X_3	X_4	Y
1	217	67	260	91	481
2	141	52	190	66	292
3	152	58	203	68	338
....					
28	245	70	218	69	469
29	141	63	193	60	252
30	177	53	183	75	338

X_1 모형에서 X_4가 추가된 경우의 추가제곱합

$$\begin{aligned} SSR(X_4 \mid X_1) &= SSR(X_1, X_4) - SSR(X_1) \\ &= SSE(X_1) - SSE(X_1, X_4) \\ &= 26746 \end{aligned}$$

$$\begin{aligned} F_0 &= \frac{[SSE(R) - SSE(F)]/(df_R - df_F)}{SSE(F)/df_F} \\ &= \frac{50795 - 24049}{24049/27} \\ &= 30.027 \end{aligned}$$

추가변수그림

✓ 추가변수그림(added variable plot)

- 중회귀모형에서 새로운 변수선택은 기존의 모형이 설명하지 못하는 부분을 새로운 변수가 들어옴으로써 추가설명력이 얼마나 유의한 가에 따라 결정

⇒ 새로운 변수의 효과를 그래프로 표현할 수 있는데, 이러한 그래프 중의 하나가 **추가변수그림**(added variable plot)임.

이를 **편회귀그림**(partial regression plot)이라고도 함.

추가변수그림 그리는 절차

독립변수가 두 개인 회귀모형에서 변수 X_2의 추가변수그림을 그리는 절차

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

① Y를 X_1으로 회귀한 후 얻어지는 잔차, $e(Y \mid X_1)$ 을 구한다.

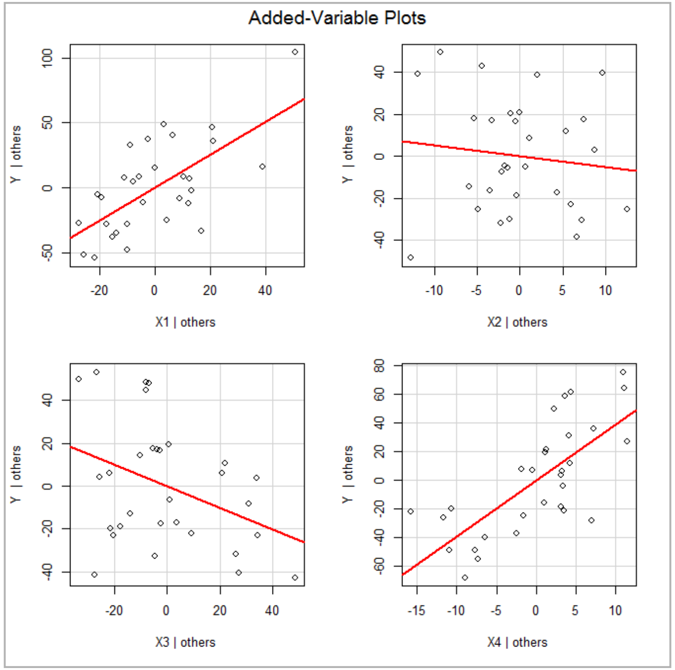
② X_2를 X_1으로 회귀한 후 얻어지는 잔차, $e(X_2 \mid X_1)$ 을 구한다.

③ 앞에서 구한 두 잔차에서 x-축을 $e(X_2 \mid X_1)$, y-축을 $e(Y \mid X_1)$으로 한 산점도를 추가변수그림이라 함.

⇒ 추가변수그림이 선형관계가 있으면 변수 X_2 는 추가적인 설명력이 있다고 판단.

R 활용 : 추가변수그림

```
> library(car)
> h4.lm = lm(Y ~ X1+X2+X3+X4, data=health)
> avPlots(h4.lm)
```



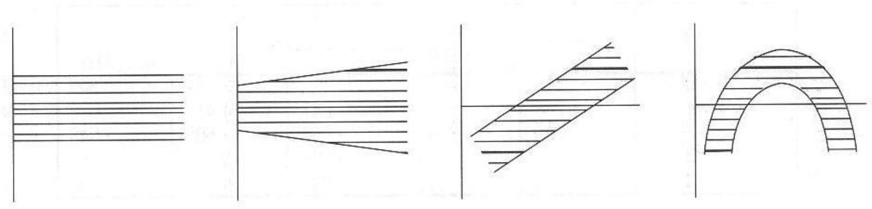
X_1 추가변수그림 : 잔차 $e(X_1 | X_2, X_3, X_4)$ 와 잔차 $e(Y | X_2, X_3, X_4)$의 산점도

해석 : 변수 X_1과 X_4에 대한 추가변수그림이 선형성이 강한 것을 볼 수 있음. 따라서 이 두 변수가 회귀모형에 매우 유의.

4 잔차검토 및 분석사례

잔차의 검토

✓ 잔차의 산점도를 그려보면 중회귀모형의 가정들이 옳았는가를 검토할 수 있음.



(a) (b) (c) (d)

ⓐ 가정에 아무런 모순이 없는 것으로 판정된다.

ⓑ 분산이 일정하지 않으며, 가중회귀(weighted regression)를 쓰거나 또는 Y_i를 변환시켜 회귀분석함이 바람직하다.

ⓒ 절편이 필요한 모형인데 절편을 사용하지 않았을 경우에 생길 수 있는 형태이다.

ⓓ 모형이 타당하지 않다. 추가적으로 독립변수의 제곱항이 필요하다. 또는 Y_i의 적절한 변환이 필요하다.

분석사례

✓ 예제자료

< NH_3를 HNO_3로 산화시키는 공정 >

X_1=공정의 작업속도(SPEED)

X_2=냉각수의 온도(TEMP)

Y = NH_3를 HNO_3로 바꿀 때 손실되는 NH_3의 함량%(LOSS)

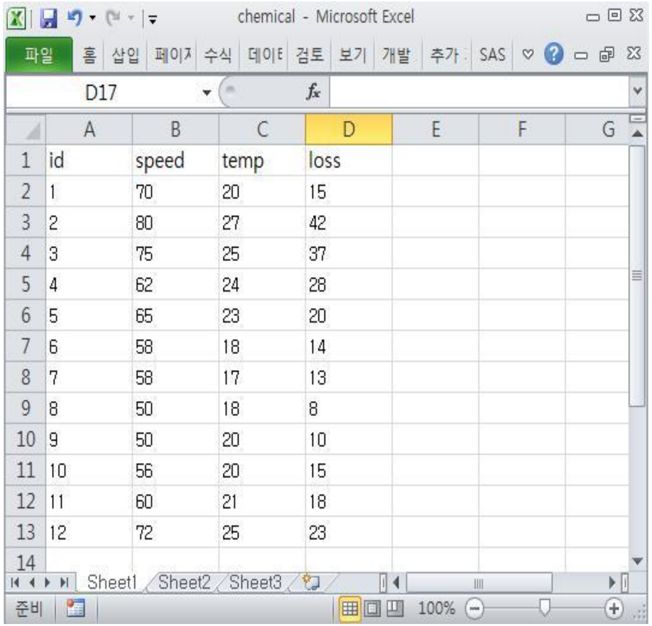
< 화학공장 데이터 >

실험번호	SPEED(X_1)	TEMP(X_2)	LOSS (Y)
1	70	20	15
2	80	27	42
3	75	25	37
4	62	24	28
5	65	23	20
6	58	18	14
7	58	17	13
8	50	18	8
9	50	20	10
10	56	20	15
11	60	21	18
12	72	25	23

1) 자료파일 읽기

엑셀파일 : chemical.xlsx

```
> install.packages("xlsx")
> library(xlsx)
> chemical = read.xlsx("c:/data/reg/chemical.xlsx", 1)
> head(chemical)
  id speed temp loss
1  1    70   20   15
2  2    80   27   42
3  3    75   25   37
4  4    62   24   28
5  5    65   23   20
6  6    58   18   14
```



	A	B	C	D
1	id	speed	temp	loss
2	1	70	20	15
3	2	80	27	42
4	3	75	25	37
5	4	62	24	28
6	5	65	23	20
7	6	58	18	14
8	7	58	17	13
9	8	50	18	8
10	9	50	20	10
11	10	56	20	15
12	11	60	21	18
13	12	72	25	23

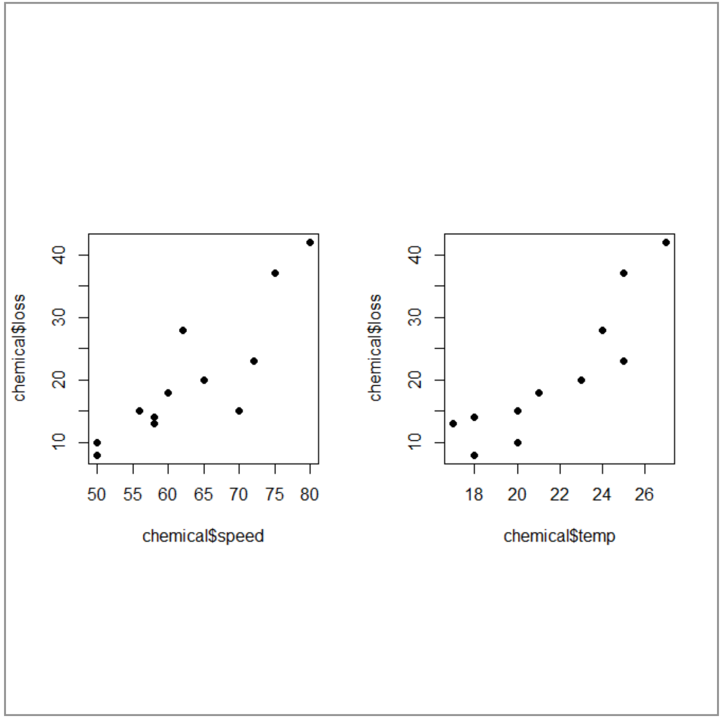
2) 기술통계량 및 상관계수 보기

```
> summary(chemical[,-1])
     speed            temp            loss      
 Min.   :50.0   Min.   :17.00   Min.   : 8.00  
 1st Qu.:57.5   1st Qu.:19.50   1st Qu.:13.75  
 Median :61.0   Median :20.50   Median :16.50  
 Mean   :63.0   Mean   :21.50   Mean   :20.25  
 3rd Qu.:70.5   3rd Qu.:24.25   3rd Qu.:24.25  
 Max.   :80.0   Max.   :27.00   Max.   :42.00  
> cor(chemical[,-1])
          speed      temp      loss
speed 1.0000000 0.8023847 0.8548423
temp  0.8023847 1.0000000 0.8953498
loss  0.8548423 0.8953498 1.0000000
```

상관계수에서 독립변수들과 종속변수간의 상관관계가 높다는 것을 알 수 있음. 또한 독립변수들간(SPEED 와 TEMP)의 상관계수가 0.802 로서 높다는 것도 알 수 있음.

3) 산점도 그리기

```
> par(mfrow=c(1,2), pty="s")
> plot(chemical$speed, chemical$loss, pch=19)
> plot(chemical$temp, chemical$loss, pch=19)
```



4) 회귀모형 적합하기

```
> chemical.lm = lm(loss ~ speed + temp,  data=chemical)
> summary(chemical.lm)
---
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -47.6243     9.4580  -5.035 0.000704 ***
speed         0.4216     0.2350   1.794 0.106360
temp          1.9217     0.6977   2.754 0.022316 *
---
Residual standard error: 4.465 on 9 degrees of freedom
Multiple R-squared:  0.8539,    Adjusted R-squared:  0.8214
F-statistic:  26.3 on 2 and 9 DF,  p-value: 0.0001741
```

추정된 회귀방정식 : $\hat{Y} = -47.624 + 0.422 \cdot speed + 1.922 \cdot temp$

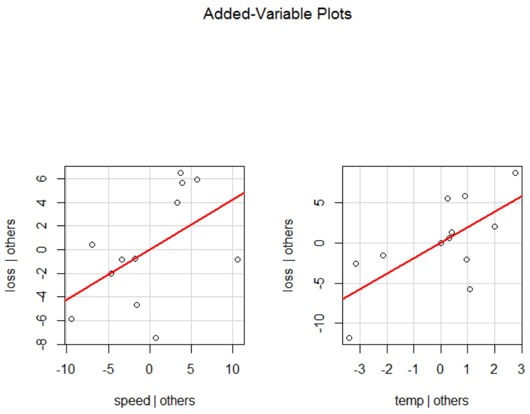
결정계수 R^2=0.854 .

변수 speed 의 p-값 = 0.106 , 변수 temp의 p-값 = 0.0223

유의수준 α=0.05를 기준으로 할 때, temp는 loss를 설명하는데 유의하나, 변수 speed는 유의하지 못함.

추가변수그림

```
> library(car)
> avPlots(chemical.lm)
```



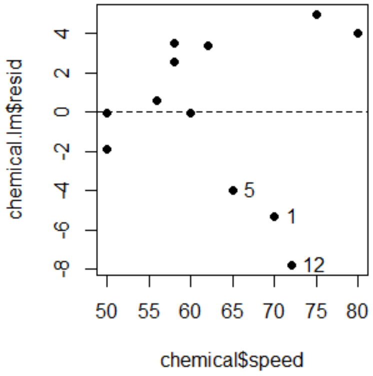
5) 분산분석표 구하기

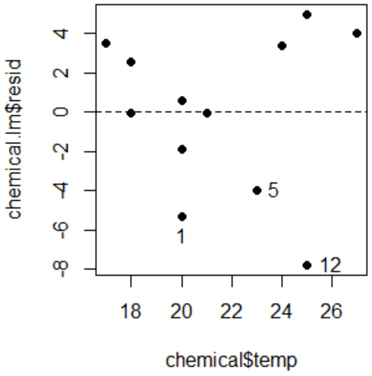
```
> anova(chemical.lm)
Analysis of Variance Table

Response: loss
          Df Sum Sq Mean Sq F value    Pr(>F)
speed      1 897.55  897.55 45.0179 8.758e-05 ***
temp       1 151.26  151.26  7.5867   0.02232 *
Residuals  9 179.44   19.94
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

요인	자유도	제곱합	평균제곱	F_0	Pr(>F)
회귀	2	1048.81	524.41	26.3	0.000174
잔차	9	179.44	19.94		
계	11	1076.99			

6) 잔차 산점도 : (독립변수, 잔차)

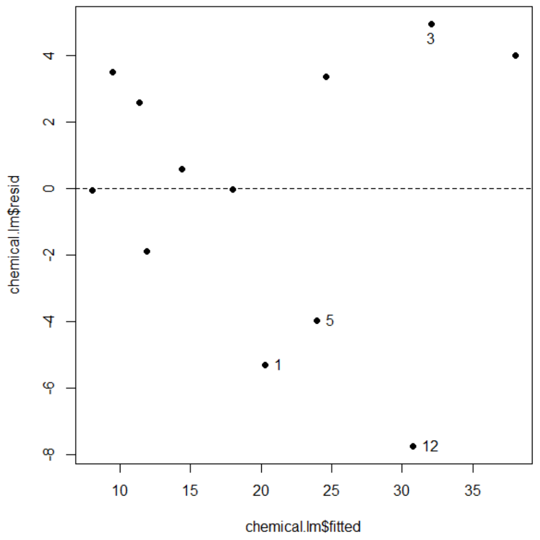
```
> par(mfrow=c(1,2), pty="s")
> plot(chemical$speed, chemical.lm$resid, pch=19)
> abline(h=0, lty=2)
> identify(chemical$speed, chemical.lm$resid)
[1]  1  5 12
> plot(chemical$temp, chemical.lm$resid, pch=19)
> abline(h=0, lty=2)
> identify(chemical$temp, chemical.lm$resid)
[1]  1  5 12
```





7) 잔차 산점도 : (추정값, 잔차)

```
> par(mfrow=c(1,1))
> plot(chemical.lm$fitted, chemical.lm$resid, pch=19)
> abline(h=0, lty=2)
> identify(chemical.lm$fitted, chemical.lm$resid)
[1]  1  3  5 12
```



다음시간 안내

6강. SAS, SPSS를 활용한 회귀모형 적합