제6강 단순회귀분석

1. 계량경제학 개요
2. 회귀분석 및 회귀모형
3. 단순회귀모형
4. 회귀모형의 응용

1. 계량경제학 개요

(1) 정의

- Econometrics = Economic Measurement(양으로 측정된 경제학)
- 경제이론, 수학, 통계적 추론 등의 분석도구를 실제현상을 분석하는데 응용하는 사회과학(A. S. Goldberger)
- 경제이론, 경제수학, 통계학을 조화시킨 학문으로 세 분야와는 엄격히 구분되는 경제학의 한 분야(R. Frisch)
- 경제관계를 경험적으로 결정하는 경제학의 한 분야

(참고) 경제관계(economic relation)의 유형

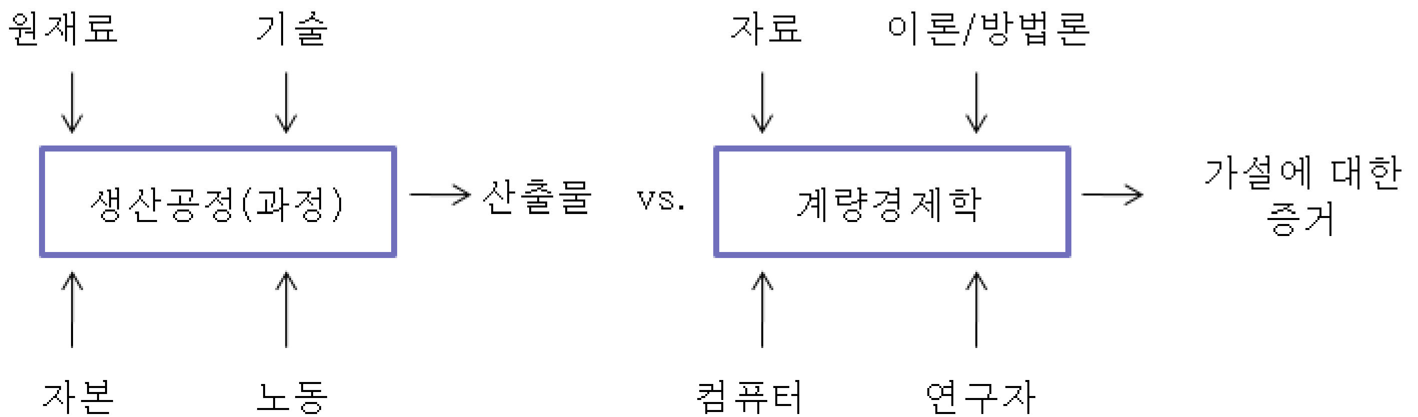
- 행위관계(behavioral relation) : $\boldsymbol{C = \alpha + \beta Y}$
- 기술관계(technological relation) : $\boldsymbol{Q = AK^{\alpha}L^{\beta}}$
- 제도관계(institutional relation) : $\boldsymbol{T = cY}$
- 항등관계(identity relation) : $\boldsymbol{Y = C + I + G + X - M}$
- 균형관계(equilibrium relation) : $\boldsymbol{D = S}$ (즉, demand=supply)

(2) 목적

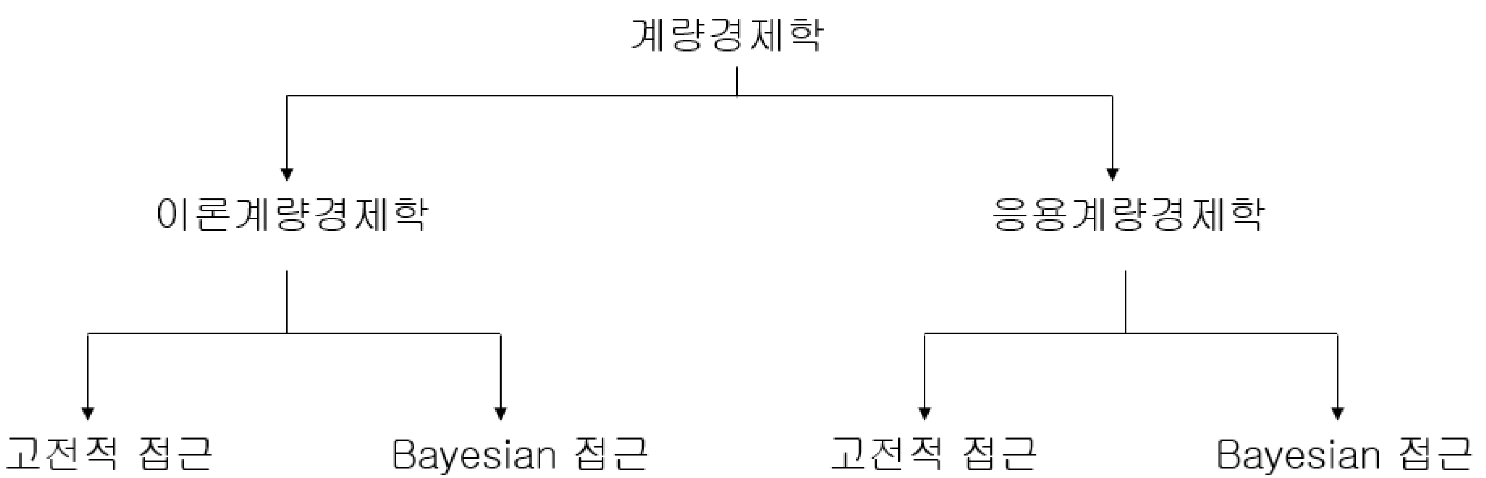
- 경제통계자료의 특징 분석 : 기술통계량, 요인분석 등 경제변수의 여러 특성을 파악
- 경제이론의 검증 : 경제학의 각 분야에서 개발된 경제이론이 현실경제를 설명하는데 적합한 자료인지를 실제 자료를 이용하여 분석
- 경제정책 분석 : 한 경제변수의 변화가 다른 경제변수에 어떤 영향을 미치는가를 분석
- 미래에 대한 예측 : 알려진 경제변수들간의 관계를 이용하여 특정 경제변수의 미래값을 예측
- 실증분석방법론의 개발 : 경제학의 실증연구에 적합한 방법론을 개발

(3) 중요성

- 계량경제학의 중요성은 원재료, 자본, 노동, 기술 등 생산요소를 사용하여 생산물을 만들어 내는 생산과정에 비유할 수 있음



(4) 분류



(5) 접근방법

① 계량모형의 설정 단계

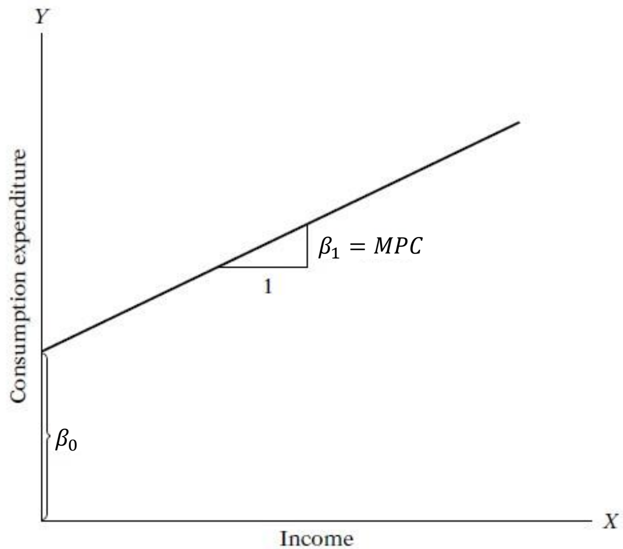
–가설의 설정, 변수의 선정, 모수의 부호와 크기, 모형의 수학적 형태 등을 고려

소비함수

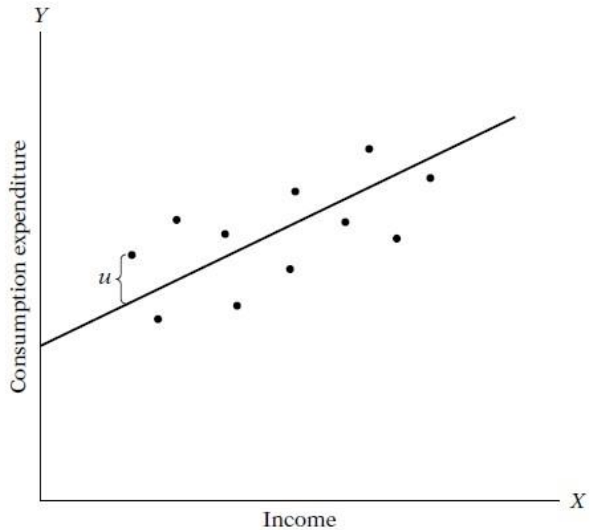
$$C = \beta_0 + \beta_1 Y \ (0 < \beta_1 < 1)$$

소비함수의 계량모형

$$C = \beta_0 + \beta_1 Y + U \ (0 < \beta_1 < 1)$$



Keynesian consumption function.



Econometric model of the Keynesian consumption function.

② 추정의 단계

-추정이란 이용 가능한 자료로 모형 내에 있는 모수(parameter)의 추정량을 얻는 것

-추정단계에서는 자료수집, 설명변수들 간의 상관관계 확인, 적절한 통계기법 등을 고려

(예: 소비함수)

-소득과 소비에 관한 자료를 모아 모형의 모수($\alpha, \beta, \sigma_u^2, \sigma_\alpha^2, \sigma_\beta^2$)를 추정

③평가의 단계

-추정결과의 통계적 유의성(statistical significance) 및 경제적 유의성(economical significance)을 살펴봄

-추정치에 대한 평가와 예측력에 대한 평가로 구분

-추정치 평가 : 경제이론, 통계, 계량경제적 기준에 의해 평가

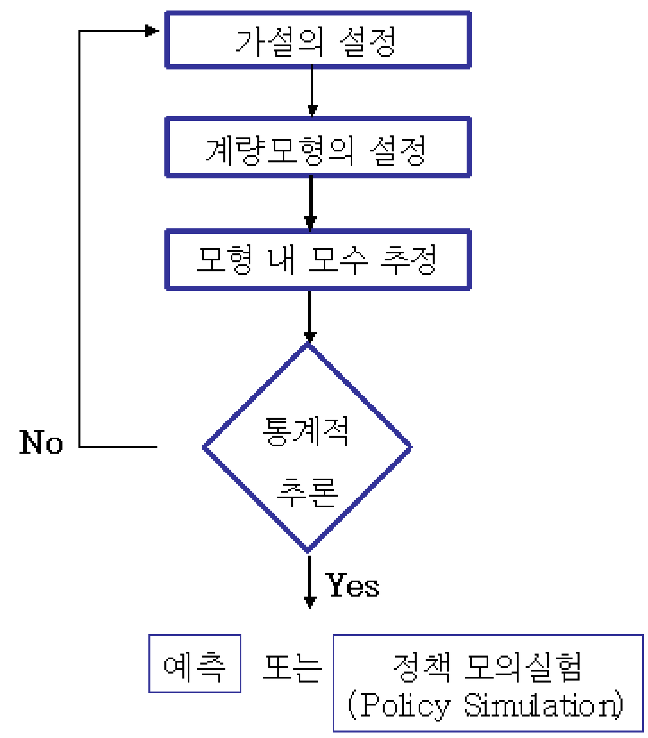
-예측력 평가 : 예측치와 실제치의 차에 대한 유의성을 고려

(예: 소비함수)

-주어진 독립변수(소득)의 값으로 종속변수(소비)의 미래 값을 예측

-소비를 바람직한 수준으로 유지하기 위해서는 소득수준이 얼마가 되어야 하는 지 등 정책 모의실험(policy simulation)을 시도

계량경제학 접근 순서도



(6) 자료의 형태

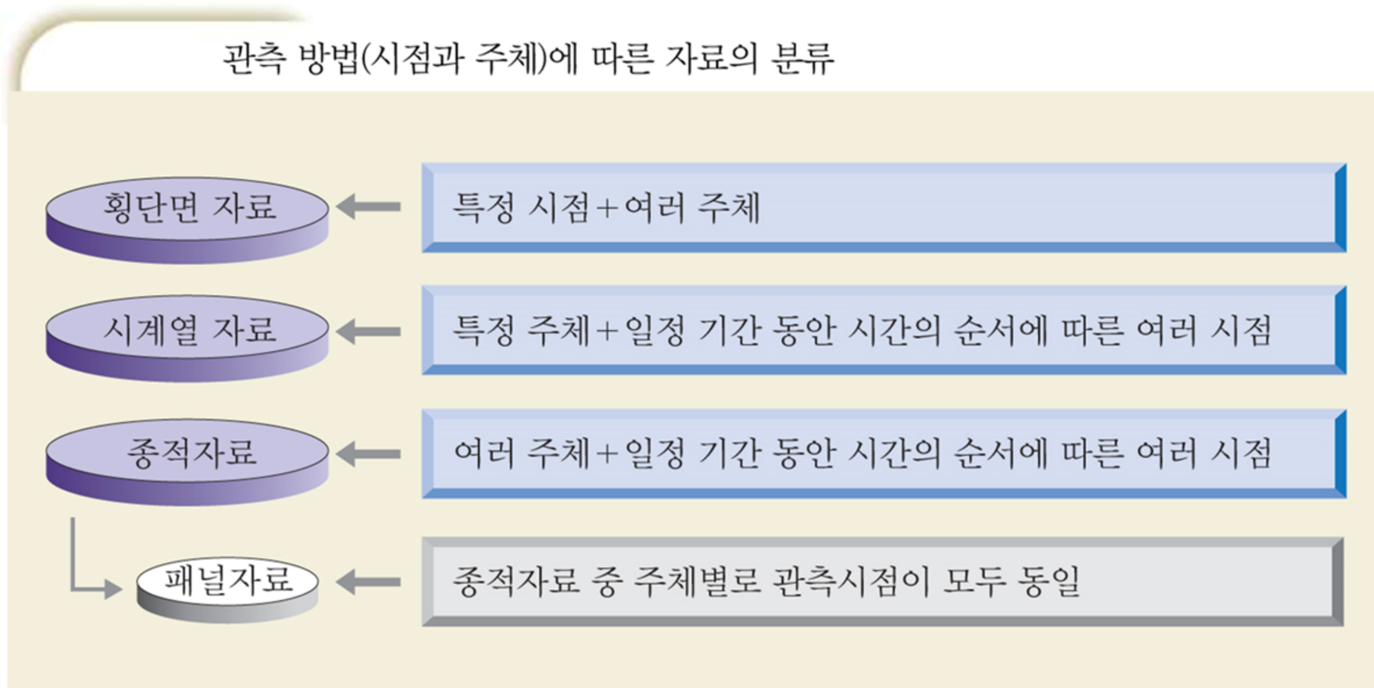
① 횡단면 자료(Cross-Sectional data) : 일정시점에서 하나 이상의 변수에 대해 수집된 자료로 미시경제변수를 측정한 자료에서 많이 발생

② 시계열 자료(Time Series data) : 일별, 주별, 월별, 분기별, 연도별 등 시간에 걸쳐 수집한 자료로 거시경제변수를 측정한 자료에서 많이 발생

③ 합동 횡단면 자료(Pooled Cross-Sectional data) : 횡단면 자료와 시계열자료가 결합된 자료

④ 패널자료(Panel data) : 동일한 횡단면 단위에 기준을 두고 시간의 흐름에 따라 수집한 자료

-(예) 특정 지역(가구)를 패널로 선택한 후에 매년 그 지역(가구)를 대상으로 주요 통계(GRDP, 교육비 지출)를 조사한 자료



(1) 회귀분석 개요

① 회귀(regression)란?

– 역사적 원천: Galton의 “보편적 회귀의 법칙”

– 현대적 의미: 종속변수와 독립변수와의 의존관계를 분석하고, 이미 알려진 독립변수의 값으로 종속변수의 평균적인 값을 추정 또는 예측

② 확정적(함수적) 관계 vs. 확률적(통계적) 관계

– 확정적 관계: $Y = \beta_0 + \beta_1 X$ vs. 확률적 관계: $Y = \beta_0 + \beta_1 X + u$

③ 회귀(regression)와 인과(causality)

–통계적 관계(회귀)가 아무리 강하다고 하더라도 반드시 인과관계를 나타내지 않으며 인과간계는 경제이론이나 선험적인 것으로부터 도출

④ 회귀분석(regression analysis)와 상관분석(correlation analysis)

회귀분석	상관분석
–독립변수는 확정변수로 가정하고 종속변수는 확률변수로 가정 –독립변수의 주어진 값으로 종속변수의 평균값을 추정·예측하는 것이 목적 –독립변수와 종속변수로 구분	–두 변수 모두 확률변수를 가정 –두 변수간의 선형성의 정도를 측정하는 것이 목적 –독립변수와 종속변수의 구분이 없음

⑤ 용어

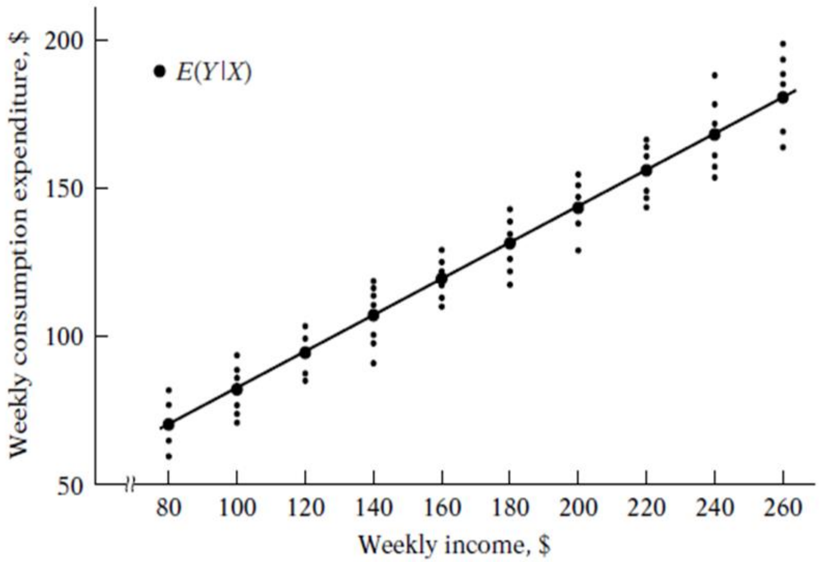
Y	X
Dependent variable(종속변수)	Independent variable(독립변수)
Explained variable(설명된 변수)	Explanatory variable(설명변수)
Predictand(예측된 변수)	Predictor(예측변수)
Regressand(피회귀변수)	Regressor(회귀변수)
Response(반응변수)	Stimulus(자극변수)
Endogenous variable(내생변수)	Exogenous variable(외생변수)
Controlled variable(통제된 변수)	Control variable(통제변수)
Outcome(결과)	Covariate(공변변수)

(2) 회귀모형의 개념

- 모집단 : 제주지역 물기업 60개
- Y(연간매출액, 단위 : 억 원)
- X(홍보비 지출액 , 단위:천 만원)

Y \ X	80	100	120	140	160	180	200	220	240	260
연간 매출액	55	65	79	80	102	110	120	135	137	150
	60	70	84	93	107	115	136	137	145	152
	65	74	90	95	110	120	140	140	155	175
	70	80	94	103	116	130	144	152	165	178
	75	85	98	108	118	135	145	157	175	180
	-	88	-	113	125	140	-	160	189	185
	-	-	-	115	-	-	-	162	-	191
합계	325	462	445	707	678	750	685	1043	966	1211
E(Y|X)	65	77	89	101	113	125	137	149	161	173

- X(홍보비 지출액)에 대한 Y(연간매출액)의 조건부 분포



Conditional distribution of expenditure for various levels of income (data of Table 2.1).

(3) 모집단 회귀함수 및 표본 회귀함수

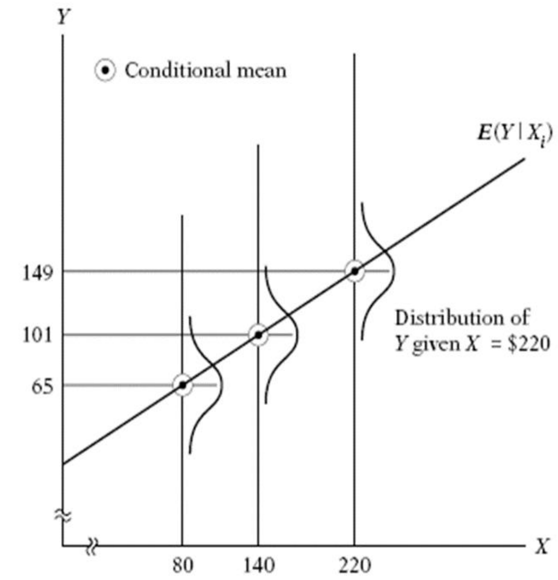
① 모집단 회귀함수

– 모집단 회귀함수: $\boldsymbol{E(Y|X_i) = \beta_0 + \beta_1 X_i}$

– 모집단 회귀함수의 확률적 형태:

$\boldsymbol{Y_i = E(Y|X_i) + u_i}$ 확정적(체계적) 요소

$\boldsymbol{Y_i = \beta_0 + \beta_1 X_i + u_i}$ ← 확률적(비체계적) 요소



② 표본 회귀함수

– 표본 회귀함수 : $\boldsymbol{\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i}$

– 표본 회귀함수의 확률적 형태 : $Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{u}_i$

– $\hat{Y}_i$: $E(Y|X_i)$의 추정량

– $\hat{\beta}_0$: β_0의 추정량

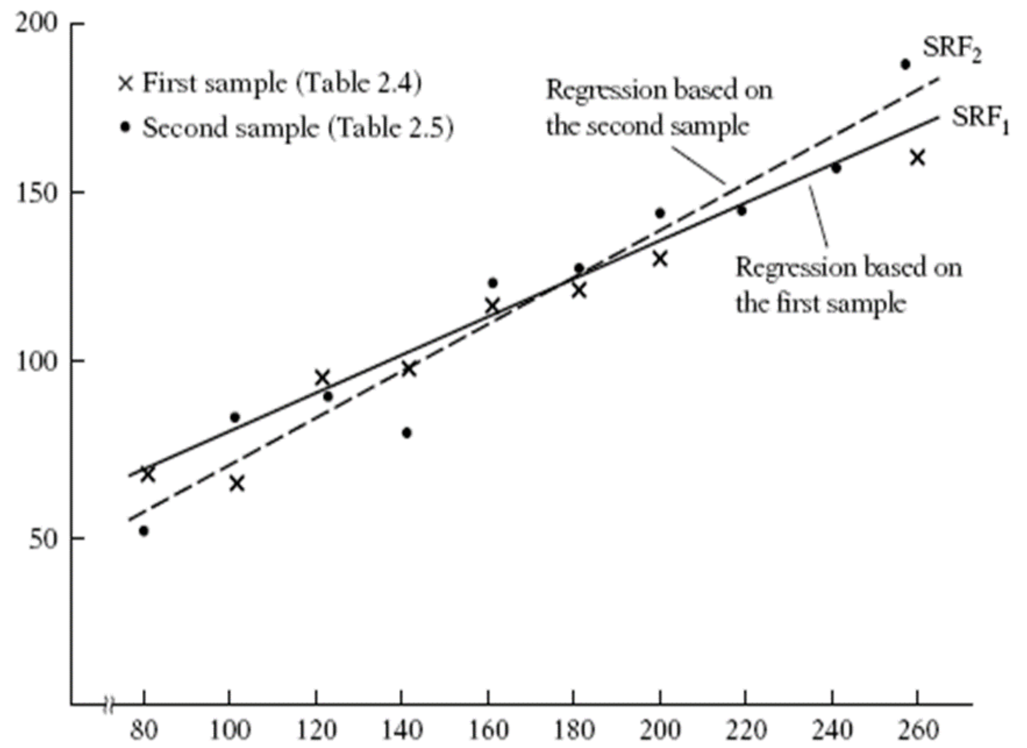
– $\hat{\beta}_1$: β_1의 추정량

확률표본1

Y	X
70	80
65	100
90	120
95	140
110	160
115	180
120	200
140	220
155	240
150	260

확률표본2

Y	X
55	80
88	100
90	120
80	140
118	160
120	180
145	200
135	220
145	240
175	260

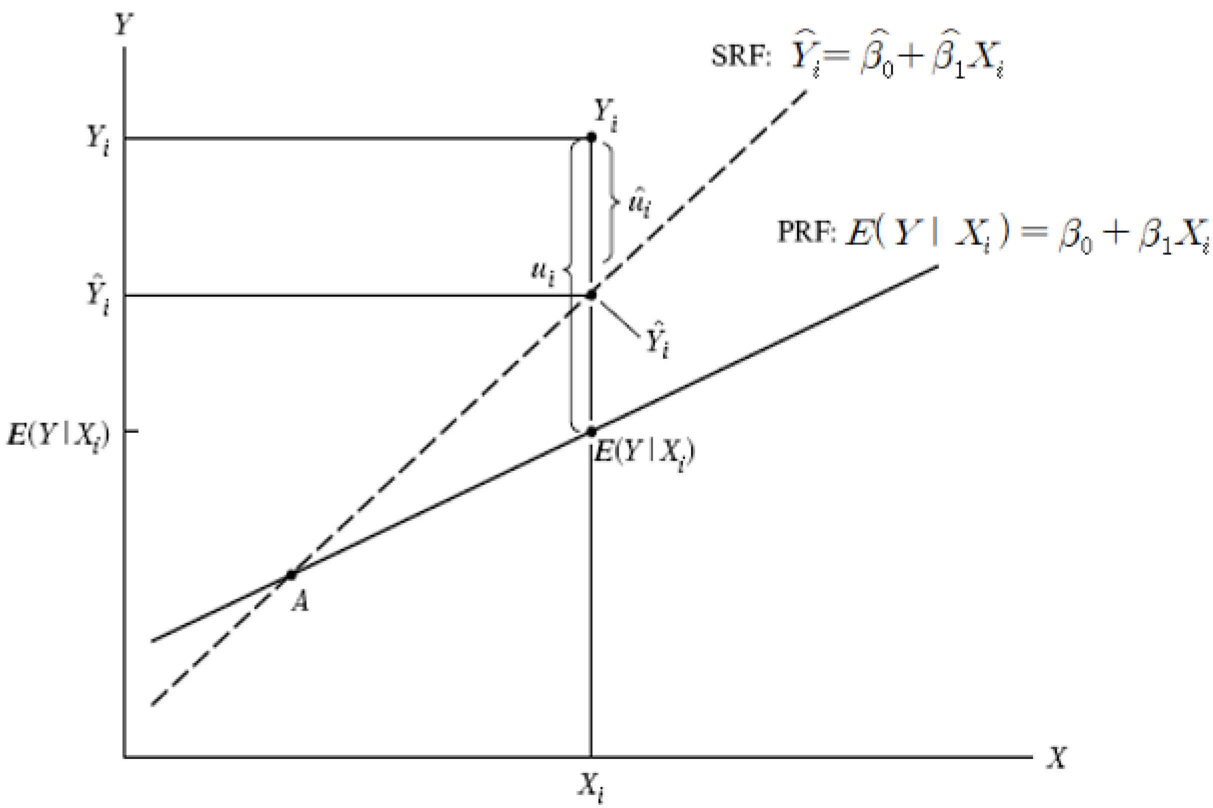


– 결론 : 회귀분석은 표본 회귀함수를 기초로 모집단 회귀함수를 추정

- 표본 회귀함수: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

(추정)

- 표본 회귀함수: $Y_i = \beta_0 + \beta_1 X_i + u_i$



(1) 모형

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

단, Y_i는 종속변수, X_i는 독립변수, β_0, β_1는 회귀계수, u_i는 교란(오차)항

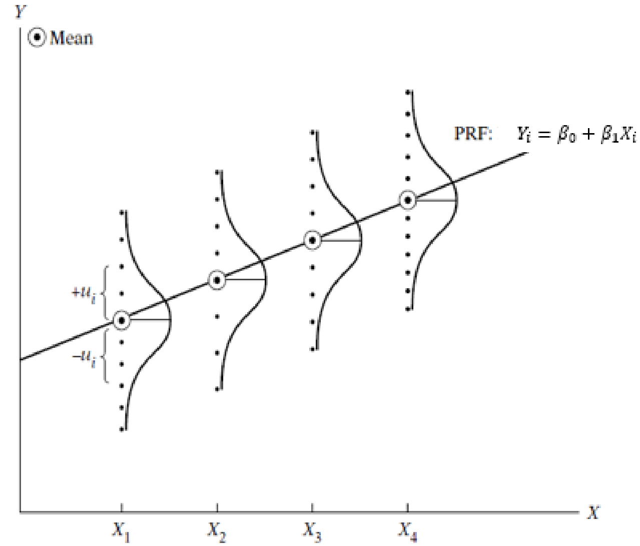
(참고) 확률적 교란항의 의미

-교란항은 생략되거나 배제된 모든 변수를 대신한다

-아무리 노력해도 설명할 수 없는 고유의 임의성(인간행동의 임의성)

-측정오차(measurement error)

-간결한 모형(parsimonious of model)



(2) 가정

① 독립변수 X는 확률변수가 아닌 확정변수임

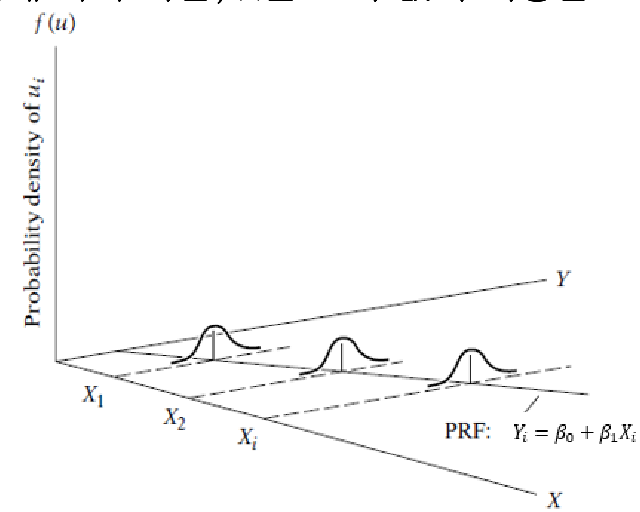
- 표본을 반복해서 잡아볼 때 X는 고정된 것으로 해 놓고 Y와 u의 값은 표본에 따라 다름, X는 오차 없이 측정됨

② 교란항의 평균은 0. 즉, $E(u_i) = 0, (i = 1, 2, 3, \ldots, n)$

③ 교란항은 모든 X에 대한 분산이 동일(동분산). 즉, $E\left(u_i^2 = \sigma_u^2\right)$

④ 모든 u값은 서로 독립임(비자기상관). 즉, $Cov(u_i, u_j) = 0,\ (i \neq j)$

⑤ u와 X는 서로 독립이다(직교조건). 즉, $Cov(u_i, X_i) = E(u_i, X_i) = 0$



(참고)직교조건의 의미

-경제적 의미 : 관계 있는 모든 설명변수들이 제대로 모형화되어 X가 Y에 미치는 영향과 u가 Y에 미치는 영향이 서로 분리되어 있다는 것을 의미

-통계적 의미 : 관계 있는 변수를 모형에 포함시키지 않으면→OLS추정량=편의 추정량

관계 없는 변수를 모형에 포함시키면→OLS추정량=비효율적 추정량

(3) 추정

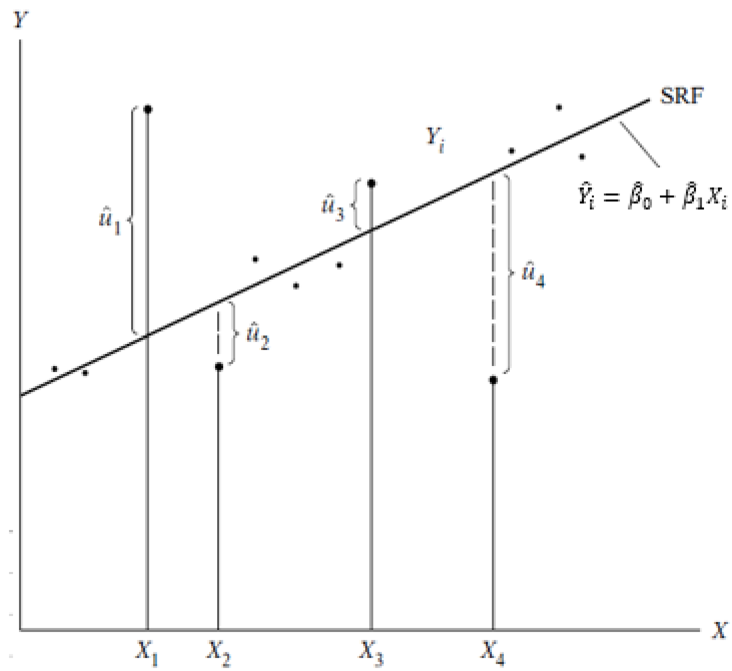
① 회귀계수

-보통최소자승법(Ordinary Least Squares : OLS) : 잔차(=실제치-예측치)의 합계가 최소가 되도록 하는 것이 바람직한데 잔차의 합이 0이 되는 식은 유일하지가 않으므로 잔차의 제곱의 합이 최소가 되게 하는 회귀식을 구하는 추정방법

(모형) $\boldsymbol{Y_i = \beta_0 + \beta_1 X_i + u_i}$

(추정회귀식) $\boldsymbol{\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i}$

(잔차) $\boldsymbol{e_i = Y_i - \hat{Y}_i}$



(보통최소자승법)

$$\underset{(\hat{\beta}_0,\hat{\beta}_1)}{\text{Min.}} \quad \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

(f.o.n.c)

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \hat{\beta}_0} = \frac{\partial \sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}{\partial \hat{\beta}_0} = 2\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)(-1) = 0$$

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \hat{\beta}_1} = \frac{\partial \sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}{\partial \hat{\beta}_1} = 2\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)(-X_i) = 0$$

정규방정식(normal equation)

$$\sum_{i=1}^{n} Y_i = \mathrm{n}\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i$$

$$\sum_{i=1}^{n} X_i Y_i = \hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2$$

정규방정식을 연립으로 풀면,

$$\sum_{i=1}^{n} X_i Y_i = \widehat{\beta_0} \sum_{i=1}^{n} X_i + \widehat{\beta_1} \sum_{i=1}^{n} X_i^2$$

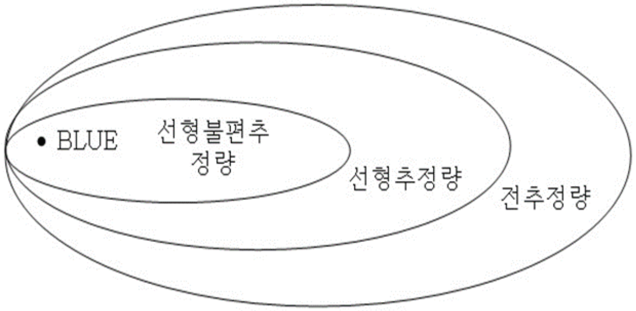
$$- \quad \bar{X} \sum_{i=1}^{n} Y_i = \mathrm{n}\widehat{\beta_0}\bar{X} + \widehat{\beta_1}\bar{X} \sum_{i=1}^{n} X_i$$

$$\sum_{i=1}^{n} X_i Y_i - \bar{X}\sum_{i=1}^{n} Y_i = \widehat{\beta_1}\left(\sum_{i=1}^{n} X_i^2 - \bar{X}\sum_{i=1}^{n} X_i\right)$$

$$\therefore \hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \bar{X}\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i^2 - \bar{X}\sum_{i=1}^{n} X_i} = \frac{cov(X,Y)}{var(X)} \qquad \bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}$$

$$\text{또는 } \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \qquad \therefore \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

-Gauss-Markov 정리 : 고전적 회귀모형에서 최소자승법으로 구한 추정량 $\hat{\beta}_0$ 및 $\hat{\beta}_1$은 선형불편추정량 중에서 최소의 분산을 갖는다. 즉, 최소자승법으로 구한 추정량은 BLUE(Best Linear Unbiased Estimator)임



(예제) 다음의 홍보비 지출액 X(단위:천만 원)와 연간 매출액 Y(단위:억 원)에 관한 자료를 이용하여 회귀계수와 추정 회귀식을 구하고 해석하라.

X	2	3	4	5	6
Y	4	4	6	6	10

(기초 계산)

$$\sum X_i = 20,\ \sum Y_i = 30,\ \sum X_i Y_i = 134,$$
$$\bar{X} = 4,\ \bar{Y} = 6,\ \sum X_i^2 = 90,\ \sum Y_i^2 = 204$$

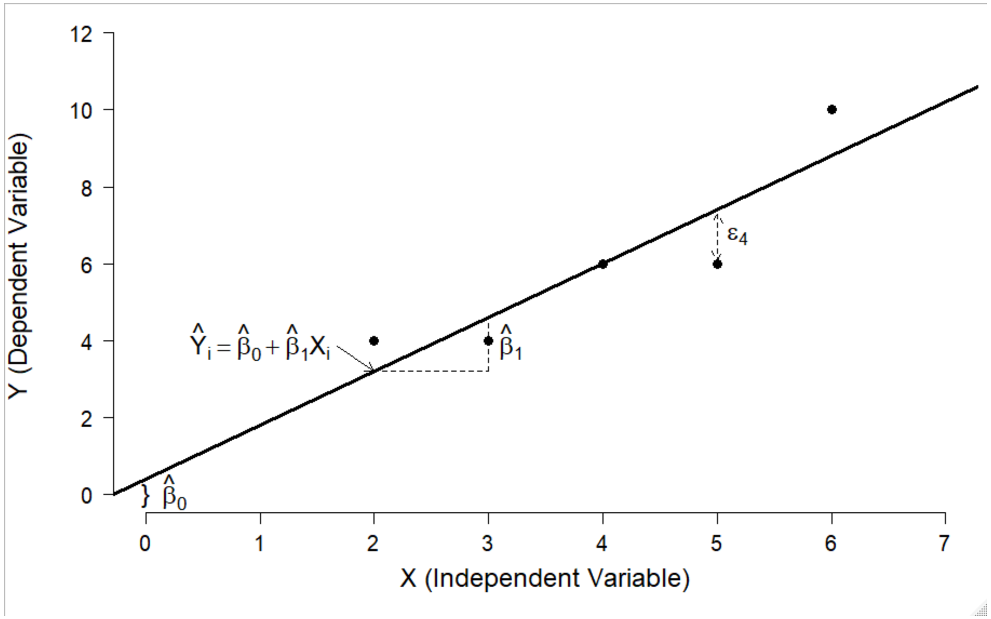
(회귀계수 계산)

$$\therefore \hat{\beta}_1 = \frac{\sum_{i=1}^n X_i Y_i - \bar{X}\sum_{i=1}^n Y_i}{\sum_{i=1}^n X_i^2 - \bar{X}\sum_{i=1}^n X_i} = \frac{134-4\times30}{90-4\times20} = \frac{14}{10} = 1.4$$

$$\therefore \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 6 - (1.4)(4) = 0.4$$

(추정 회귀식)

$$\hat{Y}_i = 0.4 + 1.4X_i$$



(해석)

-홍보비 지출액이 천 만원 증가하면 연간 매출액은 평균 1.4억 원 증가한다

② 분산

e_i =due to residual

-교란항의 분산

$\hat{\beta}_0 + \hat{\beta}_1 X_i$

$\cdot \hat{\sigma}_u^2 = \frac{\sum_{i=1}^n \hat{e}_i^2}{n-2}$ 단, $\sum_{i=1}^n \hat{e}_i^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2 - \hat{\beta}_1 \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$

-회귀계수의 분산

$\cdot \mathrm{Var}(\hat{\beta}_1) = \mathrm{E}[(\hat{\beta}_1 - E(\hat{\beta}_1))^2] = \mathrm{E}[(\hat{\beta}_1 - \beta_1)^2] = \hat{\sigma}_u^2 \frac{1}{\sum_{i=1}^n x_i^2}$

$\therefore \hat{\beta}_1 \sim (\beta_1, \hat{\sigma}_u^2 \frac{1}{\sum_{i=1}^n x_i^2})$

(교란항 및 회귀계수의 분산 계산)

$\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - \bar{X} \sum_{i=1}^n X_i$ = 90−(4)(20)=10

$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n Y_i^2 - \bar{Y} \sum_{i=1}^n Y_i$ = 204−(6)(30)=24

$\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^n X_i Y_i - \bar{X} \sum_{i=1}^n Y_i$ = 134−(4)(30)=14

$\sum_{i=1}^n \hat{e}_i^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2 - \hat{\beta}_1 \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$ = 24 − (1.4)(14) = 4.4

$\therefore \hat{\sigma}_u^2 = \frac{4.4}{3} = 1.4667$

var($\hat{\beta}_1$) = $\hat{\sigma}_u^2 \frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2}$ = (1.4667)($\frac{1}{10}$)=0.14667

③ 결정계수(coefficient of determination)

-표본으로부터 추정한 회귀선이 변수의 표본관측에 얼마나 적합한 지를 측정하는 계수로 설명력(explanatory power)이라고도 함

-결정계수는 표본관측이 추정된 회귀식에 가까운 정도를 계수로 나타낸 것으로 종속변수의 총변동과 회귀변동의 비율로 측정

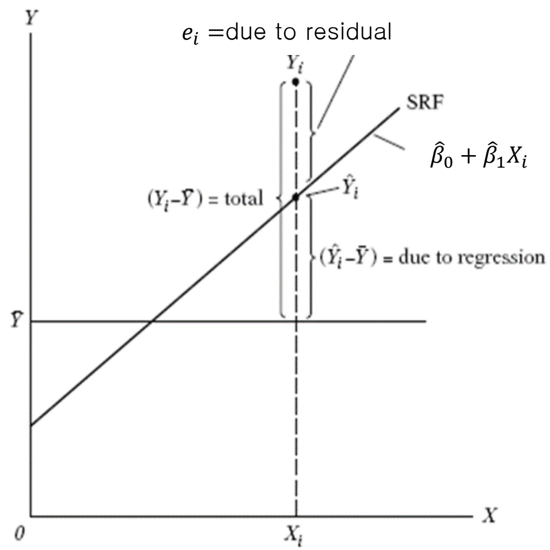
-결정계수의 값은 0과 1 사이로, 큰 값일수록 적합도 또는 설명력이 높다는 것을 의미하고 작은 값일수록 적합도 또는 설명력이 낮다는 것을 의미함

-우측 그림에서 $Y_i = \hat{Y}_i + e_i$ 또는 $Y_i - \bar{Y} = \hat{Y}_i - \bar{Y} + e_i$

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} e_i^2$$

(총변동) = 회귀변동 + 잔차변동

$$R^2 = \frac{\text{회귀변동}}{\text{총변동}} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = \frac{\hat{\beta}_1^2 \sum_{i=1}^{n}(X_i - \bar{X})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$



(결정계수의 계산)

$$\boldsymbol{R^2 = \frac{\hat{\beta}_1^2 \sum_{i=1}^{n}(X_i - \bar{X})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = \frac{19.6}{24} = 0.817}$$

(4) 최우추정법

① 배경

–보통최소자승법 외에 회귀계수를 추정하는 또 다른 방법으로는 최우추정법(Maximum Likelihood Estimation: MLE)이 있음

–보통최소자승법을 이용한 회귀계수의 추정에서는 교란항(오차항)의 분포에 대한 가정은 필요 없었는데 최우추정법을 사용하기 위해서는 교란항에 대한 가정이 필요하며 일반적으로 정규분포를 가정함

② 모형 및 가정

$- \boldsymbol{Y_i = \beta_0 + \beta_1 X_i + u_i,\ Y_i \sim NI(\beta_0 + \beta_1 X_i, \sigma^2)}$

③ 결합확률밀도함수

–$\boldsymbol{Y_1, Y_2, \ldots, Y_n}$의 결합확률밀도함수는 다음과 같음

$$\cdot\ \boldsymbol{f(Y_1, Y_2, \ldots, Y_n | \beta_0 + \beta_1 X_i, \sigma^2) = f(Y_1 | \beta_0 + \beta_1 X_i, \sigma^2)\, f(Y_2 | \beta_0 + \beta_1 X_i, \sigma^2) \ldots f(Y_n | \beta_0 + \beta_1 X_i, \sigma^2)}$$

단, $f(Y_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 X_i)^2}$(정규분포의 확률밀도함수)

–따라서, $\boldsymbol{Y_1, Y_2, \ldots, Y_n}$의 결합확률밀도함수는 다음과 같이 나타낼 수 있음

$$\cdot f(Y_1, Y_2, \ldots, Y_n | \boldsymbol{\beta_0 + \beta_1 X_i, \sigma^2}) = \frac{1}{(\sqrt{2\pi\sigma^2})^n} e^{[-\frac{1}{2}\sum\frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{\sigma^2}]} \quad \text{(a)}$$

–(a)식에서 $\boldsymbol{Y_1, Y_2, \ldots, Y_n}$이 알려져 있고, $\boldsymbol{\beta_0, \beta_1, \sigma^2}$이 알려져 있지 않을 경우 이를 우도함수(likelihood function)이라 하고 $\boldsymbol{L(\beta_0, \beta_1, \sigma^2)}$로 표기하는데 다음의 (b)식과 같음

$$\cdot L(\boldsymbol{\beta_0, \beta_1, \sigma^2}) = \frac{1}{(\sqrt{2\pi\sigma^2})^n} e^{[-\frac{1}{2}\sum\frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{\sigma^2}]} \quad \text{(b)}$$

④ 최우추정량

– 주어진 Y의 값들을 관측할 확률을 최대로 하는 모수($\boldsymbol{\beta_0, \beta_1, \sigma^2}$)을 추정하는 방법으로 즉, 우도함수((b)식)의 최댓값을 찾는 방법임

– (b)식의 미분값을 간단히 구하기 위하여 (b)식에 로그를 취하면 (c)식과 같게 됨

- $lnL = -\frac{n}{2}ln\sigma^2 - \frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum\frac{(Y_i-\beta_0-\beta_1X_i)^2}{\sigma^2}$ (c)

– (c)식을 $\boldsymbol{\beta_0, \beta_1, \sigma^2}$에 대해 편미분하고 최우추정량을 $\boldsymbol{\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\sigma}^2}$로 나타내면 (d.1)–(d.3)식과 같음

- $\frac{\boldsymbol{\partial lnL}}{\boldsymbol{\partial\widehat{\beta}_0}} = -\frac{\boldsymbol{1}}{\boldsymbol{\sigma^2}}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1X_i)(-1) = 0$ (d.1)

- $\frac{\boldsymbol{\partial lnL}}{\boldsymbol{\partial\widehat{\beta}_1}} = -\frac{\boldsymbol{1}}{\boldsymbol{\sigma^2}}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1X_i)(-X_i) = 0$ (d.2)

- $\frac{\boldsymbol{\partial lnL}}{\boldsymbol{\partial\widehat{\sigma}^2}} = -\frac{\boldsymbol{n}}{\boldsymbol{2\sigma^2}} + \frac{\boldsymbol{1}}{\boldsymbol{2\sigma^4}}\sum(\boldsymbol{Y_i - \beta_0 - \beta_1X_i})^2 = \boldsymbol{0}$ (d.3)

– (d.1)–(d.3)식으로부터 다음의 (e.1) 및 (e.2)식을 얻게 되는데 이는 최소자승법에서의 정규방정식과 동일하므로 회귀계수에 대한 OLS 추정량 및 MLE 추정량은 동일함

$$\sum Y_i = n\hat{\beta}_0 + \hat{\beta}_1\sum X_i$$

$$\sum X_iY_i = \hat{\beta}_0\sum X_i + \hat{\beta}_1\sum X_i^2$$

– (d.3)식으로부터 다음의 분산에 대한 최우추정량을 얻을 수 있는데 분산추정량은 점근적 불편추정량이 됨

$$\boldsymbol{\widehat{\sigma}^2 = \frac{1}{n}\sum(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1X_i)^2 = \frac{1}{n}\sum e_i^2}$$

(5) 가설검정

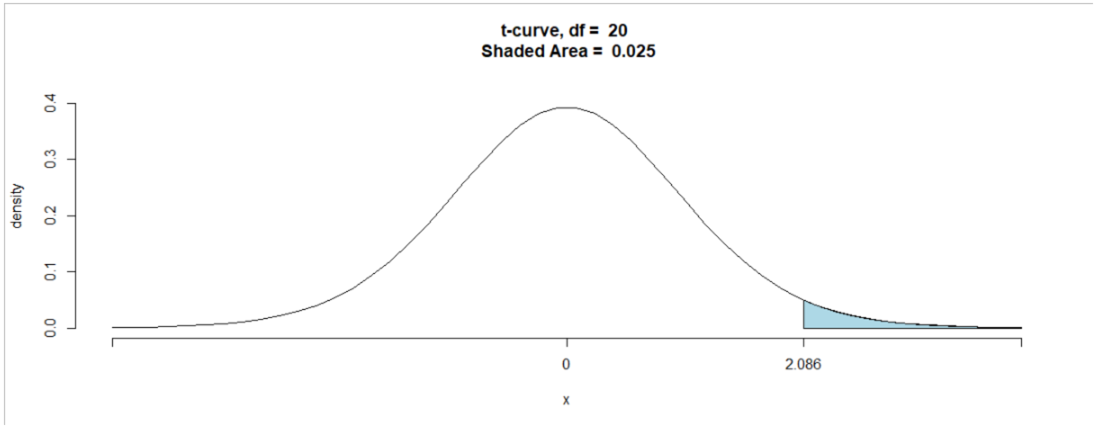
-통계적 유의성(statistical significance) : 실증분석 결과 독립변수가 종속변수에 영향을 주는 것

(모형) $\boldsymbol{Y_i = \beta_0 + \beta_1 X_i + u_i}$

(가설) $H_0: \beta_1 = 0 \;\; H_1: \beta_1 \neq 0$

(검정통계량) $t = \dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{var(\hat{\beta}_1)}} = \dfrac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} \sim t_{n-2}$

(신뢰구간) $\hat{\beta}_1 \pm t_{(n-2,\frac{\alpha}{2})}\, se(\hat{\beta}_1)$



(회귀계수에 대한 가설검정)

(추정회귀식) $\widehat{Y_i} = 0.4 + 1.4X_i$
(1.62) (0.383)

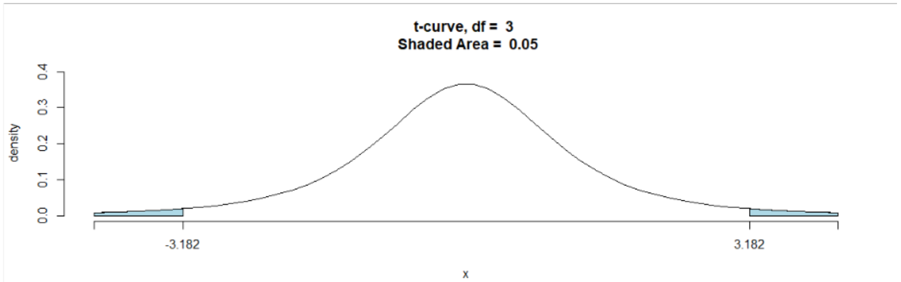
(가설) $H_0: \beta_1 = 0 \;\; H_1: \beta_1 \neq 0$

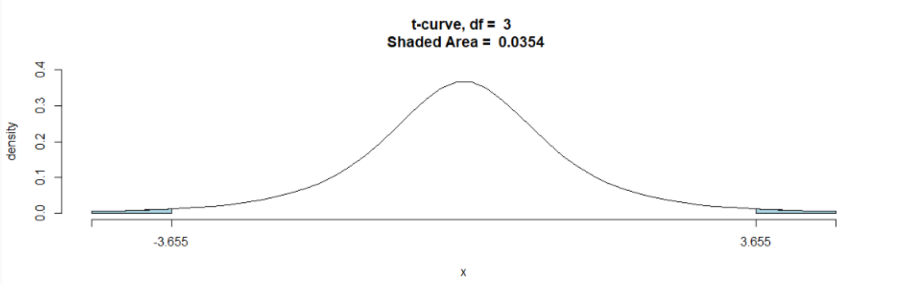
(홍보비 지출액이 연간매출액에 영향을 주지 않는다)

(검정통계량) $t = \dfrac{1.4 - 0}{0.383} = 3.655 \sim t_3$

(신뢰구간) $\hat{\beta}_1 \pm t_{(n-2,\frac{\alpha}{2})}\, se(\hat{\beta}_1)$

$= 1.4 \pm 3.182(0.383) = [0.181, 2.618]$





(6) 예측

-모형설정→모형추정→가설검정→예측이라는 단순회귀분석의 절차에 따라 주어진 독립변수의 값에 대한 종속변수의 값을 구하는 예측(forecasting, prediction)이라고 함

-모집단회귀선 위의 한 점을 예측하는 평균예측(mean prediction)과 독립변수의 값에 대응하는 개별 Y의 값을 예측하는 개별예측(individual prediction)이 있음

-예측에는 주어진 X에 대해 하나의 Y 값을 구하는 점예측(point forecasting)과 점예측에 대한 신뢰구간을 구하는 구간예측(interval forecasting)이 있음

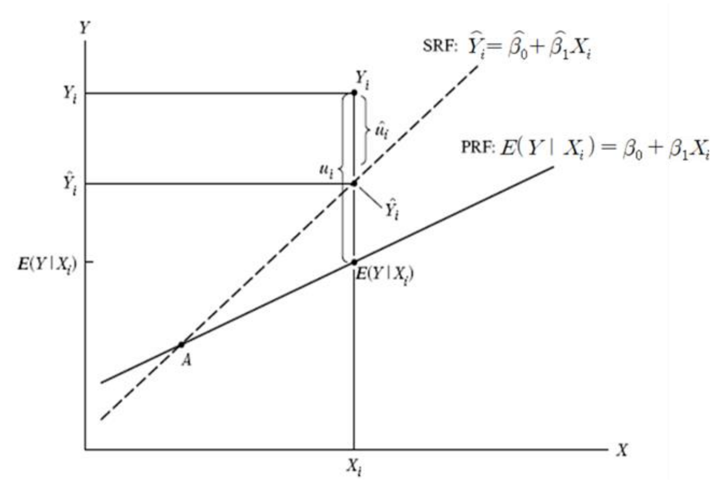
-구간예측을 위해서는 실제치와 예측치의 차이인 예측오차의 분산을 알아야 함

-일반적으로 개별예측에서 점예측치 및 예측구간대를 구함

-$\boldsymbol{Y_i = \beta_0 + \beta_1 X_i + u_i}$에서 독립변수 $\boldsymbol{X_i = X_0}$일 때 평균예측과 개별예측은 각각 다음과 같음

(평균예측) $\boldsymbol{E(Y_0|X_0) = \beta_0 + \beta_1 X_0}$

(개별예측) $\boldsymbol{Y_0 = \beta_0 + \beta_1 X_0 + u_0}$



-평균예측과 개별예측의 점 예측치(이는 추정량이 됨)는 모두 동일하며 다음과 같음

(점예측치) $\widehat{Y_0} = \hat{\beta}_0 + \hat{\beta}_1 X_0$

-예측오차는 평균예측(또는 개별예측)과 점 예측치(추정량)의 차이 이므로 각각 다음과 같음

(평균예측오차) $E(Y_0|X_0) - \widehat{Y_0} = \beta_0 + \beta_1 X_0 - (\hat{\beta}_0 + \hat{\beta}_1 X_0) = (\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)X_0$

(개별예측오차) $Y_0 - \widehat{Y_0} = \beta_0 + \beta_1 X_0 + u_0 - (\hat{\beta}_0 + \hat{\beta}_1 X_0) = (\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)X_0 + u_0$

-따라서 예측오차의 분산은 각각 다음과 같음

(평균예측오차분산) $\mathrm{var}((E(Y_0|X_0) - \widehat{Y_0}) = \mathrm{var}(\hat{\beta}_0) + X_0^2 \mathrm{var}(\hat{\beta}_1) + 2X_0 cov(\hat{\beta}_0, \hat{\beta}_1)$

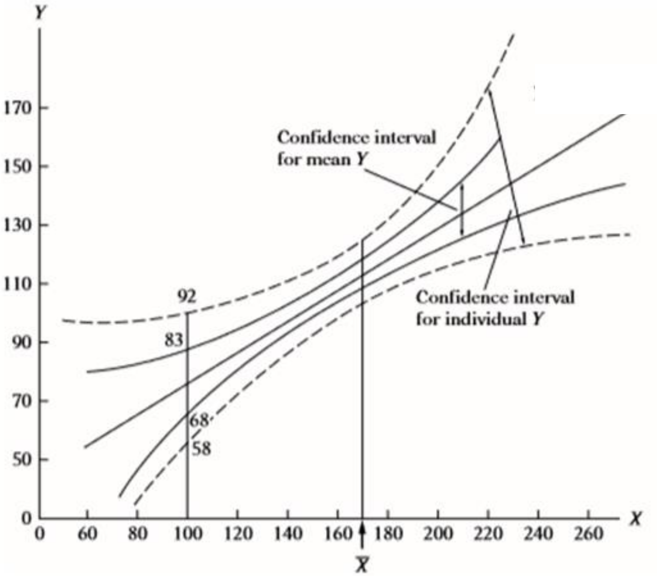
(개별예측오차분산) $var(Y_0 - \widehat{Y_0}) = \mathrm{var}(\hat{\beta}_0) + X_0^2 \mathrm{var}(\hat{\beta}_1) + 2X_0 cov(\hat{\beta}_0, \hat{\beta}_1) + \sigma_u^2$

(점예측) $\widehat{Y_0} = \hat{\beta}_0 + \hat{\beta}_1 X_0 \ (X = X_0$일 때)

(평균예측의 예측오차 분산) $\sigma_\epsilon^2 = \sigma_u^2(\frac{1}{n} + \frac{(X_0-\bar{X})^2}{\sum_{i=1}^{n} x_i^2})$

(개별예측의 예측오차 분산) $\sigma_\epsilon^2 = \sigma_u^2(1 + \frac{1}{n} + \frac{(X_0-\bar{X})^2}{\sum_{i=1}^{n} x_i^2})$

(예측구간) $\hat{\beta}_0 + \hat{\beta}_1 X_0 \pm \ t_{(n-2,\frac{\alpha}{2})}\ \sigma_\epsilon$



(예측오차의 분산에 대한 해석)

-표본의 크기(n)가 커질수록 예측오차의 분산이 작아짐. 즉, 관측자료가 많을수록 좋은 예측을 할 수 있음

-교란항의 분산이 커질수록 예측오차의 분산이 커짐. 즉, 원래 회귀모형에서 불확실성이 커서 교란항의 분산이 크면 예측이 어려울 수밖에 없음

- 독립변수의 표본평균으로부터 멀어질수록 예측오차도 커짐. 즉, 예측에 주어진 독립변수의 값이 평균으로부터 멀어질수록 표본회귀함수의 예측력은 크게 감소함

(점예측) 추정 회귀식 $\hat{Y} = \mathbf{0.4} + \mathbf{1.4}\,X$에서 홍보비 지출액이 7(천만 원)일 때 즉, $X_0 = 7$일 때 연간매출액의 개별예측치 및 평균예측치는 $\hat{Y} = \mathbf{0.4} + (\mathbf{1.4})(\mathbf{7})$ =10.2억 원

−$\hat{\sigma}_u^2$= 1.4667, 독립변수의 평균 $\bar{X} = 4$, 표본의 크기 $n = 5$, $\sum_{i=1}^{n} x_i^2 = 10$

(평균예측의 예측오차 분산) $\sigma_\epsilon^2 = \sigma_u^2\left(\frac{1}{n} + \frac{(X_0-\bar{X})^2}{\sum_{i=1}^{n} x_i^2}\right) = (1.4667)(\frac{1}{5} + \frac{(7-4)^2}{10}) = 1.61$

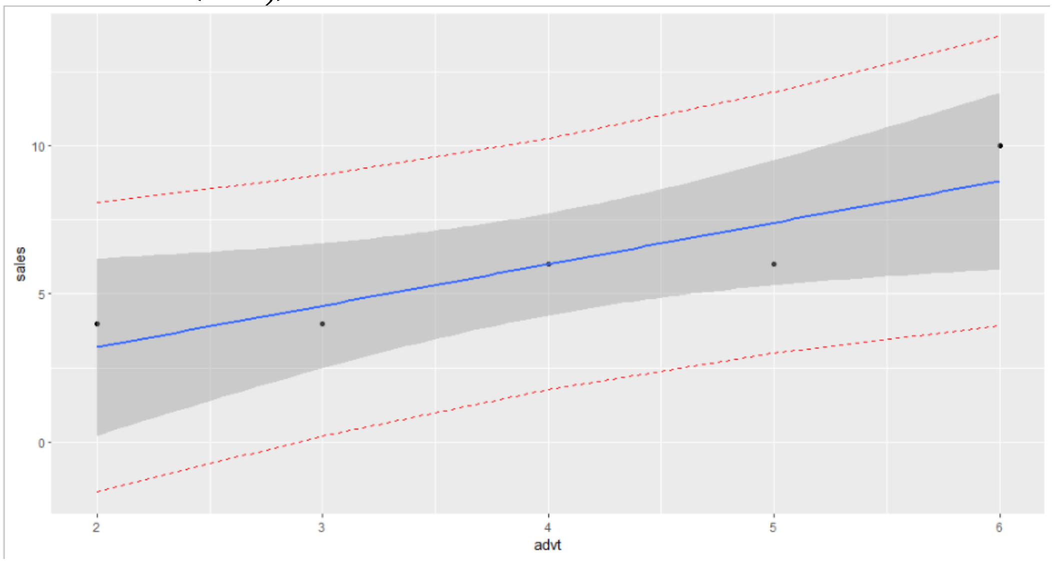
(개별예측의 예측오차 분산) $\sigma_\epsilon^2 = \sigma_u^2\left(1 + \frac{1}{n} + \frac{(X_0-\bar{X})^2}{\sum_{i=1}^{n} x_i^2}\right) = (1.4667)(1 + \frac{1}{5} + \frac{(7-4)^2}{10}) = 3.08$

(95% 예측구간) $\hat{\beta}_0 + \hat{\beta}_1 X_0 \pm t_{(n-2,\frac{\alpha}{2})}\,\sigma_\epsilon$

−점예측치가 10.2이고, 예측오차의 분산이 각각 3.08, 1.61이며, $t_{3,0.025}$=3.182

(평균예측구간) $\hat{\beta}_0 + \hat{\beta}_1 X_0 \pm t_{(n-2,\frac{\alpha}{2})}\,\sigma_\epsilon = 10.2 \pm (3.182)\sqrt{1.61} = [6.15, 14.24]$

(개별예측구간) $\hat{\beta}_0 + \hat{\beta}_1 X_0 \pm t_{(n-2,\frac{\alpha}{2})}\,\sigma_\epsilon = 10.2 \pm (3.182)\sqrt{3.08} = [4.61, 15.78]$



4. 회귀모형의 응용

(1) 회귀함수 형태 선택기준

- 경제이론에 근거한 함수형태를 선택 : 생산함수의 경우 →경제이론에 근거하여 Cobb-Douglas 생산함수를 선택
- 가능한 한 간단한 함수형태를 선택 : 모형의 설명력에 큰 차이가 없다면 경제학의 효율성 원칙에 따라 간단한 함수형태를 선택하는데 이를 간결성원칙(parsimonious principle)이라고 함
- 예측력이 좋은 함수형태를 선택 : 간결한 모형이 좋기는 하지만 모형의 현실 설명력 즉 예측력이 떨어진다면 모형의 유용성이 크게 떨어질 수밖에 없음

(2) 회귀모형의 응용

① 선형(linear)함수 모형

(모형) $Y_i = \beta_0 + \beta_1 X_i + u_i$

(기울기) $\frac{dY}{dX} = \beta_1$ ($dY = \beta_1 dX$이므로)

(탄력도) $\varepsilon_{Y|X} = \frac{\frac{dY}{Y}}{\frac{dX}{X}} = \frac{X}{Y}\frac{dY}{dX} = \beta_1 \frac{X}{Y}$

② 전대수(double log)함수 모형

(모형) $lnY_i = ln\beta_0 + \beta_1 lnX_i + u_i$ ⬅ $Y = \beta_0 X^{\beta_1}$

(기울기) $\frac{dY}{dX} = \beta_1 \frac{Y}{X}$ ($\frac{dY}{Y} = \beta_1 \frac{dX}{X}$이므로) ⬅ $\frac{d(lny)}{d(lnx)} = \frac{dy}{dx}\frac{x}{y} = \varepsilon_{y|x} = \beta_1$

(탄력도) $\varepsilon_{Y|X} = \frac{\frac{dY}{Y}}{\frac{dX}{X}} = \frac{X}{Y}\frac{dY}{dX} = \frac{X}{Y}(\beta_1 \frac{Y}{X}) = \beta_1$ $\because$ $Y = \beta_0 X^{\beta_1}$에서 $dY = \beta_1 \beta_0 X^{\beta_1 - 1} dX = \beta_1 \frac{Y}{X} dX$ 이므로

③ 쌍곡선(reciprocal)함수 모형

(모형) $Y_i = \beta_0 + \beta_1 \frac{1}{X_i} + u_i$

(기울기) $\frac{dY}{dX} = -\beta_1 \frac{1}{X^2}$ ($dY = -\beta_1 X^{-2} dX$이므로)

(탄력도) $\varepsilon_{Y|X} = \frac{\frac{dY}{Y}}{\frac{dX}{X}} = \frac{X}{Y}\frac{dY}{dX} = \frac{X}{Y}(-\beta_1 X^{-2}) = -\beta_1 \frac{1}{XY}$

④ 반로그(semi-log)함수 모형

(모형) $lnY_i = \beta_0 + \beta_1 X_i + u_i$ / $Y_i = \beta_0 + \beta_1 lnX_i + u_i$

(기울기) $\frac{dY}{dX} = \beta_1 Y$ ($\frac{dY}{Y} = \beta_1 dX$이므로) / $\frac{dY}{dX} = \beta_1 \frac{1}{X}$ ($dY = \beta_1 \frac{dX}{X}$이므로)

(탄력도) $\varepsilon_{Y|X} = \frac{\frac{dY}{Y}}{\frac{dX}{X}} = \frac{X}{Y}\frac{dY}{dX} = \frac{X}{Y}(\beta_1 Y) = \beta_1 X$ / $\varepsilon_{Y|X} = \frac{\frac{dY}{Y}}{\frac{dX}{X}} = \frac{X}{Y}\frac{dY}{dX} = \frac{X}{Y}(\beta_1 \frac{1}{X}) = \beta_1 \frac{1}{Y}$

함수	형태	기울기($\frac{dY}{dX}$)	탄력도($\frac{X}{Y}\frac{dY}{dX}$)
선형(linear)	$Y_i = \beta_0 + \beta_1 X_i + u_i$	β_1	$\beta_1 \frac{X}{Y}$
전대수(log-log)	$lnY_i = ln\beta_0 + \beta_1 lnX_i + u_i$	$\beta_1 \frac{Y}{X}$	β_1
쌍곡선(reciprocal)	$Y_i = \beta_0 + \beta_1 \frac{1}{X_i} + u_i$	$-\beta_1 \frac{1}{X^2}$	$\beta_1 \frac{1}{XY}$
반로그(semi-log)	$lnY_i = \beta_0 + \beta_1 X_i + u_i$	$\beta_1 Y$	$\beta_1 X$
반로그(semi-log)	$Y_i = \beta_0 + \beta_1 lnX_i + u_i$	$\beta_1 \frac{1}{X}$	$\beta_1 \frac{1}{Y}$