8.2장 다중회귀분석

박명현

서울대학교 학부대학

다중회귀분석(multiple regression analysis)

- 일반적으로 반응변수 Y의 변화는 두 개 이상의 설명변수 $(x_1, x_2, \cdots, x_k)$에 의하여 영향을 받는 경우가 많이 있다.

예) 광고비와 설비투자가 매출액에 미치는 영향

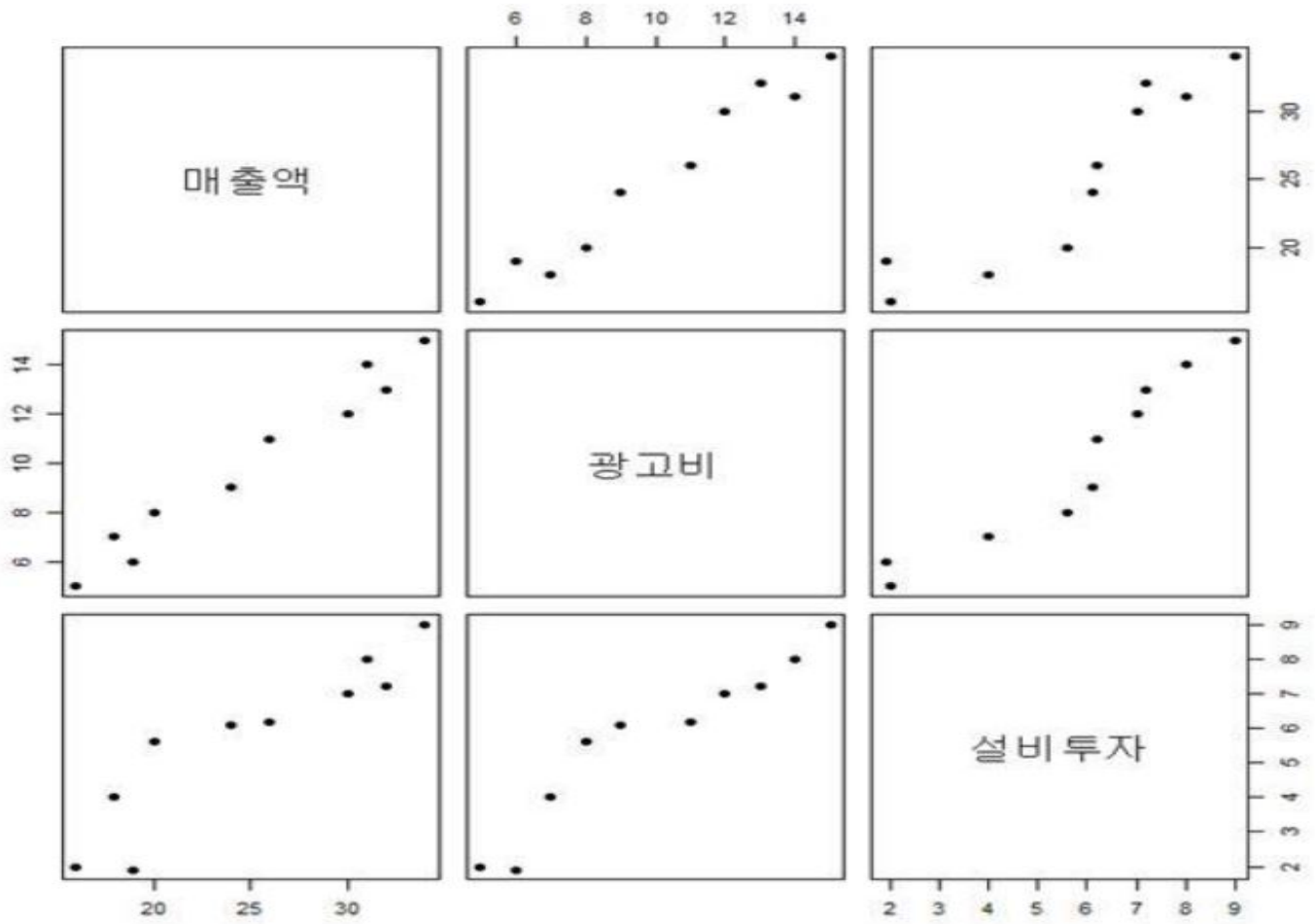
- 반응변수 Y: 매출액
- 첫 번째 설명변수 x_1: 광고비
- 두 번째 설명변수 x_2: 설비투자

$\Longrightarrow$ 하나의 설명변수만을 고려하는 단순회귀모형과 달리 두 개 이상(multiple) 의 설명변수로 반응변수를 설명

예제

- 광고비 및 설비투자와 매출액의 연관성 비교 자료(단위: 억원)

Figure: Scatter plot matrix



다중선형회귀모형(multiple linear regression model)

- 다음의 다중선형회귀모형을 고려해보자.

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i \quad (i = 1, 2, \cdots, n)$$

예) 광고비(x_{1i})와 설비투자(x_{2i})로 기업의 매출액(y_i) 설명하기

	y_매출액	x1_광고비	x2_설비투자
0	16	5	2.0
1	19	6	1.9
2	18	7	4.0
3	20	8	5.6
4	24	9	6.1
5	26	11	6.2
6	30	12	7.0
7	32	13	7.2
8	31	14	8.0
9	34	15	9.0

다중선형회귀모형

- 다중선형회귀모형

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i \quad (i = 1, 2, \cdots, n)$$

  - 반응변수, 종속변수 $Y_1, \cdots, Y_n$
  - k개의 설명변수, 독립변수 $x_{1i}, \cdots, x_{ki}$
    (확률변수가 아닌 상수로 취급)
  - $\beta_0, \beta_1, \cdots, \beta_k$: 미지의 모회귀계수(실수)
  - ϵ_i: 기댓값이 0, 분산이 σ^2인 오차항을 나타내는 확률변수
  - $\sigma^2 > 0$: 미지의 오차분산

- 다중선형회귀식

$$E(Y_i | x_{i1}, \cdots, x_{ki}) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} \quad (i = 1, 2, \cdots, n)$$

다중선형회귀모형의 가정사항

- 오차항의 가정사항: $\epsilon_i \sim_{i.i.d} N(0, \sigma^2)$ $(\sigma^2 unknown)$ $(i = 1, \cdots, n)$

  1) 오차항(ϵ_i)의 평균은 0이다: $(E(\epsilon_i) = 0)$
  2) 오차항의 분산 σ^2는 x_{ji}의 모든 값에 대하여 항상 같다.
     (등분산 가정: $Var(\epsilon_i) = \sigma^2 > 0$)
  3) 오차항 ϵ_i와 ϵ_j는 서로 독립이다 (for all $i, j; i \neq j$: 독립성)
  4) 오차항은 정규분포를 따른다. $\epsilon_i \sim N(0, \sigma^2)$

- Y변수는 각각 평균 $\beta_0 + \sum_{j=1}^{k} \beta_j x_{ji}$, 분산 σ^2인 정규분포를 하고 서로 독립이다.

$$Y_i \sim N(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ji}, \sigma^2) \ (\sigma^2 unknown)$$

최소제곱법(Least squares method)

- 최소제곱법(Least squares method)
  다중선형회귀모형 $Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ji} + \epsilon_i \ (i = 1, 2, \cdots, n)$에서 오차의 제곱합 $\sum_{i=1}^{n} \epsilon_i^2$을 최소로 하는 $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$을 그 추정값(estimated value)으로 찾는다.

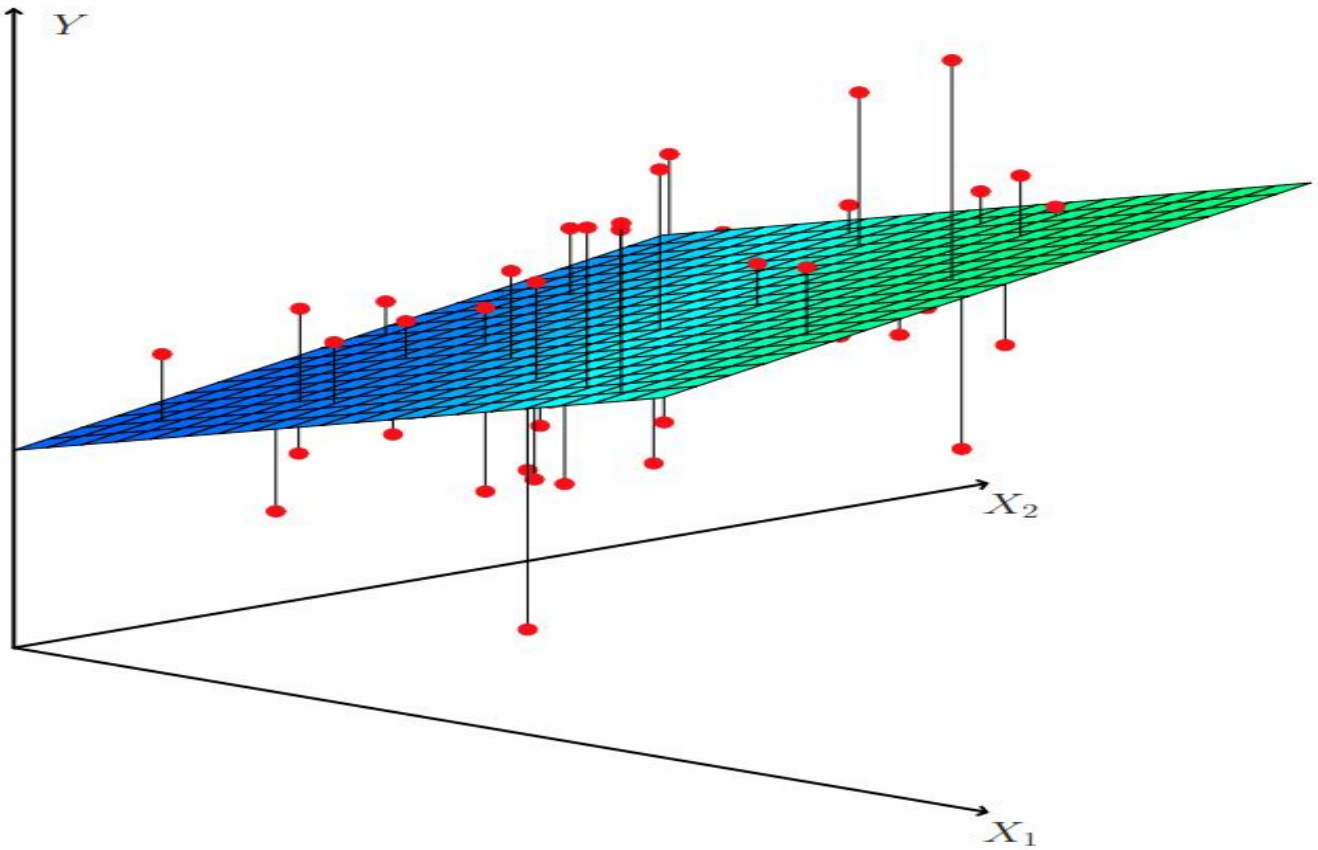
$$(\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k) = \arg\min_{(\beta_0, \beta_1, \cdots, \beta_k)} \sum_{i=1}^{n} \epsilon_i^2$$

- 최소제곱추정값 $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$는 오차의 제곱합(SS)을 $\beta_j, (j = 0, 1, \cdots, k)$에 대하여 각각 편미분하여 얻은 식을 0으로 놓은 다음의 연립방정식을 풀면된다.

$$\frac{\partial SS}{\partial \beta_j} = 0, \quad (j = 0, 1, \cdots, k)$$

$*$ 계산은 실제로 컴퓨터로 수행하게 된다.

최소제곱법(Least squares method)



- In a three-dimensional setting, with two predictors and one response, the least squares regression line becomes a plane.
- The plane is chosen to minimize the sum of the squared vertical distances between each observation (shown in red) and the plane.

$E(Y_i)$와 $\hat{y}_i$

- 추정된 다중선형회귀모형의 회귀식

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_k x_{ki} \quad (i = 1, 2, \cdots, n)$$

- $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$: 모회귀계수 $\beta_0, \beta_1, \cdots, \beta_k$의 최소제곱추정값.

- $y_i = E(y_i|x) + \epsilon_i$

  $y_i = \hat{y}_i + e_i$

- 회귀분석의 목적:
  다중선형회귀식(반응변수 Y의 기대값)을 추정하는 것.

추정계수 $\hat{\beta}_j$의 해석

- 추정계수 $\hat{\beta}_j$ $(j = 1, \cdots, k)$의 해석:

  j번째 설명변수 x_j를 제외한 나머지 설명변수 $(x_1, \cdots, x_{j-1}, x_{j+1}, \cdots, x_p)$들이 고정되어 있다는 가정 하에 x_j가 1단위 변화할 때 y의 변화량이다.

- 추정된 $\hat{\beta}_0$은 $x_1 = x_2 = \cdots = x_k = 0$일 때 Y절편으로 설명변수 값들의 범위가 0을 포함하지 않을 경우 큰 의미를 두지 않는다. 하지만 회귀식의 적합성을 높이기 위하여 모형에 포함한다.

추정계수 $\hat{\beta}_j$의 해석 예제

예) 매출액 예제에서 최소제곱법에 의한 추정회귀식은 다음과 같다. (x_1: 광고비 x_2 : 설비투자)

$$\hat{y} = 6.0985 + 2.2567x_1 - 0.643x_2$$

- 추정계수 2.2567[-0.643]의 의미: $x_2[x_1]$가 고정되어 있을 때 $x_1[x_2]$가 1단위(1억원) 증가할 때 매출액이 2.2567단위[0.643단위] 증가[감소]함을 의미한다.
- 추정계수 6.0985의 의미: $x_1, x_2 = 0$일 때의 매출액을 나타냄. 설명변수들의 범위가 0을 포함하지 않는다면 큰 의미를 두지 않는다.

오차분산 σ^2의 추정

- 잔차(residual): $e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ji}, \quad i = 1, 2, \cdots, n$
- 잔차의 값들이 작다는 것의 의미는 최소제곱회귀직선에 의한 적합이 실제 관측 결과를 잘 설명해주는 것이다.
- 다중선형회귀모형에서의 오차분산 σ^2도 오차제곱평균(MSE)으로 σ^2을 추정한다.

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-k-1} = \frac{1}{n-k-1} \sum_{i=1}^{n} e_i^2$$

이 때, $E(MSE) = \sigma^2$이다.

제곱합의 분해(Decomposition)

- 총제곱합(SST)은 잔차제곱합(SSE)과 회귀제곱합(SSR)으로 분해된다.

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

$$SST \quad = \quad SSE \quad + \quad SSR$$

- 설명변수가 k개일 때, 이들의 자유도는

$$n-1 \quad = \quad (n-k-1) \quad + \quad k$$

- 오차제곱평균(MSE), 회귀제곱평균(MSR), 총제곱평균(MST)은

$$MSE = \frac{SSE}{(n-k-1)}, \qquad MSR = \frac{SSR}{k}, \qquad MST = \frac{SST}{n-1}$$

결정계수 R^2

- 결정계수 R^2

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

  - 다중회귀분석에서는 독립변수가 추가되면 R^2가 증가함
    => 과적합(overfitting)의 문제 발생

- 수정된 결정계수 *Adjusted* R^2

$$Adj\ R^2 = 1 - \frac{MSE}{MST} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$$

  - 회귀모형에 변수를 추가하는것에 패널티를 줌. 따라서 유용한 변수를 추가해야 *Adjusted* R^2가 증가함

- $Adj\ R^2 \leq R^2$

다중선형회귀모형에 대한 유의성 F검정

- 회귀직선이 의미가 있는지를 검정하기 위해서 모든 설명변수들과 반응변수의 선형적 관계를 검토해보자.
- 귀무가설 및 대립가설

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \quad H_1 : \textit{At least one } \beta_j \neq 0 \ \ (j = 1, \cdots, k)$$

- F검정은 회귀모형의 모든 모수의 유의성에 대한 총체적검정(overall test for significance of all parameters)을 수행한다.

(일반적으로 β_0는 제외)

- 검정통계량

$$F = \frac{MSR}{MSE} = \frac{SSR/k}{SSE/(n-k-1)}$$

(SSR: 회귀식에 의한 변동량, SSE: 오차에 의한 변동량)

다중선형회귀모형에 대한 유의성 F검정

- 귀무가설 H_0가 사실일 때 F 검정통계량은 다음의 F분포를 따른다.

$$F \sim F(k, n-k-1)$$

- F 검정통계량의 관측값이 f이면 유의확률은

$$P = P\{F \geq f\}, \quad F \sim F(k, n-k-1)$$

이고 유의수준 α의 기각역은 다음과 같다.

$$f \geq F_\alpha(k, n-k-1)$$

Table: 회귀직선의 유의성검정을 위한 분산분석(ANOVA)표

요인	제곱합	자유도	평균제곱	F값	유의확률
회귀	SSR	k	MSR	f=MSR/MSE	$P\{F \geq f\}$
잔차	SSE	$n-k-1$	MSE		
계	SST	$n-1$			

다중선형회귀모형 예제

Example (예제 8.10)

어떤 플라스틱 제품의 강도(kg/cm^2)가 그 제품을 생산하는 사출 과정 중의 사출온도와 사출시간(sec)에 따라서 어떻게 변화하는가를 알아보기 위하여 실험을 한 결과 다음과 같은 자료를 얻는다.

사출온도(x_1)	195	179	205	204	201	184	210	209
사출시간(x_2)	57	61	60	62	61	54	58	61
강도(y)	81.4	122.2	101.7	145.2	135.9	64.8	92.1	113.8

사출온도와 사출시간을 각각 설명변수 x_1, x_2로 하고 강도를 반응변수 Y로 하는 다중선형회귀모형을 적용할 때, 최소제곱법에 의하여 추정된 회귀직선은 $\hat{y} = -428.8850 - 0.2338x_1 + 9.8295x_2$ 이고, 제곱합이 다음과 같이 주어졌을 때, 이 계산 결과를 이용하여 모회귀함수의 유의성 검정을 유의수준 5%에서 하고, 결정계수 R^2의 값을 구하여라.

$$SST = 5257.8788, \quad SSE = 591.1514$$

sol) 검정하고자 하는 가설은 다음과 같다.

$$H_0 : \beta_1 = \beta_2 = 0 \quad H_1 : At\ least\ one\ \beta_j \neq 0 \quad (j = 1, 2)$$

- 회귀제곱합을 구하면

$$SSR = SST - SSE = 4666.7274$$

이고, 설명변수가 $k = 2$ 이므로 SST, SSE, SSR의 자유도는 각각

$$n - 1 = 7, \quad n - k - 1 = 5, \quad k = 2$$

- 따라서, 회귀평균제곱, 잔차평균제곱과 F 검정통계량을 계산하면

$$MSR = SSR/2 = 2333.3637, \quad MSE = SSE/5 = 118.2303$$

$$f = MSR/MSE = 19.736$$

- F분포표에서 $F_{0.05}(2,5) = 5.79$이므로 기각역은 다음과 같다.

$$f \geq F_{0.05}(2,5) = 5.79$$

- 따라서, 유의수준 5%에서 모회귀함수가 유의하다고 할 수 있고, 통계패키지를 이용하면 P=0.0042임을 알 수 있다.

- 이러한 검정결과를 분산분석표로 정리하면 다음과 같다.

요인	제곱합	자유도	평균제곱	F값	유의확률
회귀	4666.7274	2	2333.3637	f=19.736	$P = 0.0042$
잔차	591.1514	5	118.2303		
계	5257.8788	7			

- 결정계수의 값은 $R^2 = SSR/SST = 0.8876$으로서, 자료 전체의 산포 중에서 약 89%가 회귀직선에 의하여 설명됨을 알 수 있다.

다중선형회귀모형 예제

Example

광고비와 설비투자가 매출액에 어떤 영향을 미치는가를 알아보고자 한다. 매출액은 반응변수 Y이고, 광고비와 설비투자는 각각 x_1, x_2이다.

광고비(x_1)	5	6	7	8	9	11	12	13	14	15
설비투자(x_2)	2	1.9	4	5.6	6.1	6.2	7	7.2	8	9
매출액(y)	16	19	18	20	24	26	30	32	31	34

최소제곱법에 의한 회귀분석 결과의 분산분석표를 이용하여 모회귀함수의 유의성검정을 유의수준 1%에서 하고, R^2의 값을 구하여라.

```
Response: y
          Df Sum Sq Mean Sq  F value    Pr(>F)
x1         1 370.95  370.95 232.5153 1.256e-06 ***
x2         1   1.89    1.89   1.1828    0.3128
Residuals  7  11.17    1.60
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

sol) 검정하고자 하는 가설은 다음과 같다.

$$H_0 : \beta_1 = \beta_2 = 0 \quad H_1 : At\ least\ one\ \beta_j \neq 0 \ \ (j = 1, 2)$$

- 회귀제곱합 SSR은 370.95+1.89=372.84이고, MSR은 186.42이다.
- 유의수준 $\alpha = 0.01$하에서 검정통계량은
  $f = 186.42/1.6 = 116.51 > F_{0.01}(2, 7)$ 이므로 귀무가설이 기각된다.
- 즉, 총체적 검정인 F검정 결과 광고비 또는 설비투자비 중 적어도 하나는 매출액에 선형적으로 유의적인 의미를 갖는 것으로 해석된다.
- 결정계수의 값은 $R^2 = SSR/SST = 372.84/384.01 = 0.9709$로서 광고비 및 설비투자비가 매출액을 선형적으로 많은 부분, 즉 전체 총 변동량의 약 97%정도를 설명하고 있다.

개별 계수 β_j에 대한 추론

- 회귀모형의 F검정결과 귀무가설 $H_0 : \beta_1 = \cdots = \beta_k = 0$이 기각되면, 최소한 하나 이상의 β_j가 0이 아닌 것이므로 개별적 변수에 대한 t유의성 검정을 실시한다.
- 귀무가설 및 대립가설

$$H_0 : \beta_j = 0 \quad H_1 : \beta_j \neq 0$$

- 표본분포

$$\frac{\hat{\beta}_j - \beta_j}{s(\hat{\beta}_j)} \sim t(n-k-1)$$

  여기서 $s(\hat{\beta}_j)$는 $\hat{\beta}_j$의 추정된 분산의 제곱근인 표준오차
- β_j에 관한 $100(1-\alpha)\%$ 신뢰구간:

$$\hat{\beta}_j \pm t_{\alpha/2}(n-k-1) \times s(\hat{\beta}_j)$$

개별 계수 β_j에 대한 유의성 t검정

- β_j에 대한 유의성 검정:
  귀무가설 $H_0 : \beta_j = 0$ 에 관한 검정통계량은 다음과 같다.

$$T = \frac{\hat{\beta}_j - 0}{s(\hat{\beta}_j)}$$

- P값 방법: 검정통계량의 관측값을 t라고 하면 유의확률은 다음과 같고, 만약 $P \leq \alpha$이면, H_0를 기각한다.

$$P = P(|T| \geq |t|)$$

- 기각역 방법: 유의수준 α에서 다음을 만족하면 H_0를 기각한다.

$$|t| \geq t_{\alpha/2}(n - k - 1)$$

유의성 t검정 예제

Example

아래 표는 광고비와 설비투자가 매출액에 미치는 영향을 조사한 자료를 이용하여 추정한 회귀모형의 결과이다.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.0985     1.3618   4.478 0.002872 **
x1            2.2567     0.4048   5.575 0.000838 ***
x2           -0.6430     0.5913  -1.088 0.312805
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.263 on 7 degrees of freedom
Multiple R-squared:  0.9709,    Adjusted R-squared:  0.9626
F-statistic: 116.8 on 2 and 7 DF,  p-value: 4.195e-06
```

광고비와 설비투자의 회귀계수들이 0과 유의적으로 다른지 유의수준 5%하에서 개별 검정을 하고, 각 계수의 95% 신뢰구간을 구하여라.

sol) 각 회귀 계수에 대한 가설은 다음과 같다.

$$H_0 : \beta_j = 0 \quad H_1 : \beta_j \neq 0$$

- 양측검정이고, $(n-k-1) = 7$이므로 t분포표로부터 $t_{0.025}(7) = 2.365$임을 알 수 있다.
- β_1에 관한 95% 신뢰구간:

$$\hat{\beta}_1 \pm t_{0.025}(7) \times s(\hat{\beta}_j) = 2.2567 \pm 2.365 \times 0.4048$$

- β_2에 관한 95% 신뢰구간:

$$\hat{\beta}_2 \pm t_{0.025}(7) \times s(\hat{\beta}_j) = -0.643 \pm 2.365 \times 0.5913$$

- β_1에 대한 검정: 검정통계량은 다음과 같고, 기각역에 속하여 유의수준 5%하에서 귀무가설 $H_0 : \beta_1 = 0$을 기각한다.

$$t = \frac{\hat{\beta}_1}{s(\hat{\beta}_1)} = \frac{2.2567}{0.4048} = 5.575 \geq t_{0.025}(7) = 2.3652$$

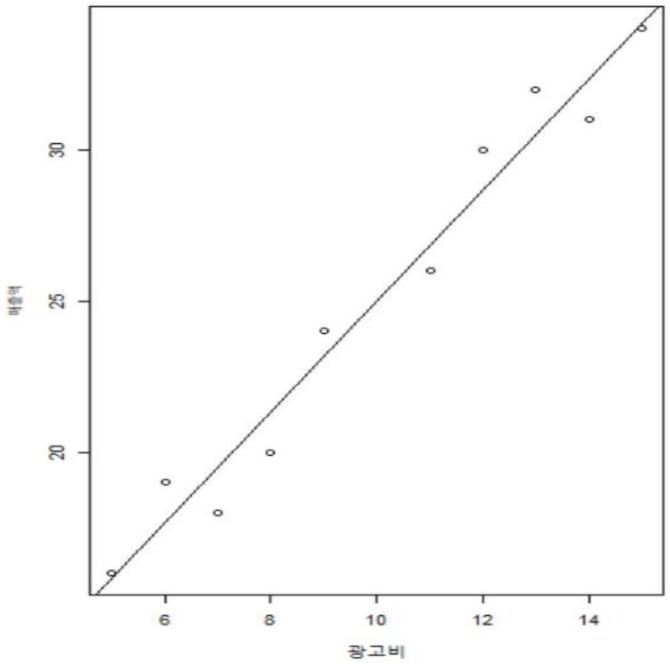
P값도 $P < \alpha = 0.05$로 귀무가설을 기각한다.

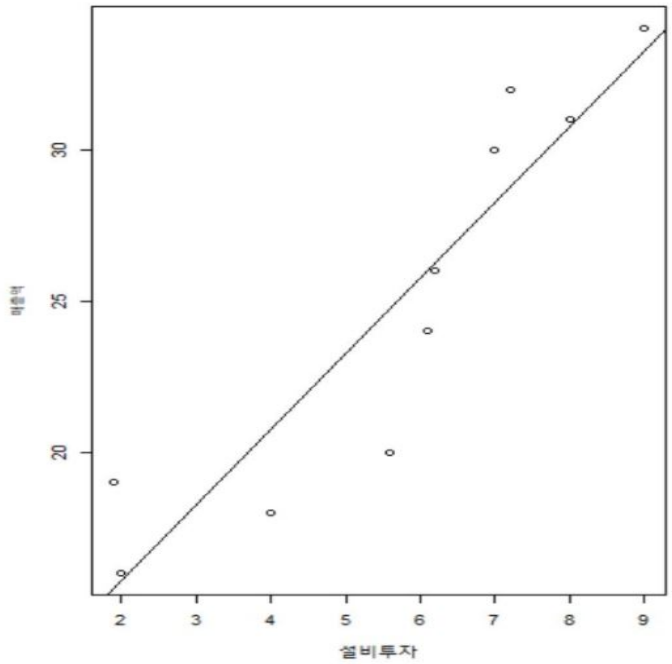
- β_2에 대한 검정: 검정통계량은 다음과 같고, 기각역에 속하지 않으므로 유의수준 5%하에서 귀무가설 $H_0 : \beta_2 = 0$을 기각할 수 없다.

$$t = \frac{\hat{\beta}_2}{s(\hat{\beta}_2)} = \frac{-0.643}{0.5913} = -1.088 \geq -t_{0.025}(7) = -2.365$$

P값도 $P > \alpha = 0.05$로 귀무가설을 기각할 수 없다. 즉, 설비투자비는 매출액에 통계적으로 비유의적인 관계로 판단된다.

다중선형회귀분석시 주의점

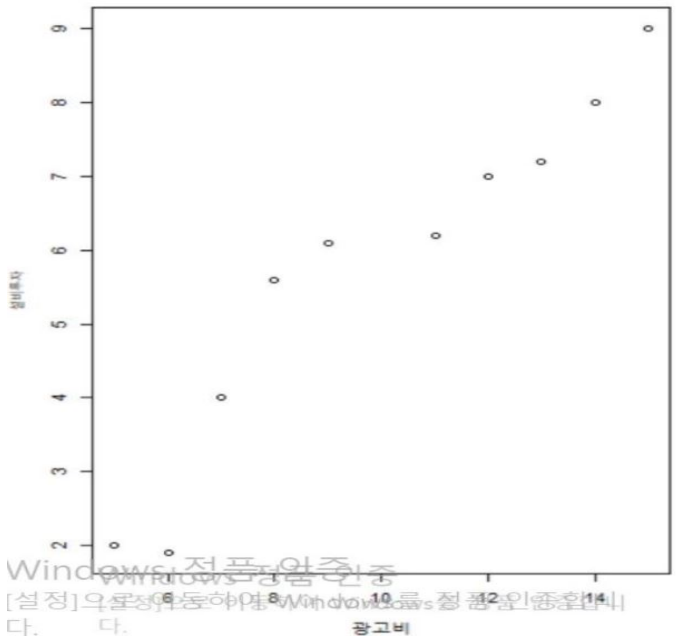




- 광고비와 매출액의 단순선형회귀모형에서 광고비는 F검정에서 통계적으로 유의적이다. ($\hat{y} = 6.63 + 1.83x_1$, $R^2 = 0.966$)
- 설비투자와 매출액도 F검정에서도 설비투자가 매우 유의적인 변수로 설명이 된다. ($\hat{y} = 10.72 + 2.50x_2$)

다중선형회귀분석시 주의점

Figure: 광고비와 설비투자의 산점도



- 광고비와 설비투자의 선형관계를 나타내는 계수인 상관계수가 0.95로 높음에 주의하자.
- 광고비와 설비투자의 높은 선형관계로 인하여 매출액을 설명하는 변수로서의 설비투자의 역할은 광고비가 전부 가지고 가서 설비투자의 역할은 더 이상 없게 된 것이다.

다중공선성(multicollinearity)

- 다중공선성: 설명변수 중 하나와 다른 설명변수들의 선형결합(linear combination)이 매우 높은 상관관계에 있을 때

  $\Longrightarrow$ 설명변수들 사이의 선형독립성(linear independence)이 성립하지 못한 상태

  예) 설명변수에 키, 앉은 키, 다리 길이가 다 있다고 가정하자. 그런데 키는 앉은 키와 다리 길이를 더한 값과 거의 일치한다. 이 경우, 다중공선성 문제가 발생한다.

- 다중공선성이 존재하면 설명변수들의 기울기를 정확히 계산하지 못하거나 아예 계산이 되지 않는다.($Var(\hat{\beta}_j) \uparrow$)

다중공선성(multicollinearity)

- 다중공선성의 진단

  1. 개별 설명변수끼리의 산점도에서 거의 직선에 가까운 선형관계를 보이는 경우
  2. 설명변수 하나를 뺄 때와 추가할 때 다른 설명변수들의 기울기 추정값들이 많이 변동하는 경우
  3. 분산분석표에 의해서 유의적으로 기울기들이 모두 0이 아니지만 문제가 되는 설명변수의 기울기는 유의적이지 않은 경우
  4. 유의성 F 검정에서는 회귀식이 유의하지만, 개별 유의성 t 검정에서는 그렇지 않은 경우 의심해 볼 수 있다.
  5. 분산확대인수(variance inflation factor)를 계산

- 처방 : 설명변수들 끼리 선형 독립성을 유지해야 한다. 이를 위해, 문제가 되는 설명변수를 제거한다.

회귀진단(regression diagnostics)

- 단순선형회귀모형에서의 진단 및 처방과 비슷
- 회귀진단 항목:

  (a) 모형의 선형성(linearity)

  (b) 오차의 등분산성(homoscedasticity), 독립성(independence), 정규성(normality)

  (c) 이상값(outlier)의 존재

  (d) 영향력이 큰 관찰치(influential observation)의 존재

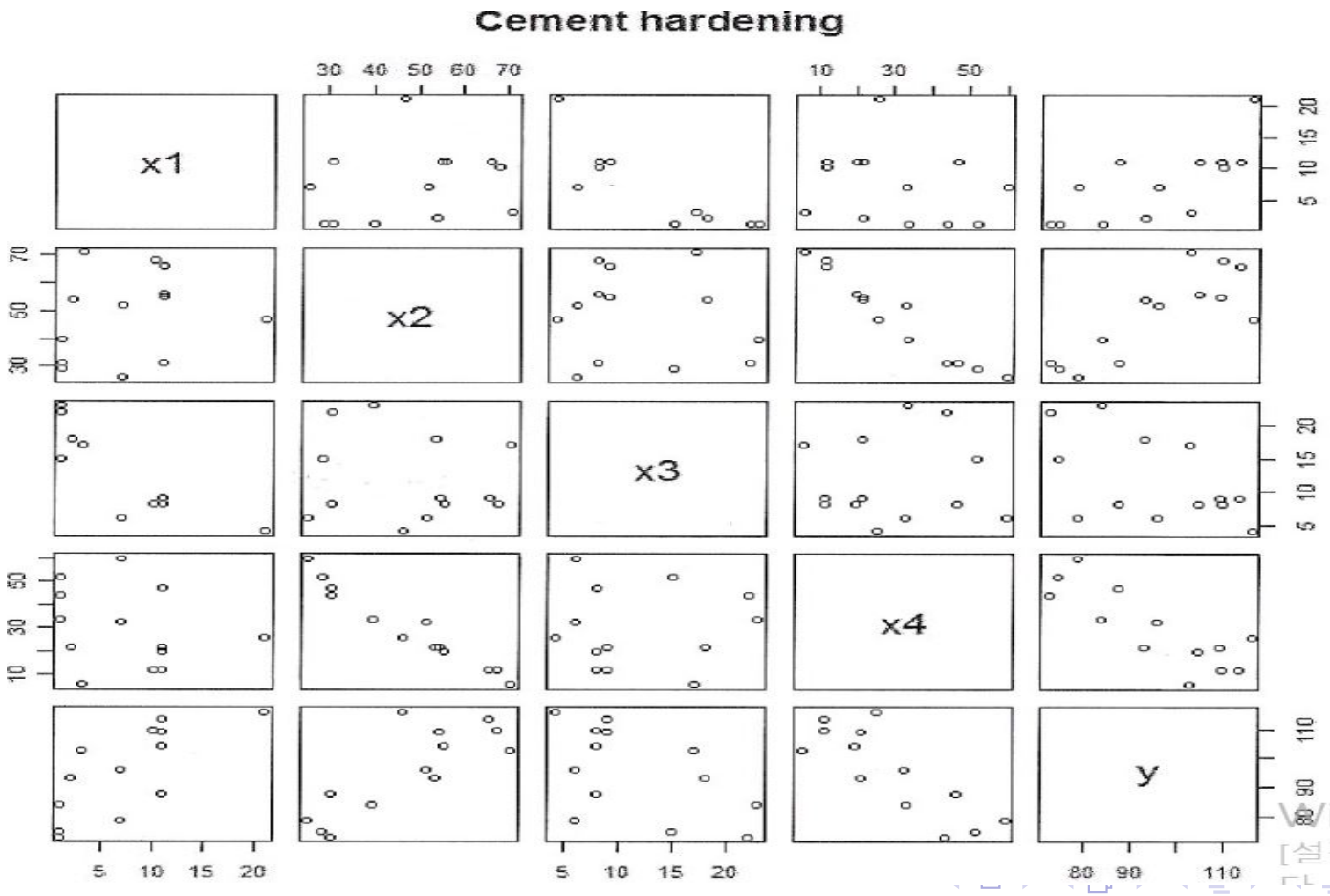
  (e) 다중공선성(multicollinearity)

- 설명변수들이 여러개이므로 산점도행렬(scatter plot matrix)을 이용할 수 있다.

산점도행렬(scatter plot matrix) 예

- 선형모형 적합성, 이상치나 영향력이 큰 관찰치 존재 여부는 단순 선형회귀분석과 같이 산점도로 확인 할 수 있다.

Figure: Y와 x_1, x_2, x_3, x_4의 산점도행렬



잔차분석(analysis of residual)

- 오차에 대한 가정사항을 진단하려면 잔차도를 그린다.

- 잔차그림은 단순회귀분석에서와 같이 (semi)스튜던트화 잔차들을 이용한다.

$$\hat{e}_{st,i} = \frac{e_i}{\widehat{sd}(e_i)} = \frac{(y_i - \hat{y}_i)}{\widehat{sd}(Y_i - \hat{Y}_i)} \quad (i = 1, \cdots, n)$$

- 그런데 다중회귀에서는 설명변수가 여러 개이므로 설명변수의 값을 가로축으로 할 수 없다. 따라서 추측값 $\hat{y}$을 가로축으로, 잔차를 세로축으로 하여 잔차도를 그린다.

$$(\hat{y}_1, \hat{e}_{st,1}), (\hat{y}_2, \hat{e}_{st,2}), \cdots, (\hat{y}_n, \hat{e}_{st,n})$$

잔차분석

- 그림으로부터 다중선형회귀모형의 적용이 타당함을 알 수 있다.

Figure: 예제 8.10 자료에 대한 잔차도