제 7 강

다중회귀모형
The Multiple Regression Model

$$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \ldots + \beta_K x_{tK} + \varepsilon_t$$

모수 β_1 은 절편(상수) 항임

β_1 에 대응되는 '변수'는 $x_{t1} = 1$.

대개 설명변수의 숫자는 K–1 ($x_{t1} = 1$을 무시함), 반면에 모수의 숫자는 K. (즉 $\beta_1 \ldots \beta_K$).

일반적 다중회귀모형

Matrix

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{T-1} \\ y_T \end{bmatrix} = \begin{bmatrix} 1 & x_{12} & \cdots & x_{1K} \\ 1 & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{T-12} & \cdots & x_{T-1K} \\ 1 & x_{T2} & \cdots & x_{TK} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{T-1} \\ \varepsilon_T \end{bmatrix}$$

$$\Downarrow$$

$$\underset{(T\times 1)}{y} = \underset{(T\times K)}{X} \cdot \underset{(K\times 1)}{\beta} + \underset{(T\times 1)}{\varepsilon}$$

다중회귀모형에 대한 가정

y_t

1. $y_t = \beta_1 + \beta_2 x_{t2} + \ldots + \beta_K x_{tK} + \varepsilon_t$
2. $E(y_t) = \beta_1 + \beta_2 x_{t2} + \ldots + \beta_K x_{tK}$
3. $var(y_t) = \sigma^2$
4. $cov(y_t, y_s) = 0,\ t \neq s$

5 x_{tk} 는 확률변수가 아니며, 다른 설명변수들의 정확한 선형함수가 아님

6. $y_t \sim N(\beta_1+\beta_2 x_{t2} + \ldots + \beta_K x_{tK}, \sigma^2)$

다중회귀모형에 대한 가정

ε_t

1. $y_t = \beta_1 + \beta_2 x_{t2} + \ldots + \beta_K x_{tK} + \varepsilon_t$
2. $E(\varepsilon_t) = 0$
3. $var(\varepsilon_t) = \sigma^2$
4. $cov(\varepsilon_t, \varepsilon_s) = 0,\ t \neq s$

5. x_{tk} 는 확률변수가 아니며 다른 설명변수들의 정확한 선형함수가 아님

6. $\varepsilon_t \sim N(0, \sigma^2)$

두 개의 설명변수

개별 모수들의 의미

$$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \varepsilon_t$$

$$\frac{\partial E(y_t)}{\partial x_{t2}} = \beta_2, \quad : \beta_2$$ x_{t3} 를 일정하게 놓을 때, x_{t2}의 한 단위 변화가 y의 평균값에 미치는 영향

즉 x_{t2}에서의 한 단위 변화의 y의 평균값에 대한 '직접적(direct)' 혹은 '순(net)' 영향

$$\frac{\partial E(y_t)}{\partial x_{t3}} = \beta_3 \quad : \beta_3$$ x_{t2} 를 일정하게 놓을 때, x_{t3} 에 있어서의 한 단위 변화가 y의 평균값에 미치는 영향

일정하게 놓을 때(Holding constant):

x_{t2} 의 y의 변화에 있어서의 진정한 공헌을 평가하기 위해, x_{t3}의 영향을 통제함을 의미함

상관관계가 있는 설명변수

$$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \varepsilon_t$$

y_t = 산출량 x_{t2} = 자본(기계) x_{t3} = 노동

기계 한대 당 5명의 노동력

기계당 노동력의 수가 변화하지 않는다면 기계 혹은 노동력이 산출량의 변화를 얼마나 설명하는가를 구분하는 것이 불가능하게 됨

두 개의 설명변수

최소제곱추정

$$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \varepsilon_t$$

$$S \equiv S(\beta_1, \beta_2, \beta_3) = \sum_{t=1}^{T} (y_t - \beta_1 - \beta_2 x_{t2} - \beta_3 x_{t3})^2$$

Define: $y_t^* = y_t - \bar{y}$

$$x_{t2}^* = x_{t2} - \bar{x}_2$$

$$x_{t3}^* = x_{t3} - \bar{x}_3$$

두 개의 설명변수

최소제곱추정량

$$b_1 = \bar{y} - b_2\bar{x}_2 - b_3\bar{x}_3$$

$$b_2 = \frac{(\sum y_t^* x_{t2}^*)(\sum x_{t3}^{*2}) - (\sum y_t^* x_{t3}^*)(\sum x_{t2}^* x_{t3}^*)}{(\sum x_{t2}^{*2})(\sum x_{t3}^{*2}) - (\sum x_{t2}^* x_{t3}^*)^2}$$

$$b_3 = \frac{(\sum y_t^* x_{t3}^*)(\sum x_{t2}^{*2}) - (\sum y_t^* x_{t2}^*)(\sum x_{t3}^* x_{t2}^*)}{(\sum x_{t2}^{*2})(\sum x_{t3}^{*2}) - (\sum x_{t2}^* x_{t3}^*)^2}$$

Least squares estimation of β_2 now depends upon <u>both</u> x_{t2} and x_{t3} .

최소제곱추정량의 성질

1. 회귀선 (면) $(\bar{y},\ \bar{x}_2,\ \bar{x}_3)$을 통과함

i.e., $b_1 = \bar{y} - b_2\bar{x}_2 - b_3\bar{x}_3$

==> $\bar{y} = b_1 + b_2\bar{x}_2 + b_3\bar{x}_3$

2. 불편추정량: $E(b_i) = \beta_i$
3. 일치추정량: $b_i \rightarrow \beta_i$ (확률수렴) As $T \rightarrow \infty$
4. **BLUE (Gauss-Markov Theorem)**
5. $\Sigma\hat{\varepsilon}_t = 0$
6. $\Sigma\hat{\varepsilon}_t x_{t2} = \Sigma\hat{\varepsilon}_t x_{t3} = 0$ } $(\Sigma\hat{\varepsilon}_t x_{tk} = 0)$
7. $\Sigma\hat{\varepsilon}_t \hat{y}_t = 0$

가우스-마코프 정리(Gauss-Markov Theorm)

다중회귀모형의 가정들 하에서, 최소제곱추정량은 모든 선형불편 추정량들 가운데 가장 작은 분산을 갖는다.

이는 최소제곱추정량이 최우수선형 불편추정량(Best Linear Unbiased Estimators,BLUE)임을 의미함

분산

$$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \varepsilon_t$$

$$\text{var}(b_2) = \frac{\sigma^2}{(1-r_{23}^2)\sum(x_{t2}-\bar{x}_2)^2}$$

$$\text{var}(b_3) = \frac{\sigma^2}{(1-r_{23}^2)\sum(x_{t3}-\bar{x}_3)^2}$$

$r_{23} = 0$ 일 때, 이는 단순 선형 회귀모형에서의 공식으로 환원됨

$$\text{where } r_{23} = \frac{\sum(x_{t2}-\bar{x}_2)(x_{t3}-\bar{x}_3)}{\sqrt{\sum(x_{t2}-\bar{x}_2)^2}\sqrt{\sum(x_{t3}-\bar{x}_3)^2}}$$

LS추정량의 분산은 다음의 경우 작아짐:

1. 오차항의 분산, σ^2 가 작을 수록:

   ← 분자에 σ^2가 있음

2. 표본의 크기, T 가 클 수록:

   ← 분모에 $\sum_{t=1}^{T}(x_{t2}-\bar{x}_2)^2$ 가 있음

3. 설명변수의 값들이 퍼져 있을 수록:

   ← 분모에 $(x_{t2}-\bar{x}_2)^2$ 가 있음

4. 설명변수들간의 상관관계가 0에 가까울 수록:

   ← 분모에 $0 \le 1- r_{23}^2 \le 1$가 있음

x_2와 x_3 가 가깝게 연관되어 있는 경우 ==>
$var(b_2)$와 $var(b_3)$ 가 매우 커지게 됨. β_2 와 β_3 의 참값들을 알아 내기가 힘듦

다중회귀모형에서 설명변수와 관련한 추가적 가정이 필요한 이유를 다른 각도에서 알 수 있음:
독립변수들간에 정확한 선형관계가 없어야 함.
(완전한 공선성(collinearity)이 없어야 함,
i.e., $x_k \neq ax_j+b$)

공분산

$$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \varepsilon_t$$

$$\text{cov}(b_2, b_3) = \frac{-r_{23}\sigma^2}{(1-r_{23}^2)\sqrt{\sum(x_{t2}-\bar{x}_2)^2}\sqrt{\sum(x_{t3}-\bar{x}_3)^2}}$$

$$\text{where } r_{23} = \frac{\sum(x_{t2}-\bar{x}_2)(x_{t3}-\bar{x}_3)}{\sqrt{\sum(x_{t2}-\bar{x}_2)^2}\sqrt{\sum(x_{t3}-\bar{x}_3)^2}}$$

공분산에 영향을 미치는 요인들

두 모수에 대한 추정량간의 공분산은
다음의 경우 그 절대값이 커짐

1. 오차항의 분산, σ^2가 클 수록.
2. 표본의 크기, T가 작을 수록.
3. 변수들의 값이 덜 퍼져 있을 수록
4. 상관계수가, r_{23} 높을 수록.

분산-공분산 행렬(variance-covariance matrix)

$$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \varepsilon_t$$

The least squares estimators b_1, b_2, and b_3 have covariance matrix:

$$\text{cov}(b_1,b_2,b_3) = \begin{bmatrix} \text{var}(b_1) & \text{cov}(b_1,b_2) & \text{cov}(b_1,b_3) \\ \text{cov}(b_1,b_2) & \text{var}(b_2) & \text{cov}(b_2,b_3) \\ \text{cov}(b_1,b_3) & \text{cov}(b_2,b_3) & \text{var}(b_3) \end{bmatrix}$$

정규분포의 가정

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \ldots + \beta_K x_{Kt} + \varepsilon_t$$

$$y_t \sim N\left[(\beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \ldots + \beta_K x_{Kt}), \sigma^2\right]$$

또는 $\varepsilon_t \sim N(0, \sigma^2)$

b_k 는 y_t의 선형 함수임:

$$b_k \sim N\left[\beta_k, \text{var}(b_k)\right]$$

$$z = \frac{b_k - \beta_k}{\sqrt{\text{var}(b_k)}} \sim N(0,1) \quad \text{for } k = 1,2,\ldots,K$$

오차항 분산의 추정

오차항 분산에 대한 불편 추정량:

$$\hat{\sigma}^2 = \frac{\sum \hat{\varepsilon}_t^2}{T-K}$$

K: 추정해야하는 모수의 수

카이제곱 분포로의 변환:

$$\frac{(T-K)\,\hat{\sigma}^2}{\sigma^2} \sim \chi_{T-K}$$

t 분포

일반적으로 b_k 의 분산 $var(b_k)$ 는 알려져 있지 않으며, 따라서 이를 $\hat{\sigma}^2$ 를 σ^2대신 사용한 $\hat{var}(b_k)$로 추정함

$$t = \frac{b_k - \beta_k}{\sqrt{\hat{var}(b_k)}} = \frac{b_k - \beta_k}{se(b_k)}$$

t 는 자유도(df)=T–K인 t분포를 함.

구간추정

$$P\left[-t_c \le \frac{b_k - \beta_k}{se(b_k)} \le t_c\right] = 1 - \alpha$$

t_c 는 자유도가 T-K인 t분포에서 $P(t > t_c) = \alpha/2$ 와 같이 주어지는 임계값(critical value)임.

$$P\left[b_k - t_c\, se(b_k) \le \beta_k \le b_k + t_c\, se(b_k)\right] = 1 - \alpha$$

구간추정 또는 신뢰구간:

$$\left[b_k - t_c\, se(b_k),\ b_k + t_c\, se(b_k)\right]$$

개별모수에 대한 가설검정

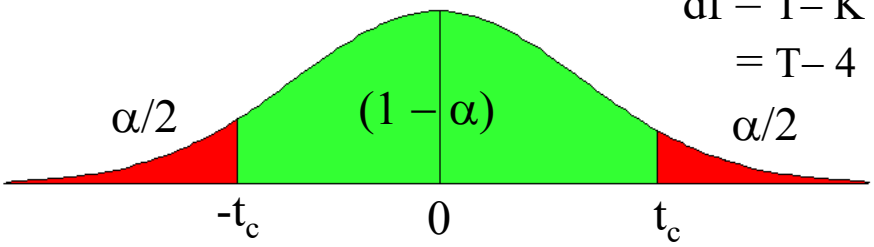
양측검정

$$y_t = \beta_1 + \beta_2 X_{t2} + \beta_3 X_{t3} + \beta_4 X_{t4} + \varepsilon_t$$

$H_0: \beta_2 = c$
$H_1: \beta_2 \neq c$

$$t = \frac{b_2 - c}{se(b_2)} \sim t_{(T-K)}$$

$df = T - K$
$= T - 4$



개별모수에 대한 가설검정

단측검정

$$y_t = \beta_1 + \beta_2 X_{t2} + \beta_3 X_{t3} + \beta_4 X_{t4} + e_t$$

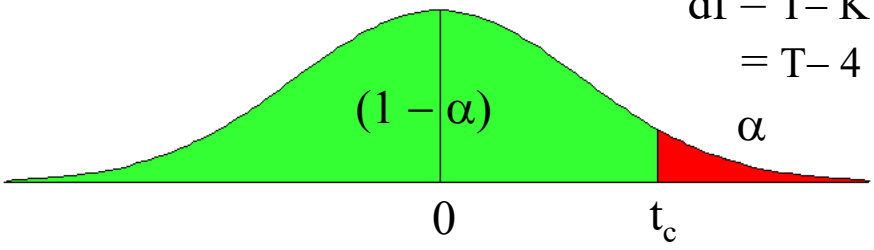
$H_0: \beta_3 = c$

$H_1: \beta_3 > c$

$$t = \frac{b_3 - c}{se(b_3)} \sim t_{(T-K)}$$

$df = T - K$

$= T - 4$



적합도 (Goodness of fit)

결정계수

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{t=1}^{T}(\hat{y}_t - \bar{y})^2}{\sum_{t=1}^{T}(y_t - \bar{y})^2}$$

$$0 \le R^2 \le 1$$

조절된 결정계수

Adjusted Coefficient of Determination

Original:

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

Adjusted:

$$\bar{R}^2 = 1 - \frac{SSE/(T-K)}{SST/(T-1)}$$

Situation

Each week the management of a Bay Area Rapid Food hamburger chain must decide how much money should be spent on advertising their products, and what specials (lower prices) should be introduced for that week.

How does total revenue change as the level of advertising expenditure changes? Does an increase in advertising expenditure lead to an increase in total revenue? If so, is the increase in total revenue sufficient to justify the increased advertising expenditure?

Management is also interested in pricing strategy. Will reducing prices lead to an increase or decrease in total revenue? If a reduction in price leads a decrease in total revenue then demand is price inelastic; If a price reduction leads to an increase in total revenue then demand is price elastic.

Economic and Econometric Model

We initially hypothesize that total revenue, *tr*, is linearly related to price, *p*, and advertising expenditure, *a*. The economic model is:

$$tr = \beta_1 + \beta_2 p + \beta_3 a$$

The economic model (7.1.1) describes the expected behavior of many individual franchises. As such we should write it as

$$E(tr) = \beta_1 + \beta_2 p + \beta_3 a$$

To allow for a difference between observable total revenue and the expected value of total revenue we add a *random error term*,

$$\varepsilon = tr - E(tr)$$

Denoting the tth weekly observation by the subscript t, we have

$$tr_t = E(tr_t) + e_t = \beta_1 + \beta_2 p_t + \beta_3 a_t + \varepsilon_t$$

Hypothesis Testing

Does the change in advertising expenditure change total revenue?

$$H_0 : \beta_3 = 0,\ H_1 : \beta_3 \neq 0$$

Does an increase in advertising expenditure lead to an increase in total revenue?

$$H_0 : \beta_3 \leq 0,\ H_1 : \beta_3 > 0$$

Is the increase in total revenue sufficient to justify the increased advertising expenditure?

$$H_0 : \beta_3 \leq 1,\ H_1 : \beta_3 > 1$$

Will reducing prices lead to an increase total revenue?

$$H_0 : \beta_2 \geq 0,\ H_1 : \beta_2 < 0$$

Will reducing prices lead to a decrease total revenue?

$$H_0 : \beta_2 \leq 0,\ H_1 : \beta_2 > 0$$