제4장. 다중선형회귀분석

4.1 다중선형회귀모형 - 설명변수가 두 개인 경우

- 설명변수 X_1, X_2와 일변량 반응변수 Y 간에 선형회귀모형은 다음과 같다.

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

β_0는 절편, β_1은 X_1과 관련된 회귀계수, β_2는 X_2와 관련된 회귀계수이며

오차항 $\epsilon \sim iid\ N(0, \sigma^2)$을 가정한다.

- 벡터와 행렬로 표현하면

$$Y = X\beta + \epsilon$$

여기서 $Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$, $x_i = \begin{pmatrix} 1 \\ x_{i1} \\ x_{i2} \end{pmatrix}$, $X = \begin{pmatrix} x_1' \\ \vdots \\ x_n' \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{12} \\ \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} \end{pmatrix}$, $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$, $\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$ 으로

Y와 ϵ는 $n \times 1$ 벡터, x_i와 β는 3×1 벡터이고 X는 $n \times 3$ 행렬이다.

- 회귀계수의 추정량을 구하기 위해 다음의 오차제곱합을

$$S = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2})^2$$

최소화하도록 $\widehat{\beta_0}, \widehat{\beta_1}, \widehat{\beta_2}$을 구해보자. 각 모수에 대해 편미분($\frac{\partial S}{\partial \beta_0}, \frac{\partial S}{\partial \beta_1}, \frac{\partial S}{\partial \beta_2}$)하여 0으로 놓은 후에 $\beta_0, \beta_1, \beta_2$ 에 대한 연립방정식을 풀면 된다.

- *설명변수의 수가 증가하면 동시에 풀어야 하는 연립방정식의 개수도 증가한다.*
- *그러나, 행렬을 이용하면 표현이 간결하고 이해하기도 쉽다.*

4.2 다중선형회귀모형 - 설명변수가 여러 개인 경우

- k개의 설명변수 $X_1, X_2, \cdots, X_k$와 일변량 반응변수 Y 간에 선형회귀모형은

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon$$

여기서 β_j, $j = 1, 2, \cdots, k$은 회귀계수이며, 오차항 $\epsilon \sim iid\ N(0, \sigma^2)$이다.
즉, 오차항 ϵ이 정규분포를 따르면 Y도 정규분포를 따르게 된다.

그러므로 주어진 X에 대하여 Y의 기댓값과 분산을 구하면 다음과 같다.

1) $E(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$, $i = 1, 2, \cdots, n$

2) $Var(Y_i) = \sigma^2$, $i = 1, 2, \cdots, n$

3) $Cov(Y_i, Y_l) = 0$, $i \neq l$

- 다중선형회귀모형을 행렬과 벡터로 다음과 같이 나타낼 수 있다.

$$Y = X\beta + \epsilon$$

또한,

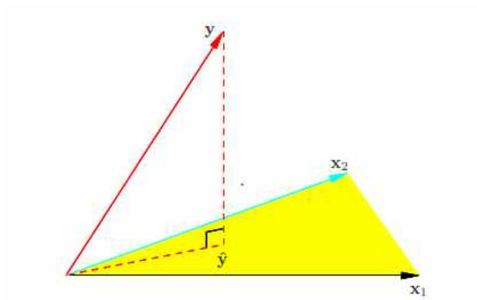
$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & & & \vdots & \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_0 \\ \epsilon_1 \\ \vdots \\ \epsilon_k \end{pmatrix}$$

Y와 ϵ는 $n \times 1$ 벡터, X는 $n \times (k+1)$ 행렬과 β는 $(k+1) \times 1$ 벡터이다.

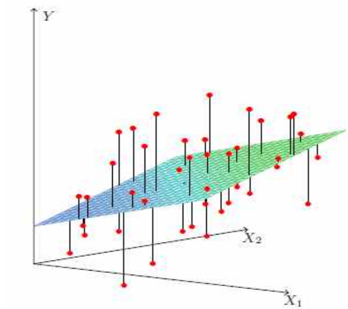
* <그림 4.1>은 반응관측값 Y를 설명변수 X_1과 X_2가 만드는 공간으로 투영시켜 반응 추정값을 구하는 기하학적인 표현을 보여준다.

<그림 4.2>는 선형최소제곱법을 이용한 회귀직선 적합을 공간에서 표현한 그림이다.

<그림 4.1>



<그림 4.2>



(그림출처 : Hastie, Tibshirani & Freidman, 2001)

4.3 모수추정

1. **최소제곱법을 이용한 모수 추정** : 가장 널리 이용되는 방법

회귀계수 β의 추정치 $\hat{\beta}$를 구하기 위해 오차제곱합($S(\beta)$)

$$\begin{aligned} S(\beta) &= \sum_{i=1}^{n} \epsilon_i^2 \\ &= \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_k X_{ik})^2 \\ &= (Y - X\beta)'(Y - X\beta) \end{aligned}$$

을 최소화하는 $\hat{\beta}$를 구하면 [정리4.1]과 같다.

[정리4.1] X가 완전계수(full rank) $k+1 \le n$일 때, β의 최소제곱추정량은

$$\hat{\beta} = (X'X)^{-1} X' Y$$

이다.

$H = X(X'X)^{-1}X'$ 라고 하면,

$$\widehat{Y} = X\hat{\beta} = HY$$

$$e = Y - \widehat{Y} = [I - X(X'X)^{-1}X']Y = (I - H)Y$$ 이며,

$X'e = 0$ 과 $\widehat{Y}'e = 0$을 만족한다.

또한 잔자체곱합은

$$\begin{aligned} S(\hat{\beta}) &= \sum_{i=1}^{n}(Y_i - \widehat{\beta_0} - \widehat{\beta_1}X_{i1} - \cdots - \widehat{\beta_k}X_{ik})^2 \\ &= e'e \\ &= Y'[I - X(X'X)^{-1}X']Y \\ &= Y'Y - Y'X\hat{\beta} \end{aligned}$$

이다.

*** 단순선형회귀모형에 대해 추정량을 이용해 회귀계수벡터를 구해보자.**

$$Y = X\beta + \epsilon, \text{ 즉 } \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} \\ 1 & X_{21} \\ \vdots & \vdots \\ 1 & X_{n1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} ,$$

Y와 ϵ는 $n \times 1$ 벡터, X는 $n \times 2$ 행렬이고 β는 2×1 벡터이다.

$\hat{\beta} = (X'X)^{-1} X'Y$을 이용하여

$$(X'X) = \begin{pmatrix} n & \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} X_i & \sum_{i=1}^{n} X_i^2 \end{pmatrix}, \qquad X'Y = \begin{pmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} X_i Y_i \end{pmatrix}$$

$$(X'X)^{-1} = \frac{1}{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2} \begin{pmatrix} \sum_{i=1}^{n} X_i^2 & -\sum_{i=1}^{n} X_i \\ -\sum_{i=1}^{n} X_i & n \end{pmatrix}$$

$$= \frac{1}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} X_i^2 & -\overline{X} \\ -\overline{X} & 1 \end{pmatrix}$$

구할 수 있다.

따라서 β의 최소제곱추정량은 아래와 같다.

$$\hat{\beta} = \begin{pmatrix} \widehat{\beta_0} \\ \widehat{\beta_1} \end{pmatrix} = (X'X)^{-1}X'Y = \frac{1}{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2} \begin{pmatrix} \sum_{i=1}^{n} X_i^2 & -\sum_{i=1}^{n} X_i \\ -\sum_{i=1}^{n} X_i & n \end{pmatrix} \begin{pmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} X_i Y_i \end{pmatrix}$$

$$= \begin{pmatrix} \overline{Y} - \widehat{\beta_1}\overline{X} \\ \frac{\sum X_i Y_i - n\overline{X}\,\overline{Y}}{\sum (X_i - \overline{X})^2} \end{pmatrix}$$

2. 우도함수를 이용한 모수 추정

오차항이 정규분포를 따른다는 가정하에 관측값에 대한 결합우도함수를 벡터와 행렬을 이용

$$
\begin{aligned}
& P(Y|\beta,\sigma^2,X) \\
&= (2\pi\sigma^2)^{-\frac{n}{2}}\exp\left\{-\frac{(Y-X\beta)'(Y-X\beta)}{2\sigma^2}\right\} \\
&= (2\pi\sigma^2)^{-\frac{n}{2}}\exp\left\{-\frac{1}{2\sigma^2}\left\{(\beta-\hat{\beta})'(X'X)(\beta-\hat{\beta})+Y'[I_n-X(X'X)^{-1}X']Y\right\}\right\}
\end{aligned}
$$

(참고) $(Y-XB)'(Y-XB) = Y'Y - Y'X\beta - X'\beta'Y + \beta'X'X\beta$

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

$$= (\beta - \hat{\beta})' (X'X)(\beta - \hat{\beta}) + Y' [I_n - X(X'X)^{-1} X']$$

여기서 $\hat{\beta} = (X'X)^{-1} X' Y$.

- 우도함수를 최대로 하는 추정량, 즉 지수함수를 최소로 하는 최대우도 추정량은 $\hat{\beta} = (X'X)^{-1} X' Y$이다.

또한, 로그 우도함수 LL을 β에 대해 미분하여 ($\frac{\partial LL}{\partial \beta} = 0$) 구한 최대우도추정량은 $\hat{\beta} = (X'X)^{-1} X' Y$이다.

- 로그 우도함수 LL을 σ^2에 대해 미분하여($\frac{\partial LL}{\partial \sigma^2} = 0$) 구한 최대우도추정량은

$$\widehat{\sigma^2_{ML}} = \frac{(Y - X\hat{\beta})'(Y - X\hat{\beta})}{n} = \frac{1}{n} [Y'Y - Y'X\hat{\beta} - \hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}$$

$$= \frac{1}{n} \left\{ Y' [I_n - X(X'X)^{-1} X'] Y \right\}$$

여기서 I_n은 $n \times n$ 항등행렬이다.

3. 추정량의 성질

[정리4.2] 가우스-마코프 정리(Gauss-Markov Theorem)

- 선형회귀모형에서 오차항의 기댓값이 0이고 서로 독립이며 등분산인 경우, 회귀계수에 대한 최적선형불편추정량(best linear unbiased estimator : BLUE)은 최소제곱추정량이다.

(증명)

● 회귀계수의 최소제곱추정량 $\hat{\beta}=(X'X)^{-1}X'Y$는 다음의 특성을 가짐.

1) $E(\hat{\beta})=\beta$

$Cov(\hat{\beta})=\sigma^2(X'X)^{-1}$

2) 잔차벡터 $e=(e_1,e_2,\cdots,e_n)'$에 대해

$E(e)=0,\quad Cov(e)=\sigma^2(X'X)^{-1}$

3) 분산추정량

$$s^2=\frac{\hat{\epsilon}'\hat{\epsilon}}{n-(k+1)}=\frac{Y'\left[I-X(X'X)^{-1}X'\right]Y}{n-k-1}=\frac{Y'[I-H]Y}{n-k-1}$$

에 대해 $E(s^2)=\sigma^2$ 을 만족하므로 s^2은 σ^2의 불편추정량이다.

4.4 회귀모형추정

반응변수가 정규분포 $N(X\beta, \sigma^2 I)$를 따른다고 가정할 경우,

1. 회귀계수에 대한 검정

다음과 같은 가설에

$$H_{0j} : \beta_j = 0 \qquad H_{1j} : \beta_j \neq 0, \qquad j = 1, 2, \cdots, k$$

대해 검정통계량은

$$t_j = \frac{\hat{\beta}_j}{\widehat{Var}(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{s\sqrt{q_{jj}}} \quad \sim \; t(n-k-1)$$

여기서 q_{jj}는 $(X'X)^{-1}$의 j번째 대각선상에 놓인 값이다.

유의수준 α에서의 양측검정법은

$$|t_j| \geq t_{n-k-1}(\alpha/2) \text{이면 } H_{0j} \text{를 기각한다.}$$

H_{0j}을 기각하는 경우 $\beta_j = 0$이라고 할 수 없으므로 추정된 회귀계수가 회귀모형에 기여함

2. 회귀계수에 대한 신뢰구간

각 회귀계수에 대한 신뢰구간은 t-분포를 이용하여 구할 수 있다.

β_j에 대한 $100(1-\alpha)\%$ 신뢰구간은

$$\widehat{\beta}_j \pm t_{n-k-1}(\alpha/2)\sqrt{\widehat{Var}(\widehat{\beta}_j)}$$

으로 구한다.

[정리4.3] $Y = X\beta + \epsilon$, $rank(X) = k+1$이고, $\epsilon \sim N_n(0, \sigma^2 I)$ 일 때,

회귀계수 β의 최대우도추정량과 최소제곱법추정량은 일치한다. 또한

$\hat{\beta}$의 분포는

$$\hat{\beta} = (X'X)^{-1}X'Y \sim N_{k+1}(\beta, \sigma^2(X'X)^{-1})$$

로 $(k+1)$ 변량 정규분포를 따른다.

잔차에 대해서는 $e = Y - X\hat{\beta}$일 때,

$$n\widehat{\sigma^2} = e'e \sim \sigma^2 \chi^2_{n-k-1}$$

여기서 $\widehat{\sigma^2}$는 σ^2의 최대우도추정량이다.

[정리4.4] $Y = X\beta + \epsilon$, $rank(X) = k+1$이고, $\epsilon \sim N_n(0, \sigma^2 I)$일 때,

회귀계수 β에 대한 $100(1-\alpha)\%$ 신뢰영역은

$$(\beta - \hat{\beta})'(X'X)(\beta - \hat{\beta}) \le (k+1)s^2 F_{k+1, n-k-1}(\alpha)$$

여기서 $F_{k+1, n-k-1}(\alpha)$는 F-분포의 오른쪽 꼬리부분의 확률이고, $s^2 = MSE$이다.

β_j에 대한 $100(1-\alpha)\%$ 동시신뢰구간은

$$\beta_j \pm \sqrt{\widehat{Var}(\widehat{\beta}_j)} \sqrt{(k+1)F_{k+1, n-k-1}(\alpha)}, \qquad j = 0, 1, \cdots, k$$

여기서 $\sqrt{\widehat{Var}(\widehat{\beta}_j)}$는 $s^2(X'X)^{-1}$의 대각선상에서의 j번째 원소

4.5 모형에 대한 적합도 검정

1. 결정계수 : 회귀모형의 적합도에 대한 측도

- 전체변동 = 회귀모형에 의한 변동 + 오차에 의한 변동

$$
\begin{aligned}
* \ (Y-\overline{Y})'(Y-\overline{Y}) &= [(\widehat{Y}-\overline{Y})+(Y-\widehat{Y})]'[(\widehat{Y}-\overline{Y})+(Y-\widehat{Y})] \\
&= (\widehat{Y}-\overline{Y})'(\widehat{Y}-\overline{Y})+(Y-\widehat{Y})'(Y-\widehat{Y}) \\
&= \sum_{j=1}^{n}(\widehat{Y}_j-\overline{Y})^2+\sum_{j=1}^{n}e_j^2
\end{aligned}
$$

(TSS(전체제곱합) = SSR(회귀제곱합) + SSE(잔차제곱합))

결정계수 $R^2 = \dfrac{\sum_{j=1}^{n}(\widehat{Y}_j-\overline{Y})^2}{\sum_{j=1}^{n}(Y_j-\overline{Y})^2} = 1-\dfrac{\sum_{j=1}^{n}e_j^2}{\sum_{j=1}^{n}(Y_j-\overline{Y})^2}$

즉, 전체 변동 중 적합된 회귀식에 의해 설명되는 비율

그러나, 설명변수 개수가 증가할수록 R^2은 무조건 증가한다. 이런 점을 보완한 수정된 결정계수(adjusted R^2)

$$R_{adj}^2 = 1-(1-R^2)(\frac{n-1}{n-k-1}) = 1-\frac{SSE/(n-k-1)}{TSS/(n-1)}$$

여기서 n은 관측개체수, k는 설명변수의 개수이다.

- 모형의 유의성 검정

$H_0 : \beta = 0 \qquad H_1 : \beta \neq 0$

<분산분석표>

	SS	DF	MS	F	유의확률
Model	SSR	k	SSR/k	$F_0 = MSR/MSE$	$F_{k,n-k-1}(\alpha)$
Error	SSE	$n-k-1$	$SSE/(n-k-1)$		
Total	SST	$n-1$	$SST/(n-1)$		

분산분석표에서 검정통계량

$$F_0 = \frac{(TSS - SSE)/k}{SSE/(n-k-1)}$$

$F_0 \geq F_{k,n-k-1}(\alpha)$이면 H_0를 기각한다.

2. 회귀계수벡터에 대한 검정

다중회귀모형에서 회귀계수 벡터 $\beta = \begin{pmatrix} \beta_0 \\ \beta_X \end{pmatrix}$ 로 표현 할 수 있다. β_X에 대한 유의성 검정이 모형의 유의성검정이 된다.

$$H_0 : \beta_X = 0$$

에 대한 F-검정통계량

$$F = \frac{SSR/k}{SSE/(n-k-1)} \overset{H_0}{\sim} F(k, n-k-1)$$

$F \geq F_{k,n-k-1}(\alpha)$이면 H_0를 기각한다.

4.6 신뢰구간과 예측구간

1. 반응평균과 반응평균에 대한 신뢰구간

설명변수 $X_* = (X_{*1}, X_{*2}, \cdots, X_{*k})'$에서의

종속변수 $Y_* = \beta_0 + \beta_1 X_{*1} + \beta_2 X_{*2} + \cdots + \beta_k X_{*k}$의 반응평균은

$$\begin{aligned} y_* &= E[Y_*] \\ &= \beta_0 + \beta_1 X_{*1} + \beta_2 X_{*2} + \cdots + \beta_k X_{*k} \\ &= x_*' \beta \end{aligned}$$

이고, 여기서 $x_* = (1, X_{*1}, X_{*2}, \cdots, X_{*k})'$이다.

y_*의 추정값은 $\hat{y}_* = x_*' \hat{\beta}$,

Y_*의 불편추정량은 $E[Y_*] = \widehat{y_*}$, $Var[\widehat{y_*}] = \sigma^2 x_*' (X'X)^{-1} x_*$

$$\therefore \widehat{y_*} = x_*' \beta \sim N(y_*, \sigma^2 x_*' (X'X)^{-1} x_*)$$

X_*에서 반응평균에 대한 추정값 $\hat{y}_*$에 대한 $100(1-\alpha)\%$ 신뢰구간은

$$\widehat{y_*} \pm t_{n-k-1}(\alpha/2)se[\widehat{y_*}]$$

여기서 $se[\widehat{y_*}] = s\sqrt{x_*'(X'X)^{-1}x_*}$, $s = \sqrt{MSE_0}$ 이다.

2. 예측과 예측구간

반응값이 관측되지 않은 새로운 설명변수 값 $X_* = (X_{*1}, X_{*2}, \cdots, X_{*k})'$에서 반응변수에 대한 예측모형은

$$\hat{y}_* = x_*'\hat{\beta}$$

여기서 $x_* = (1, X_{*1}, X_{*2}, \cdots, X_{*k})'$, $\hat{\beta} = (X'X)^{-1}X'Y$이다.

$\widehat{Y_*}$는 $\hat{y}_*$의 예측값이라 할 때, 예측값에 대한 신뢰구간을 예측구간이라 한다.

$$\text{확률변수 } \widehat{Y_*} = x_*'\beta + \epsilon$$

즉, $\widehat{Y_*} \sim N(x_*'\beta, \sigma^2(1 + x_*'(X'X)^{-1}x_*))$

$\widehat{Y}_*$에 대한 $100(1-\alpha)\%$ 예측구간은

$$\widehat{Y}_* \pm t_{n-k-1}(\alpha/2)se[\widehat{Y}_*]$$

여기서 $se[\widehat{Y}_*] = s\sqrt{(1+x_*'(X'X)^{-1}x_*)}$, $s = \sqrt{MSE_0}$

4.7 잔차분석

- **잔차** = 관측값과 추정값의 차이
  = 회귀모형에 의하여 설명되지 않은 변동의 크기를 나타내는 오차에 대한 추정값

- **선형회귀모형에서 사용되는 가정들**

선형성: 입력변수와 출력변수간의 관계가 선형적

등분산성: 오차의 분산이 입력변수와 무관하게 일정

독립성: 오차들이 서로 독립

정규성: 오차의 분포가 정규분포

- **선형모형을 자료에 적합하기 전에 위의 3가지 가정(오차항에 대한 가정)이 만족되는지를 확인하여야 한다.**

→ 잔차분석을 통하여 위 가정들을 만족하는지 체크

4.8 R 활용 다중회귀분석

[예제4.1] R 시스템에 내장되어 있는 환경 데이터 airquality에 대한 다중회귀분석 실행
airquality data set은 New Yo가 도시의 153일 동안 오존, 일조량, 기온과 풍속 데이터이다.

1) 데이터 파악

2) 데이터 산점도

3) 오존을 반응변수로 한 다중회귀모형
log(오존)을 반응변수로 한 다중회귀모형

4) 잔차정규성검정

5) 잔차독립성검정

6) 오존 데이터의 회귀계수에 대한 신뢰영역

7) 새로운 데이터에 대한 추정

4.9 다항회귀모형

- 설명변수와 반응변수 사이의 곡선 관계를 다항식을 이용해 표현

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \epsilon$$

[예제 4.2] 8마리의 실험쥐에 대해 A약의 용량을 달리하여 투여한지 2주 후 몸무게 증가량을 측정하여 <표4.3> 데이터를 얻었다. 이와 같은 데이터에 대해 이차항이 포함된 다항회귀모형을 적합하려 한다.

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon,\ \epsilon \sim iid\, N(0, \sigma^2)$$

투여량	1	2	3	4	5	6	7	8
몸무게 증가량	1.0	1.2	1.8	2.0	3.8	4.3	6.5	9.0

[예제 4.2] 철(Fe) 성분 함유에 따라 부식 정도를 측정한 데이터에 대해 이차항이 포함된 다항회귀모형

$$Y = \beta_0 + \beta_1 X + \epsilon\,,\ \epsilon \sim iid\, N(0, \sigma^2)$$

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$$

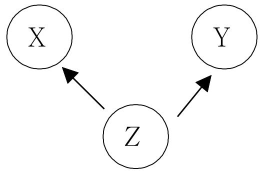
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \beta_5 X^5 + \epsilon$$

을 적합해보고 비교하고자 한다.

4.10 설명변수와 반응변수의 상관성과 관련한 문제

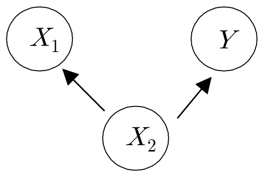
- 회귀분석 시 변수들 간의 관련성 파악

1) 그럴 듯한 관계(spurious relationship) : X와 Y



중간에 매개된 변수 Z가 X와 Y를 공통적으로 관련

2) 다중공선성 관계(collinear relationship) : X_2는 X_1과 Y와 각각 상관있을 경우



3) 끼어드는 관계(intervening relationship)