4강

중회귀모형 (1)

정보통계학과 김성수교수

학습목차

1. 2.1 중회귀모형
2. 2.2 중회귀모형의 추정
3. 2.3 회귀방정식의 신뢰성

1 중회귀모형

중회귀모형의 기본개념

✓ 중회귀모형

종속변수의 변화를 설명하기 위하여 두 개 이상의 독립변수가 사용되는 선형회귀모형을 **중선형회귀**(multiple linear regression model)라 부르며, 간단히 **중회귀모형**(multiple regression model)이라고도 함.

- **독립변수의 수가 k개인 중회귀모형**

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

$\beta_0, \beta_1, \cdots, \beta_k$: 회귀계수

$\varepsilon_i \sim N(0, \sigma^2)$ 이고 서로 독립

$i = 1, 2, \cdots, n$

여기서 $\beta_0, \beta_1, \cdots, \beta_k$ 는 모집단의 회귀계수이고,

ε_i 는 반응변수 Y_i 를 측정할 때 발생하는 오차.

행렬을 이용한 중회귀모형

✓ 중회귀모형에서 독립변수가 2개인 경우

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \qquad i = 1, 2, \cdots, n$$

n개의 오차 $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ 은 서로 독립이고, 각각 $N(0, \sigma^2)$ 의 분포를 따른다고 가정.

✓ 벡터 표현

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \\ &= (1, \ X_{i1}, \ X_{i2}) \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \varepsilon_i \end{aligned}$$

행렬을 이용한 중회귀모형

✓ 중회귀모형의 행렬 표현

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\ , \quad \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma^2)$$

$$\underset{n\times 1}{\boldsymbol{Y}} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \underset{n\times 3}{\boldsymbol{X}} = \begin{pmatrix} 1 & X_{11} & X_{12} \\ 1 & X_{21} & X_{22} \\ \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} \end{pmatrix}, \quad \underset{3\times 1}{\boldsymbol{\beta}} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}, \quad \underset{n\times 1}{\boldsymbol{\varepsilon}} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

여기에서

$$E(\boldsymbol{\varepsilon}) = \boldsymbol{0}$$

$$\begin{aligned} Var(\boldsymbol{\varepsilon}) &= E[(\boldsymbol{\varepsilon} - E(\boldsymbol{\varepsilon}))(\boldsymbol{\varepsilon} - E(\boldsymbol{\varepsilon}))'] \\ &= E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] \\ &= \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix} = \boldsymbol{I}\sigma^2 \end{aligned}$$

행렬을 이용한 중회귀모형

<표본상점의 총판매액 자료>

상점번호	광고료 (단위;100만원)	상점의 크기 (단위: 10 m^2)	총판매액 (단위:1000만원)
1	4.2	4.5	9.3
2	8.5	12.0	18.5
3	9.3	15.0	22.8
4	7.5	8.5	17.7
5	6.3	7.4	14.6
6	12.2	18.5	27.9
7	6.5	5.5	12.5
8	10.4	16.5	25.2
9	5.8	3.7	10.8
10	9.2	13.5	20.5
11	7.2	5.2	14.9
12	8.5	15.0	19.2
13	10.6	14.4	22.5
14	13.9	13.3	28.4
15	12.7	12.5	25.6

회귀모형 $Y = X\beta + \varepsilon$

여기서,

$$Y = \begin{pmatrix} 9.3 \\ 18.5 \\ 22.8 \\ 17.7 \\ 14.6 \\ 27.9 \\ 12.5 \\ 25.2 \\ 10.8 \\ 20.5 \\ 14.9 \\ 19.2 \\ 22.5 \\ 28.4 \\ 25.6 \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 4.2 & 4.5 \\ 1 & 8.5 & 12.0 \\ 1 & 9.3 & 15.0 \\ 1 & 7.5 & 8.5 \\ 1 & 6.3 & 7.4 \\ 1 & 12.2 & 18.5 \\ 1 & 6.5 & 5.5 \\ 1 & 10.4 & 16.5 \\ 1 & 5.8 & 3.7 \\ 1 & 9.2 & 13.5 \\ 1 & 7.2 & 5.2 \\ 1 & 8.5 & 15.0 \\ 1 & 10.6 & 14.4 \\ 1 & 13.9 & 13.3 \\ 1 & 12.7 & 12.5 \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

2 중회귀모형의 추정

회귀계수의 추정

✓ 최소제곱법

- 중회귀모형에서 β의 최소제곱추정량 $\hat{\beta}$ 는 단순회귀와 마찬가지로 다음과 같은 오차제곱함을 최소로 하는 β 를 구하면 됨.

$$S=\sum_{i=1}^{n}\epsilon_i^2=\sum_{i=1}^{n}(Y_i-\beta_0-\beta_1X_{i1}-\beta_2X_{i2}-\cdots-\beta_kX_{ik})^2$$

$$S=\sum_{i=1}^{n}\epsilon_i^2=\epsilon'\epsilon=(Y-X\beta)'(Y-X\beta)$$
$$=Y'Y-Y'X\beta-\beta'X'Y+\beta'X'X\beta$$

$\Rightarrow \beta$ 에 관해 미분 $\quad \frac{\partial S}{\partial \beta}=-2X'Y+2X'X\beta=0$

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

회귀계수의 추정

✓ 참고 : 행렬의 미분법

$$c = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \text{라 할 때, } \quad c'x = c_1x_1 + c_2x_2 + \cdots + c_nx_n$$

1) $$\frac{\partial(c'x)}{\partial x} = \begin{bmatrix} \frac{\partial(c'x)}{\partial x_1} \\ \frac{\partial(c'x)}{\partial x_2} \\ \vdots \\ \frac{\partial(c'x)}{\partial x_n} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix} = c$$, 마찬가지로 $$\frac{\partial(c'x)}{\partial c} = x$$

2) $n \times n$ 행렬 A : 대칭행렬인 경우

$$\frac{\partial(x'Ax)}{\partial x} = 2Ax$$

R 활용 : 행렬 연산

```
> market2 = read.table("c:/data/reg/market-2.txt", header=T)
> head(market2,3)
  ID  X1   X2    Y
1  1 4.2  4.5  9.3
2  2 8.5 12.0 18.5
3  3 9.3 15.0 22.8
> X = market2[,c(2:3)]
> X = cbind(1, X)
> Y = market2[,4]
> X = as.matrix(X)
> Y = as.matrix(Y)
> XTX = t(X) %*% X
> XTX
       1      X1      X2
1   15.0  132.80  165.50
X1 132.8 1280.80 1610.68
X2 165.5 1610.68 2149.49
> XTXI = solve(XTX)
> XTY = t(X) %*% Y
```

```
> beta = XTXI %*% XTY
> beta = round(beta,3)
> beta
   [,1]
1 0.850
X1 1.558
X2 0.427
```

$$\hat{\beta} = (X'X)^{-1}X'Y = \begin{pmatrix} 0.850 \\ 1.558 \\ 0.427 \end{pmatrix}$$

적합된 선형회귀식 :

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$
$$= 0.850 + 1.558 X_1 + 0.427 X_2$$

R 활용 : 행렬 연산

- 참고 : 적합된 선형회귀식을 이용하면 X_1과 X_2에 대하여
  총 판매액의 기대값 E(Y)를 측정할 수 있음.

예) 광고료가 1000만원 (X_1=10) 이고 상점의 크기가 100m^2 (X_2=10) 인
상점의 평균 총판매액의 추정값은

$$\hat{Y} = 0.850 + 1.558 \times 10 + 0.427 \times 10 = 20.7$$

으로 약 207백만원이 되리라고 추정.

잔차의 성질

✓ 잔차

추정된 회귀식의 값 $\hat{Y}_i$ 과 관찰값 Y_i 의 차이를 잔차, $e_i = Y_i - \hat{Y}_i$ 라 함.

추정값과 잔차벡터의 행렬표현 :

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

$$\begin{aligned}e &= Y - \hat{Y} = Y - X\hat{\beta} \\ &= Y - X(X'X)^{-1}X'Y \\ &= \left(I - X(X'X)^{-1}X'\right)Y \\ &= (I - H)Y\end{aligned}$$

여기서, $H = X(X'X)^{-1}X'$ 을 나타내며, 이를 햇행렬(hat matrix)이라 함.

잔차의 성질

햇행렬은 다음을 만족하는 멱등행렬(idempotent matrix)임

$$H^2 = HH = H$$

$$H' = H$$

➢ 잔차의 성질

(1) 잔차의 합은 0 . $\sum e_i = 0$

(2) 잔차의 독립변수에 대한 가중합은 0 . 즉 $\sum_{i=1}^{n} X_{ij} e_i = 0, \ (j = 1, 2, \cdots, k)$

(3) 잔차의 추정값에 대한 가중합도 0 . 즉, $\sum \hat{Y}_i e_i = 0$

(4) 중회귀모형 $Y = X\beta + \varepsilon$ 에서 오차항 ε 는 $N(0, I\sigma^2)$ 의 분포를 하며, ε_i 간에는 서로 상관관계가 없이 서로 독립적이나, 잔차 e_i 간에는 상관관계가 일반적으로 존재함.

잔차의 성질

$e' = (e_1, e_2, \cdots, e_n)$ 의 기댓값 벡터와 분산-공분산행렬

$$E(e) = E[(I - H)Y] = (I - H)E(Y)$$
$$= [I - X(X'X)^{-1}X'](X\beta) = 0$$

$$E(Y) = X\beta$$

$$Var(e) = (I - H)Var(Y)(I - H)'$$
$$= (I - H)(I\sigma^2)(I - H)'$$
$$= (I - H)(I - H)'\sigma^2 = (I - H)\sigma^2$$
$$= [I - X(X'X)^{-1}X']\sigma^2$$

$$Var(Y) = I\sigma^2$$

⇒ $Var(e)$ 는 대각선행렬이 1 이 아니며,

e_i 와 e_j 간에 공분산이 존재.

3 회귀방정식의 신뢰성

회귀방정식의 신뢰성

✓ 추정된 회귀방정식의 신뢰성 측도

(1) 분산분석표에 의한 F-검정

(2) 결정계수(Coefficient of determination)

(3) 잔차평균제곱(residual mean squares)

(4) 추정된 회귀계수들의 분산

(5) 종속변수의 추정량 의 분산

분산분석표에 의한 F-검정

✓ **총제곱합 SST**

$$SST = \sum (Y_i - \overline{Y})^2 = \sum Y_i^2 - n(\overline{Y})^2$$
$$= \boldsymbol{Y}'\boldsymbol{Y} - n(\overline{Y})^2$$

, $n(\overline{Y})^2 = \frac{1}{n}(\sum Y_i)^2 = \boldsymbol{Y}'\boldsymbol{1}\boldsymbol{1}'\boldsymbol{Y}/n = \boldsymbol{Y}'(\frac{\boldsymbol{J}}{n})\boldsymbol{Y}$

, $\boldsymbol{J}$ 행렬은 모든 요소가 1인 $n \times n$ 행렬

$$\Rightarrow \quad SST = \boldsymbol{Y}'\boldsymbol{Y} - n(\overline{Y})^2$$
$$= \boldsymbol{Y}'(\boldsymbol{I} - \frac{\boldsymbol{J}}{n})\boldsymbol{Y}$$

분산분석표에 의한 F-검정

✓ **잔차제곱합 SSE**

잔차벡터 $e=(I-H)Y$

$$SSE=\sum(Y_i-\hat{Y}_i)^2=\sum e_i^2=e'e$$
$$=[(I-H)Y]'[(I-H)Y]$$
$$=Y'(I-H)'(I-H)Y$$
$$=Y'(I-H)Y$$

분산분석표에 의한 F-검정

✓ **회귀제곱합 SSR**

$$
\begin{aligned}
SSR &= \sum (\hat{Y}_i - \bar{Y})^2 \\
&= \sum \hat{Y}_i^2 - n(\bar{Y})^2 \\
&= \hat{\boldsymbol{Y}}' \hat{\boldsymbol{Y}} - n(\bar{Y})^2
\end{aligned}
$$

여기서 $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{H}\boldsymbol{Y}$ 이므로

$$
\begin{aligned}
SSR &= \hat{\boldsymbol{\beta}}' \boldsymbol{X}' \boldsymbol{Y} - n(\bar{Y})^2 \\
&= \boldsymbol{Y}' \boldsymbol{H} \boldsymbol{Y} - \boldsymbol{Y}' \left(\frac{\boldsymbol{J}}{n}\right) \boldsymbol{Y} \\
&= \boldsymbol{Y}' \left(\boldsymbol{H} - \frac{\boldsymbol{J}}{n}\right) \boldsymbol{Y}
\end{aligned}
$$

분산분석표에 의한 F-검정

✓ **변동의 분해**

✓ 분산분석표

〈중회귀의 분산분석표〉

요인	자유도	제곱합	평균제곱	F_0
회귀	k	SSR	$MSR = \frac{SSR}{k}$	$\frac{MSR}{MSE}$
잔차	$n-k-1$	SSE	$MSE = \frac{SSE}{n-k-1}$	
계	$n-1$	SST		

$F_0 = \frac{MSR}{MSE}$: 회귀방정식이 유의한가를 검정하기 위한 검정통계량

귀무가설 $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$

대립가설 H_1 : 최소한 하나의 $\beta_i \neq 0$, $i = 1, 2, \cdots, k$

: 유의수준 α에서 만약 F_0의 값이 $F_0 > F(k, n-k-1; \alpha)$ 이면 귀무가설을 기각하며, 회귀방정식이 유의(significant)하다는 것을 의미.

: R 분석 결과에서는 검정통계량 F_0 의 유의확률 p-값을 이용하여 검정하면 됨.

R 활용 예

```
> market2 = read.table("c:/data/reg/market-2.txt", header=T)
> head(market2, 2)
  ID  X1   X2    Y
1  1 4.2  4.5  9.3
2  2 8.5 12.0 18.5
> market2.lm = lm(Y ~ X1+X2, data=market2)
> 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 ***
---
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 회귀식 : $\hat{Y}=0.85041+1.55811X1+0.4273X2$

결정계수 : 0.9799

F-값=292.5 이고, 유의확률 p-값=6.597×10^{-11} 로서 적합된 중회귀모형이 이 데이터를 설명하는데 유의함.

(이는 귀무가설 $H_0:\beta_1=\beta_2=0$ 이 기각되므로 β_1과 β_2가 동시에 영이 되지는 않을 것이라는 의미임)

R 활용 예 : 분산분석표

```
> anova(market2.lm)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X1         1 485.57  485.57 559.283 1.955e-11 ***
X2         1  22.30   22.30  25.691 0.0002758 ***
Residuals 12  10.42    0.87
---
```

중회귀모형의 분산분석 결과 해석

$$SS(X1) = 485.57$$

$$SS(X2|X1) = 22.30$$

, 여기서 $SS(X2|X1)$는 변수 $X1$이 적합된 후, 변수 $X2$가 추가되었을 때의 추가제곱합을 의미.

회귀제곱합 $SS(X1, X2) = SS(X1) + SS(X2|X1)$

〈분산분석표〉

요인	자유도	제곱합	평균제곱	F_0	Pr(>F)
회귀	2	507.87	253.94	292	6.597e-11
잔차	12	10.42	0.87		
계	14	518.29			

결정계수

✓ 결정계수(Coefficient of determination)

• 회귀모형에 의하여 설명되는 변동 SSR이 총변동 SST에 비하여 어느 정도인가를 나타내 주는 값

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

• 의미 : R^2의 값이 1에 가까운 값을 가지면 추정된 회귀식이 관찰점들을 잘 설명해주고 있으며, R^2의 값이 0에 가까운 작은 값을 갖게 되면 그와 같은 회귀식을 추정하게 되는 의의는 거의 없음.

결정계수

✓ 중상관계수(multiple correlation coefficient)

- 단순회귀모형에서 결정계수 : 두 변수의 상관계수의 제곱과 같음.

- 중회귀모형의 결정계수 : 반응변수 Y_i와 추정값 $\hat{Y}_i$의 상관계수의 제곱과 같음.

- 따라서 결정계수의 제곱근

$$R = \sqrt{R^2}$$

을 중상관계수(multiple correlation coefficient)라고 함.

R 활용 : 중상관계수

```
> market2.lm = lm(Y ~ X1+X2, data=market2)
> summary(market2.lm)
---
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
```

```
> names(market2.lm)
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"
> yhat = market2.lm$fitted
> cor(market2$Y, yhat)
[1] 0.9898983
> cor(market2$Y, yhat)^2
[1] 0.9798986
```

R 활용 : 중상관계수

✓ 수정결정계수(adjusted R-squared)

- 독립변수가 추가하게 되면 결정계수는 항상 증가함. 왜냐하면 총제곱합 SST는 고정된 값이고, 잔차제곱합 SSE는 독립변수가 추가될수록 작아지게 되므로 결정계수는 증가하는 성질을 가지고 있음.
- 따라서 독립변수들을 추가하다 보면 과다한 적합을 할 수 있게 되므로 변수선택과 같은 모형개발이라는 입장에서 볼 때 두 모형을 비교하기 위한 결정계수로서 수정결정계수가 자주 이용됨.

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

- 수정결정계수는 설명력이 떨어지는 독립변수가 추가될 때는 감소하는 성질을 가지고 있으므로 모형선택의 관점에서 유용하게 이용됨.

R 결과

```
> market2.lm = lm(Y ~ X1+X2, data=market2)
> summary(market2.lm)
---
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
```

결정계수 : 0.9799
수정결정계수 : 0.9765

R 결과

✓ 잔차평균제곱(residual mean squares)

- 잔차평균제곱 MSE

$$MSE = \frac{SSE}{n-k-1} = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-k-1}$$

- MSE의 기대값 $E(MSE) = \sigma^2$ 으로 MSE는 σ^2의 불편추정량이 됨.
- MSE의 값이 작으면 작을수록 관찰값 Y_i들이 추정값 $\hat{Y}_i$ 과 차이가 거의 없다는 것을 의미하며, 추정된 회귀방정식을 믿을 수 있게 됨.

R 결과

```
> market2.lm = lm(Y ~ X1+X2, data=market2)
> summary(market2.lm)
---
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
```

잔차평균제곱근 $\sqrt{MSE}$
0.9318

```
> anova(market2.lm)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X1         1 485.57  485.57 559.283 1.955e-11 ***
X2         1  22.30   22.30  25.691 0.0002758 ***
Residuals 12  10.42    0.87
---
```

```
> sqrt(10.42/12)
[1] 0.931844
```

다음시간 안내

5강. 중회귀모형 (2)