

상관분석과 회귀분석

CE730

조남운

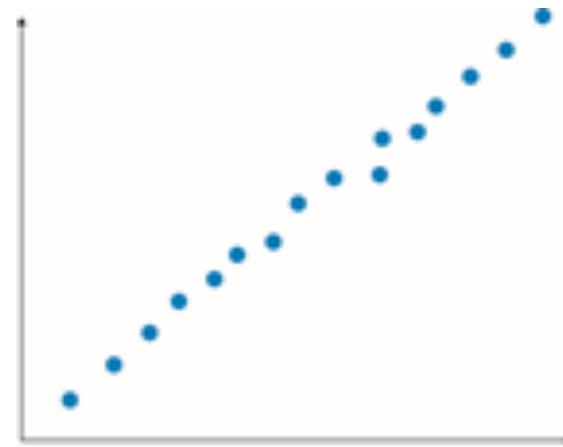
목차

- 상관분석
 - 상관계수
 - 표본상관계수의 검정
- 회귀분석
 - 단순회귀모형
 - 다중회귀모형
- 회귀분석의 응용
 - 시장모형
 - 자본자산가격결정모형

산점도 Scatter Plot

- 두 변수의 관계를 시각적으로 나타낸 것
- 각 변수가 하나의 축을 담당함
- 탐색적 자료분석에서 사용
 - 엄밀하게 분석하기 전 대략적인 변수간의 관계를 파악하는 분석

강한 양의 선형관계



Strong Linear Relationship

약한 양의 선형관계



Weak Linear Relationship



No Linear Relationship

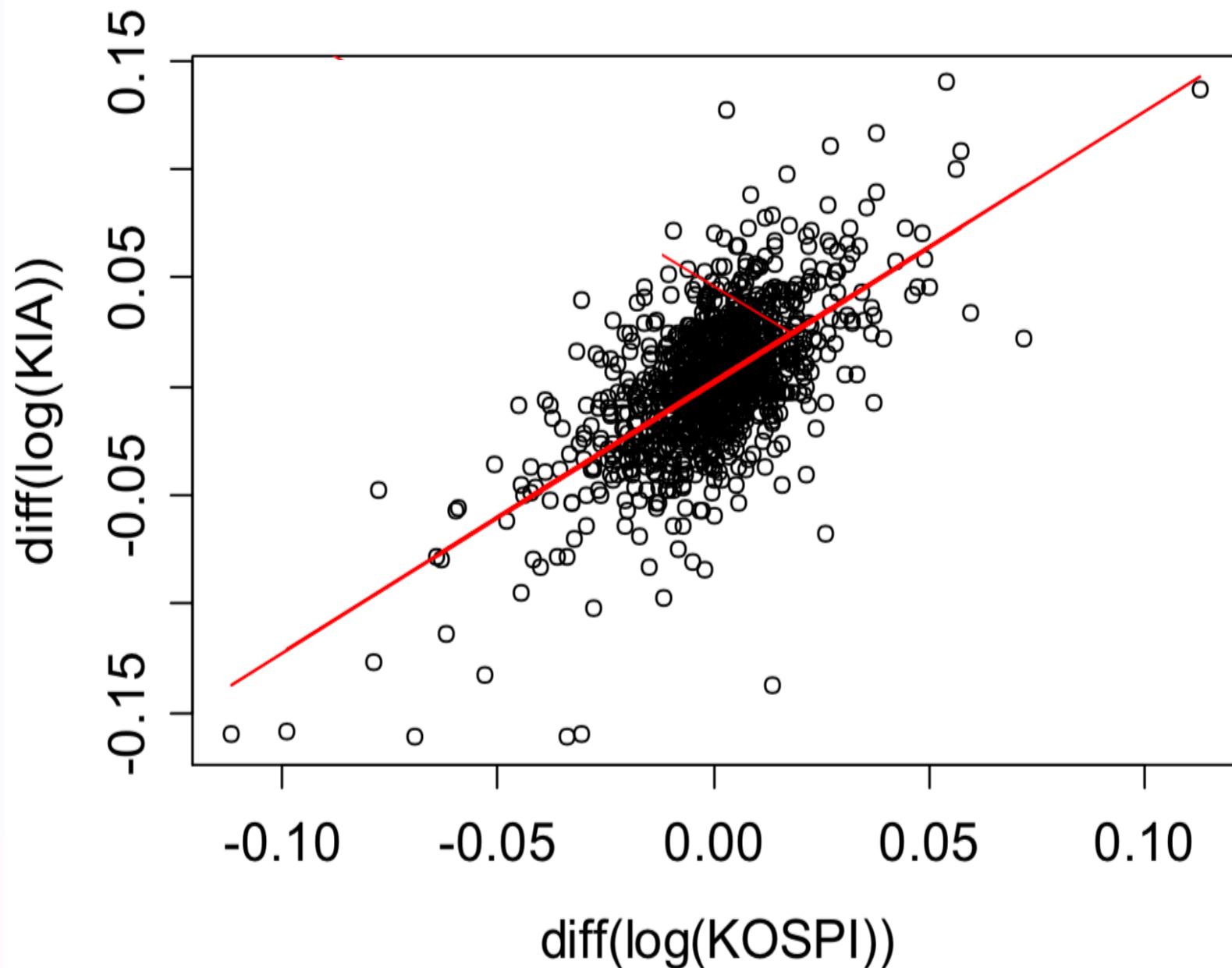
무관계



Nonlinear Relationship

비선형관계

기아자동차 주식과 KOSPI 주식의 산점도



상관계수

$$\rho_{XY} := \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}$$

모상관계수

- 엄밀한 분석을 위해서는 선형관계에 대한 정보를 객관적으로 표현할 수 있는 지표가 필요
 - 선형관계의 종류: 양/음
 - 선형관계의 강도: 강한 관계/약한 관계
- 상관계수는 이 조건을 충족하는 지표

$$r_{XY} := \frac{s_{XY}}{s_X s_Y}$$

표본상관계수

상관계수의 의미

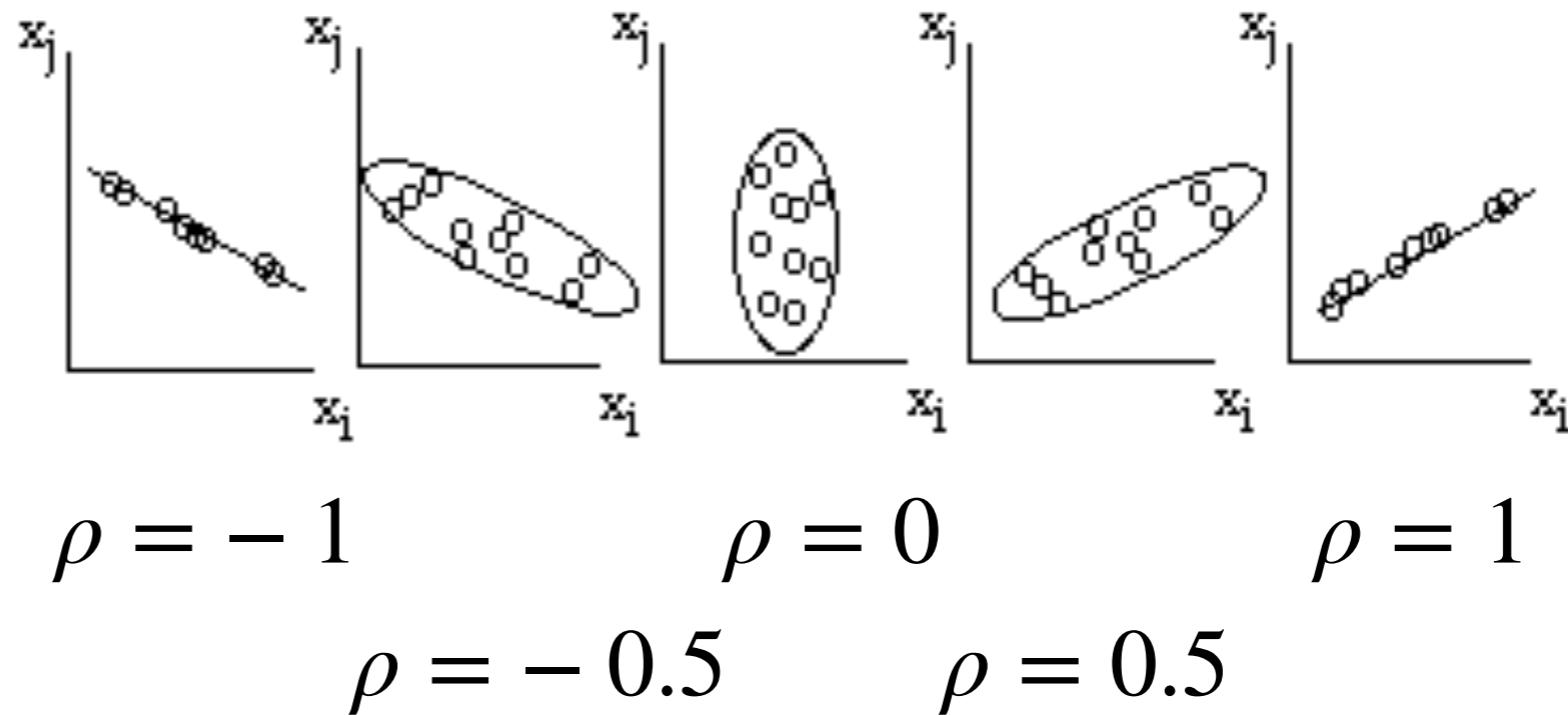
- 선형관계의 종류: 양/음
 - $\rho > 0 \Rightarrow$ 양의 상관관계
 - $\rho = 0 \Rightarrow$ 무관계
 - $\rho < 0 \Rightarrow$ 음의 상관관계
- 선형관계의 강도: 강한 관계/약한 관계
 - 절대치가 1에 가까울 수록 강한 관계
 - 절대치가 0에 가까울수록 약한 관계

상관계수

Correlation Coefficient

$$\rho_{XY} := \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

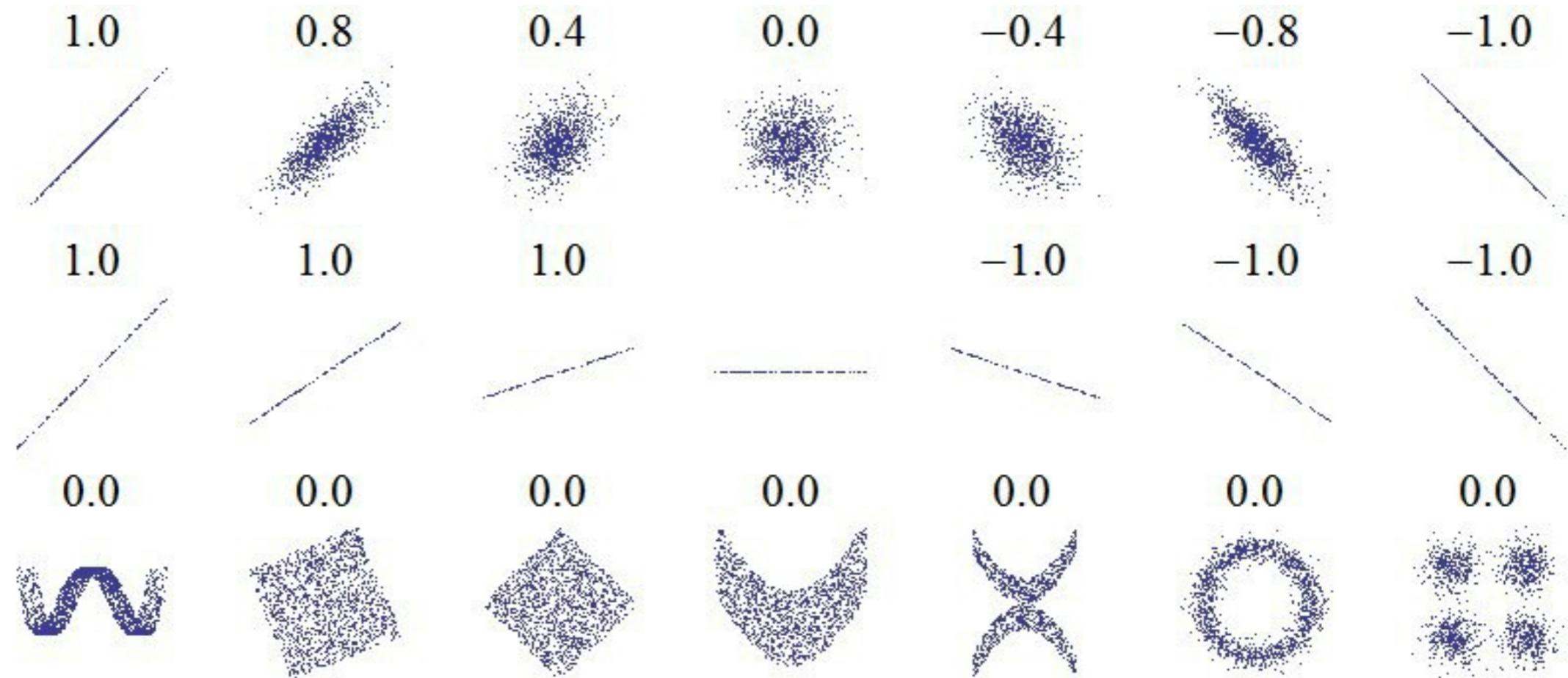
$$= \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$



- 공분산을 각 변수의 표준편차로 나누어 scale effect를 제거함

ρ 와 다양한 분포

- 주의: 기울기와는 무관함



<https://stats.stackexchange.com/questions/194636/is-it-correct-to-use-correlation-coefficient-in-this-case>

상관계수의 계산

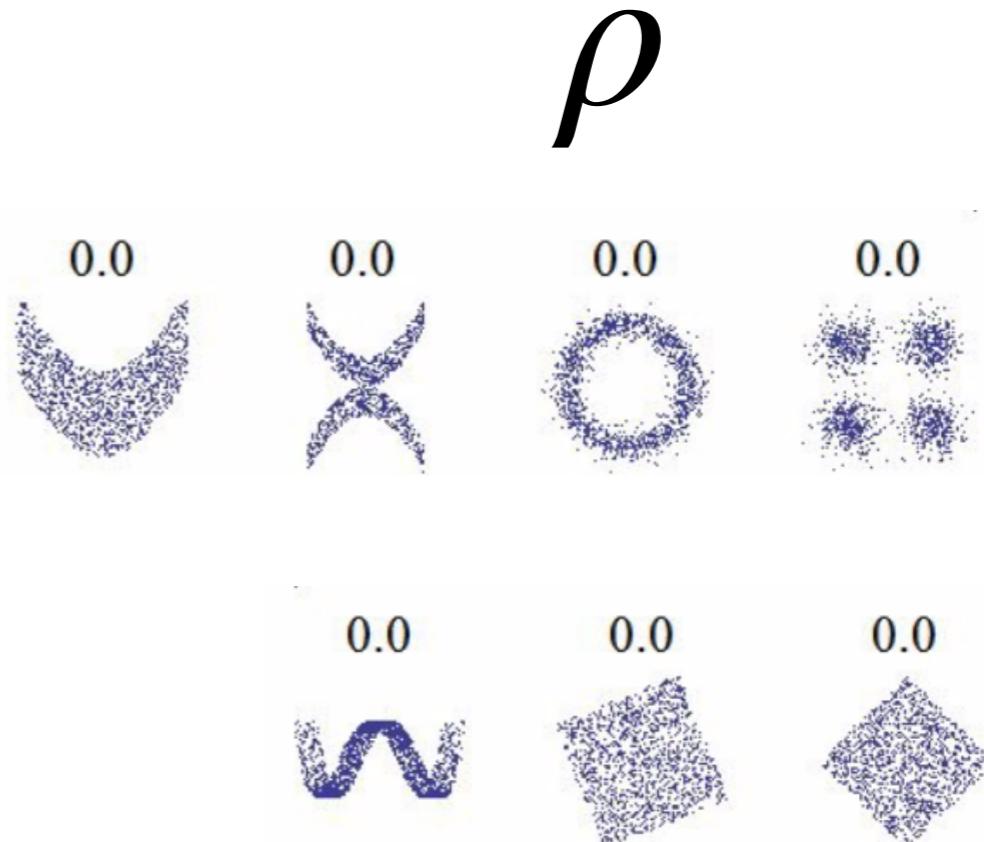
$$r_{XY} := \frac{s_{XY}}{s_X s_Y}$$

- 상관계수의 정의에 입각하여
(표본) 표준편차를 계산하면
됨
- Ex8,1

$$\rho_{XY} := \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

상관계수와 관련한 주의사항

- 상관계수는 선형관계만을 포착함
- 비선형관계는 매우 강하다 할지라도 포착하지 못함
- 상관계수 = 0 의 의미:
 - 관계가 없다 (X)
 - 선형관계가 아니다 (O)



표본상관계수의 검정

- 알고자 하는 것
 - 두 변수 X, Y 는 관계가 있는가?
- 전제조건
 - 두 모집단이 이변량정규분포를 따른다는 가정

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1 - \rho_{xy}^2}} \exp\left(\frac{-A}{2(1 - \rho_{xy}^2)}\right)$$

$$A := \frac{(x - \mu_x)^2}{\sigma_x^2} - \frac{2\rho_{xy}(x - \mu_x)(y - \mu_y)}{\sigma_x\sigma_y} + \frac{(y - \mu_y)^2}{\sigma_y^2}$$

표본상관계수 검정: 귀무가설, 대립가설

- 귀무가설
 - 두 변수는 상관관계가 없다
 \Rightarrow 상관계수=0
- 대립가설
 - 두 변수는 상관관계가 있다
 \Rightarrow 상관계수 $\neq 0$

$$H_0 : \rho_{XY} = 0$$

$$H_1 : \rho_{XY} \neq 0$$

$$t = \frac{r_{XY}\sqrt{n-2}}{\sqrt{1 - r_{XY}^2}} \sim t(df = n-2)$$

Ex 8.2

- 검정절차
 - 데이터로부터 통계량 계산 (H_0 기반)
 - 분포표상에서 critical value와 비교
 - 혹은 p-value와 유의수준 비교
 - 혹은 CI 산출
 - 기각역에 존재 $\Rightarrow H_0$ 기각
 - 그렇지 않을 경우 $\Rightarrow H_0$ 기각하지 않음

선형회귀분석

Linear Regression

Analysis

회귀분석

- 종속변수를 설명변수의 선형 결합으로 설명하고자 하는 분석법
- 관심있는 변수 (y) - 종속변수
- 종속변수의 크기를 설명하는 외생변수들
 - 설명변수

$$\mathbf{x} := \{x_1, x_2, \dots, x_k\}$$

$$\mathbf{x}_i := \{x_{1i}, x_{2i}, \dots, x_{ki}\}$$
$$i = 1, 2, \dots, n$$

n개 중 i번째 데이터

Methodology Overview

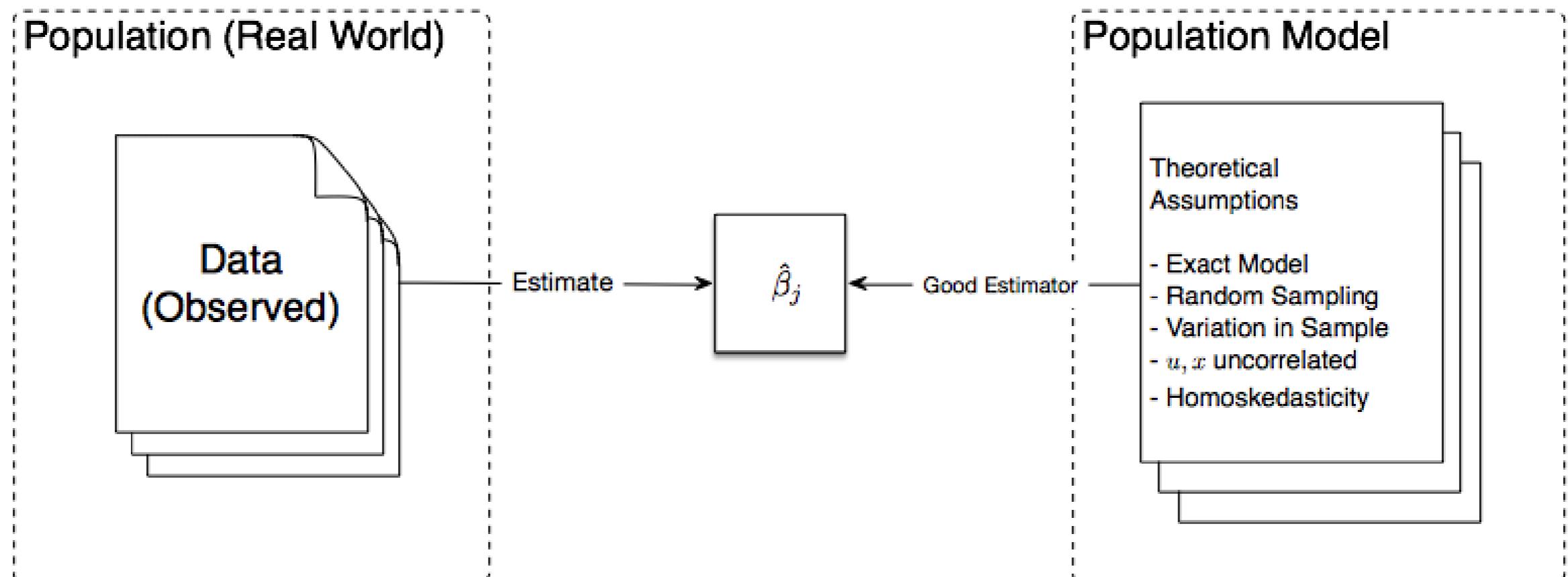


Figure : Big Picture of Estimation

선형회귀모형

- 설명되지 않는 부분의 존재를 오차항 (error term: ϵ)으로 표현
 - 인지되지 않은 설명변수, 관측오차 등으로 구성됨
- $i=1$ 인 경우: 단순선형회귀모형

$$y_i = \beta_0 + \boldsymbol{\beta} \mathbf{x}_i + \epsilon_i$$

$$= \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \epsilon_i$$

Predicted

Unpredicted

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

단순선형회귀모형

Simple Linear Regression (SLR) Model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

외생/내생변수, 독립/종속변수

- 선형회귀모형에서 y 는 설명되는 변수
 - 모형으로 설명되는 변수
 \Rightarrow 내생변수, 종속변수
 - x 는 설명하는 변수
 - 모형으로는 설명하지 못하는 변수
 \Rightarrow 외생변수, 독립변수

SLR 관련 용어들

TABLE 2.1 Terminology for Simple Regression

y	x
Dependent variable	Independent variable
Explained variable	Explanatory variable
Response variable	Control variable
Predicted variable	Predictor variable
Régressand	Regressor

기본 용어들

$\hat{\epsilon}_i$: residual

ϵ : error term

β_i : parameter

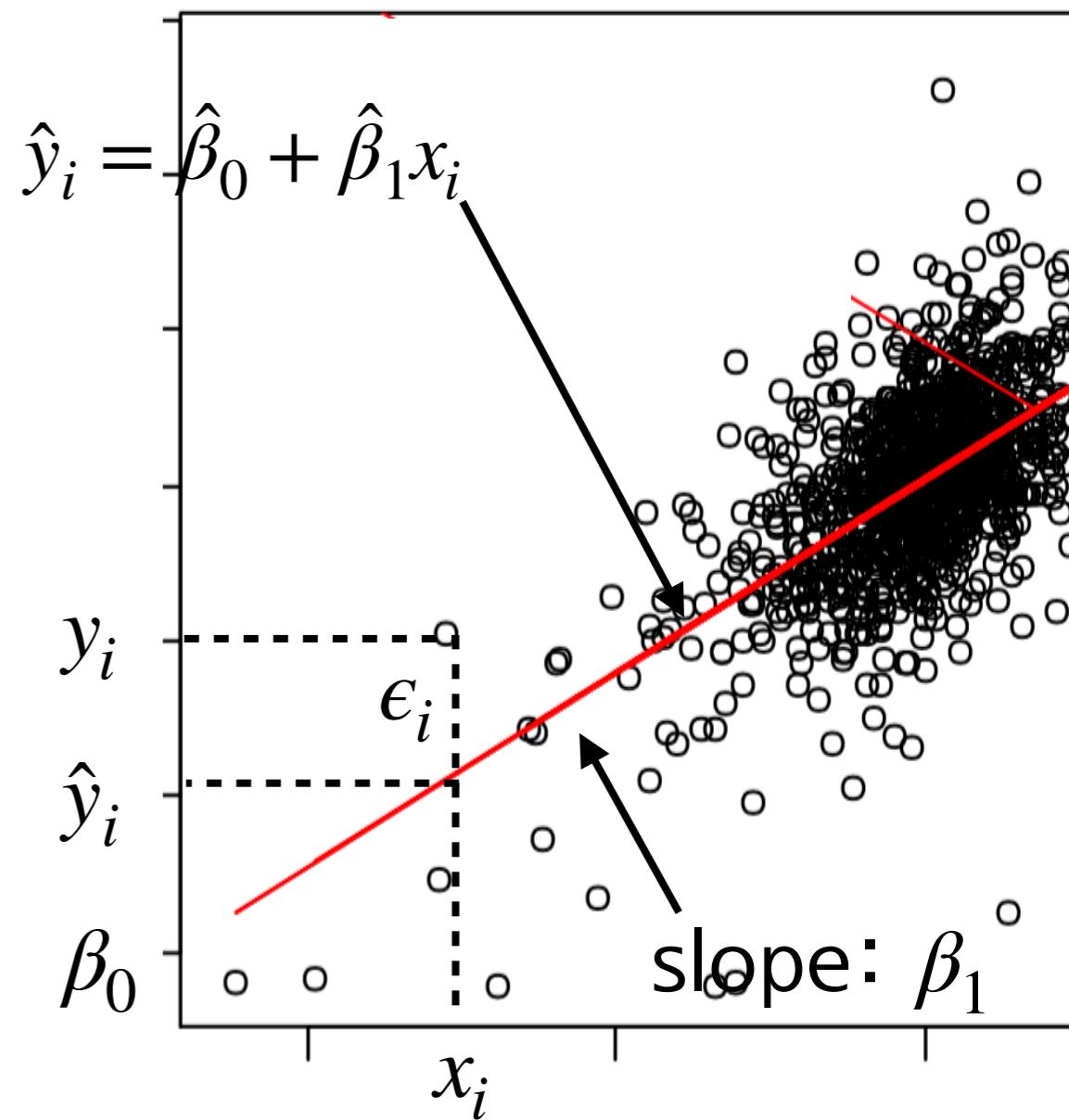
$\hat{\beta}_i$: estimate

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\epsilon}_i$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

hat은 추정량을 의미

hat이 없는 변수는 모수를 의미



회귀계수 β 의 추정

- 주어진 데이터를 “가장 잘” 설명하는 계수들(β)과 잔차(ϵ)를 찾는 문제
- “가장 잘”的 기준
 - 다양할 수 있음
 - 가장 많이 사용하는 것은 오차항의 제곱합(SSR)을 최소로 만드는 것
- Ordinary Least Squares (OLS)
 - LS에도 전제조건에 따라 다양한 버전 존재 (GLS 등)

$$\arg \min_{\beta_0, \beta_1} \sum_i^n \epsilon_i^2$$

$$\arg \min_{\beta_0, \beta_1} \sum_i^n (y_i - \beta_0 - \beta_1 x_i)^2$$

SSR: Sum of Squared Residuals

OLS: SLR case

- FOC:

β_0, β_1 로 편미분한 값이 0

- 그렇게 만드는 β 값:

$$\hat{\beta}_0, \hat{\beta}_1$$

- SOC:

한 번 더 편미분한 매트릭스
(Hessian)가 양정부호(PD:
Positive Definite)

$$\arg \min_{\beta_0, \beta_1} \sum_i^n (y_i - \beta_0 - \beta_1 x_i)^2$$

$$SSR := \sum_i^n (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\frac{\partial SSR}{\partial \beta_0} = -2 \sum_i^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\frac{\partial SSR}{\partial \beta_1} = -2 \sum_i^n x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

FOC (계속)

$$\frac{\partial SSR}{\partial \beta_0} = -2 \sum_i^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$-\sum_i^n \hat{\beta}_0 + \sum_i^n (y_i - \hat{\beta}_1 x_i) = 0$$

$$\hat{\beta}_0 = \frac{1}{n} \sum_i^n (y_i - \hat{\beta}_1 x_i) = \bar{y} - \hat{\beta}_1 \bar{x}$$

FOC (계속)

$$\frac{\partial SSR}{\partial \beta_1} = -2 \sum_i^n x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\sum_i^n x_i(y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_i) = 0$$

$$\sum_i^n [x_i(y_i - \bar{y})] - \hat{\beta}_1 \sum_i^n [x_i(x_i - \bar{x})] = 0$$

$$\hat{\beta}_1 = \frac{\sum_i^n [x_i(y_i - \bar{y})]}{\sum_i^n [x_i(x_i - \bar{x})]}$$

$$\hat{\beta}_1 = \frac{\sum_i^n [(x_i - \bar{x} + \bar{x})(y_i - \bar{y})]}{\sum_i^n [(x_i - \bar{x} + \bar{x})(x_i - \bar{x})]}$$

FOC (계속)

$$\hat{\beta}_1 = \frac{\sum_i^n [(x_i - \bar{x} + \bar{x})(y_i - \bar{y})]}{\sum_i^n [(x_i - \bar{x} + \bar{x})(x_i - \bar{x})]}$$

$$\hat{\beta}_1 = \frac{\sum_i^n [(x_i - \bar{x})(y_i - \bar{y}) + \cancel{\bar{x}(y_i - \bar{y})}]}{\sum_i^n [(x_i - \bar{x})(x_i - \bar{x}) + \cancel{\bar{x}(x_i - \bar{x})}]}$$

$$\bar{x} := \frac{1}{n} \sum_i^n (x_i - \bar{x} + \bar{x}) = \frac{1}{n} \sum_i^n (x_i - \bar{x}) + \bar{x} \Rightarrow \frac{1}{n} \sum_i^n (x_i - \bar{x}) = 0$$

$$\therefore \hat{\beta}_1 = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_i^n (x_i - \bar{x})^2}$$

SLR: SOC

- 2×2 matrix 가 양정부호 (PD)이기 위해서는 오른쪽에 기술된 두 행렬식(det)이 모두 0보다 커야 함
- 데이터가 모두 양수인 경우 SOC 통과
- 따라서 FOC를 만족하는 $\hat{\beta}_0, \hat{\beta}_1$ 는 OLS조건을 충족 하는 유일한 값임

$$D_{\beta_0, \beta_1}^2 E = 2 \begin{pmatrix} n & \sum_i^n x_i \\ \sum_i^n x_i & \sum_i^n x_i^2 \end{pmatrix}$$

$$\text{Det}(n) > 0$$

$$\text{Det} \begin{pmatrix} n & \sum_i^n x_i \\ \sum_i^n x_i & \sum_i^n x_i^2 \end{pmatrix} > 0$$

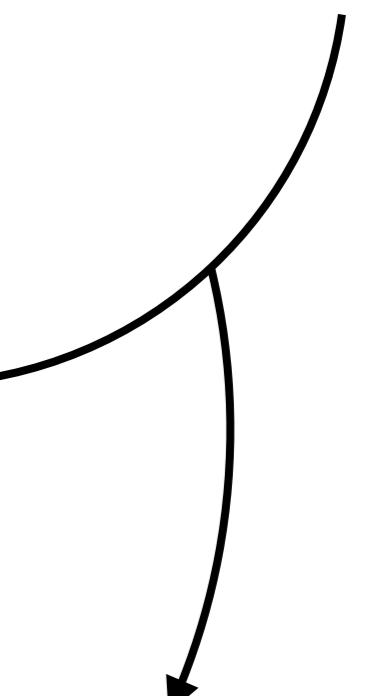
OLS for SLR: 결론

$$\hat{\beta}_1 = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_i^n (x_i - \bar{x})^2} = \frac{s_{XY}}{s_X^2}$$

- β_0 의 추정량은 설명변수값이 0일 때의 종속변수값을 추정
- β_1 의 추정량은 설명변수값 1 증가에 따른 종속변수의 평균적인 변화량을 추정

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$r_{XY} = \frac{s_{XY}}{\sqrt{s_X^2 s_Y^2}}$$


$$\hat{\beta}_1 = r_{XY} \sqrt{\frac{s_Y^2}{s_X^2}}$$

다중회귀모형

Multivariate Linear Regression (MLR) Model

선형대수(행렬) 기초

Matrix

Definition (Matrix)

Matrix is a rectangular array of numbers (scalars)

Let $a_{ij} \in \mathbb{R}$ or $A_{ij} \in \mathbb{R}$ be the i th row and j th column element of matrix A

Definition (Equal)

$$A = B \iff \begin{cases} \text{same size} \\ a_{ij} = b_{ij} \quad \forall i, j \end{cases}$$

Addition, Subtraction

Let A, B be $n \times k$ matrices and $r \in \mathbb{R}$

Definition (Addition)

$$(A + B)_{ij} := a_{ij} + b_{ij} \quad \forall i, j$$

Important note: the first $+$ and the second $+$ are not same operators

Definition (Subtraction)

$$(A - B)_{ij} := a_{ij} - b_{ij} \quad \forall i, j$$

Multiplications of Matrices

Definition (Scalar Multiplication)

$$(rA)_{ij} := rA_{ij} \quad \forall i, j$$

Let A be $n \times k$ matrix and B be $k \times m$ matrix. Then AB is $n \times m$ matrix.

Definition (Matrix Multiplication)

$$(AB)_{ij} := A_{i1}B_{1j} + A_{i2}B_{2j} + \cdots + A_{ik}B_{kj} = \sum_{r=1}^k A_{ir}B_{rj}$$

For $n \times n$ matrices, identity matrix I_n is a multiplicative identity.

$$AI = IA = A$$

Laws of Matrix Algebra

Laws of Matrix Algebra

$$(A + B) + C = A + (B + C) \quad (\text{Associative Law for Addition})$$

$$(AB)C = A(BC) \quad (\text{Associative Law for Multiplication})$$

$$A + B = B + A \quad (\text{Commutative Law for Addition})$$

$$A(B + C) = AB + AC \quad (\text{Distributive Law})$$

$$(A + B)C = AC + BC \quad (\text{Distributive Law})$$

Important Note: $AB \neq BA$

Transpose

Definition (Transpose)

$A^\top (n \times m)$ is a transpose of $A (m \times n)$ if:

$$(A^\top)_{ij} := A_{ji} \quad \forall i, j$$

Some researchers denote X^T by X'

$$(A \pm B)^\top = A^\top \pm B^\top$$

$$(A^\top)^\top = A$$

$$(rA)^\top = rA^\top$$

$$(AB)^\top = B^\top A^\top$$

(Theorem 8.1)

MLR: 기본 용어들

TABLE 3.1 Terminology for Multiple Regression

y	x_1, x_2, \dots, x_k
Dependent variable	Independent variables
Explained variable	Explanatory variables
Response variable	Control variables
Predicted variable	Predictor variables
Regressand	Regressors

Deriving $\hat{\beta}$ through Matrix Algebra

- Matrix form is much more simple and intuitive.
- Basic understanding of linear algebra is needed (Appendix D)

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + u_i \quad (\text{i-th observation})$$

$$y_i = (1 \quad x_{i1} \quad \cdots \quad x_{ik}) \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + u_i = \mathbf{x}_i \boldsymbol{\beta} + u_i \quad (\text{Matrix form})$$

- Aggregating all observations $i = 1, 2, \dots, n$ and define \mathbf{X} as:

$$\mathbf{X} := \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$$

Matrix Notation for all n observations

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \quad (\text{E3})$$

$$\mathbf{y} := \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{u} := \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

- All matrices above are not random matrix (*i.e.*, not matrix version of random variable, but ordinary matrix) if data is given.

다중회귀모형의 행렬표현

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & x_{N2} & \cdots & x_{Nm} \end{pmatrix} \quad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}$$

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{y} \quad \text{OLS estimator}$$

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (X'X)^{-1}X'(X\boldsymbol{\beta} + \boldsymbol{\epsilon}) = (X'X)^{-1}X'X\boldsymbol{\beta} + (X'X)^{-1}X'\boldsymbol{\epsilon} \\ &= \boldsymbol{\beta} + (X'X)^{-1}X'\boldsymbol{\epsilon} \end{aligned}$$

OLS estimate $\hat{\beta}$: Matrix Representation

SLR과 MLR의 결과는 서로 통하는 곳이 있음

$$\therefore \hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

- Similar structure with SLR

$$\sum_i (x_i - \bar{x})^2 > 0 \Rightarrow \hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

- $\sum_i (x_i - \bar{x})^2 \leftrightarrow \mathbf{X}'\mathbf{X}$
- $\sum_i (x_i - \bar{x})(y_i - \bar{y}) \leftrightarrow \mathbf{X}'\mathbf{y}$
- In fact, matrix representation of SSR is:

$$SSR := \sum_i^n \hat{u}_i^2 = \hat{\mathbf{u}}'\hat{\mathbf{u}} = (\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta})$$

변수들의 의미

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$
$$y = \mathbf{x}\boldsymbol{\beta} + u$$

- β_0 : intercept
- $\beta_j, j = 1, 2, \dots, k$: slope parameter associated with x_j
- u : error term

Graphical Expression of OLS estimating in Multiple Linear Regression

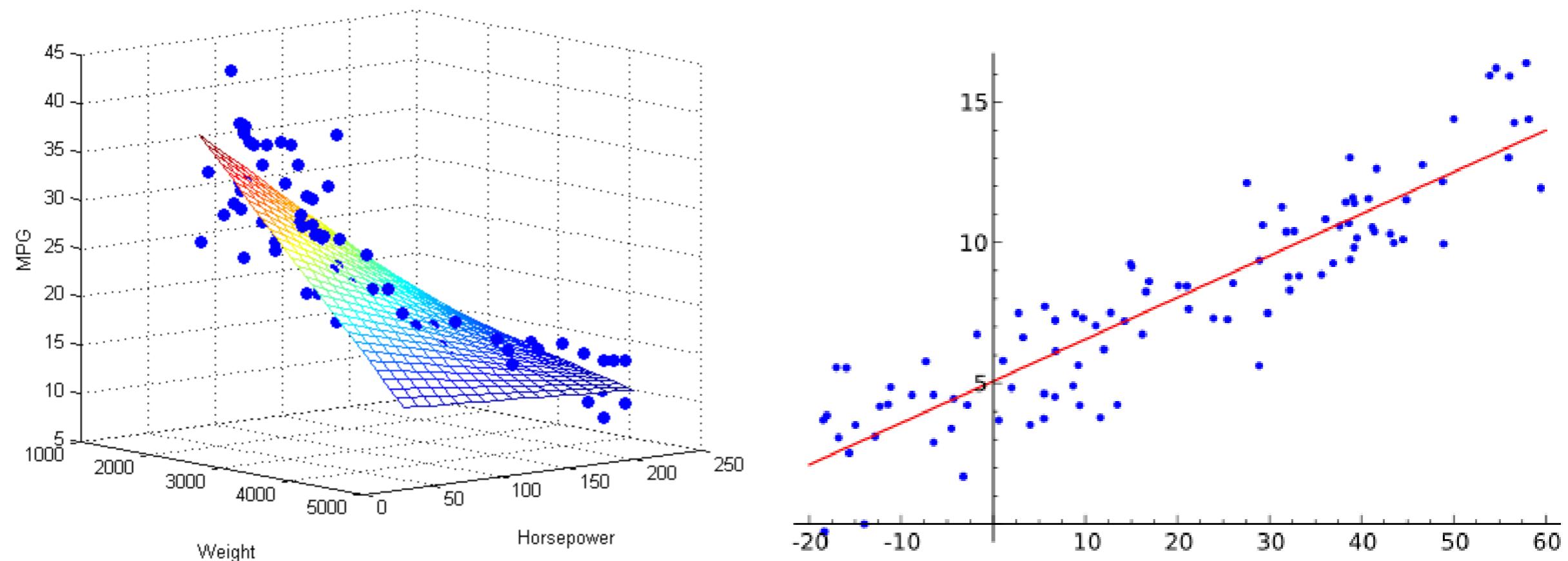


Figure : Estimating $MPG = \hat{\beta}_0 + \hat{\beta}_1 Weight + \hat{\beta}_2 Horseposer$

모수 β 추정에 관련된 필수 가정들

- (A1) 다중공선성이 없어야 한다
(설명변수 사이에 선형 관계가 없어야 한다)
- (A2) $X'X/n$ 이 n 의 증가에 따라 상수행렬로 수렴해야 한다
(X 가 충분히 잘 퍼져 있어 y 에 대한 설명을 제공할 수 있어야 한다)
- (A3) $E(\varepsilon) = 0$

모수 β 추정에 관련된 가정들

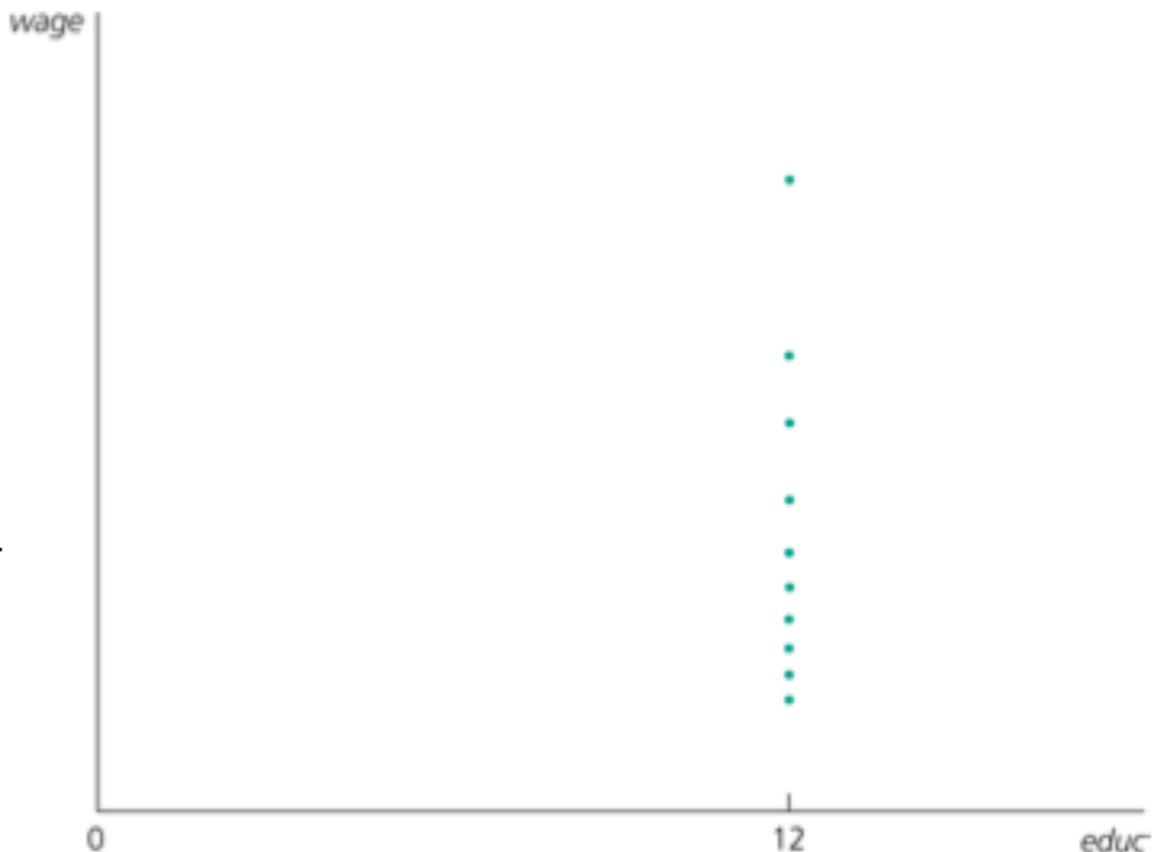
- (A4) X 는 확률변수가 아니라 상수이다
혹은 X 가 확률변수라도 $\text{Cov}(X, \varepsilon) = 0$
(X 가 관측값일 경우 자동적으로 충족됨)
- (A5) ε_i 는 서로 독립 & $\text{Var}(\varepsilon_i) = \sigma^2 < \infty$

A1: No Perfect Multicollinearity

- 설명변수들 중 일부가 서로 완전한 선형관계일 경우, $X'X$ 의 역행렬이 존재하지 않음
- 이는 OLS의 해가 유일하지 않음을 의미
- 해법: 완전한 선형관계가 존재한다면, 관계 변수 중 하나를 제거함으로써 문제 해결 가능

$$A2: \mathbf{X}'\mathbf{X}/n < \infty$$

- x_i 가 상수, 즉 모두 같은 값
이어서는 안됨
 - 이러한 경우 상수인 x_i 는
 y 에 대한 설명력이 없음



A3: $E(\varepsilon)=0$

- 불편성(Unbiasedness)을 충족하기 위한 필수조건
 - 가장 중요한 조건
 - 이것을 만족하지 못할 경우 편의(bias)가 발생함

A5: 잔차간 독립(A5-1), $\text{Var}(\varepsilon_i) = \sigma^2$ 동분산(Homoskedasticity) (A5-2)

- A5를 위반할 경우 β 의 추정량은 의사회귀(spurious regression)가 되어 잘못된 추정을 하게 됨

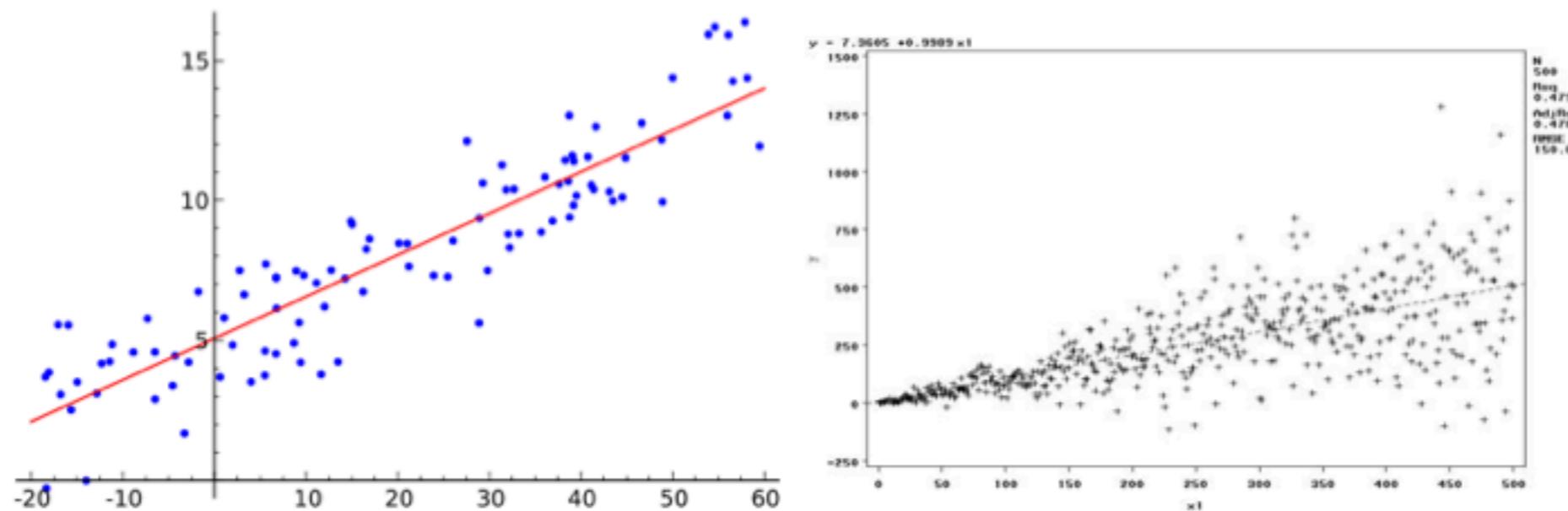


Figure : Left:homoskedasticity, right:heteroskedasticity

OLS 추정량 β 의 성질

행렬로 표현된 확률변수의 연산법칙

- 문자표현관행
- 두꺼운 소문자: 벡터 (vector)
- 대문자: 행렬 (matrix)
- 프라임(') :
Matrix Transpose

M1

$$E(A\mathbf{y}) = AE(\mathbf{y})$$

M2

$$\text{Var}(A\mathbf{y}) = A\text{Var}(\mathbf{y})A'$$

OLS 추정량 $\hat{\beta}$ 의 기대치

Regression Model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

OLS estimate

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

$$E(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{y}) \quad \text{A4, M1}$$

$$= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})$$

OLS estimate

$$= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\boldsymbol{\epsilon})$$

$$= \boxed{I}$$

$$= 0$$

$$A4$$

$$= \boldsymbol{\beta}$$

A3

- A1-A5를 만족할 경우 OLS 추정량 $\hat{\boldsymbol{\beta}}$ 는 모수 $\boldsymbol{\beta}$ 에 대한 불편추정량

OLS 추정량 hat β 의 분산

Regression Model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

OLS estimate

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}\end{aligned}$$

$$Var(\hat{\boldsymbol{\beta}}) = Var(\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon})$$

$$= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Var(\boldsymbol{\epsilon})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \quad \text{M2}$$

$$= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \quad \text{A5}$$

$$= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

$$Var(\hat{\boldsymbol{\beta}}) = \begin{pmatrix} Var(\hat{\beta}_0) & Cov(\hat{\beta}_0, \hat{\beta}_1) \\ Cov(\hat{\beta}_0, \hat{\beta}_1) & Var(\hat{\beta}_1) \end{pmatrix}$$

Var-Cov Matrix
(SLR)

$(X'X)^{-1}$ in SLR

$$(X'X)^{-1} = \frac{1}{n \sum (x_i - \bar{x})^2} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ \sum x_i & n \end{pmatrix}$$

$$\begin{aligned} Var(\hat{\beta}) &= \sigma^2 (X'X)^{-1} = \begin{pmatrix} Var(\hat{\beta}_0) & Cov(\hat{\beta}_0, \hat{\beta}_1) \\ Cov(\hat{\beta}_0, \hat{\beta}_1) & Var(\hat{\beta}_1) \end{pmatrix} \\ &= \frac{\sigma^2}{n \sum (x_i - \bar{x})^2} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ \sum x_i & n \end{pmatrix} \end{aligned}$$

$$Var(\hat{\beta}) = \sigma^2(X'X)^{-1} = \begin{pmatrix} Var(\hat{\beta}_0) & Cov(\hat{\beta}_0, \hat{\beta}_1) \\ Cov(\hat{\beta}_0, \hat{\beta}_1) & Var(\hat{\beta}_1) \end{pmatrix}$$

$$= \frac{\sigma^2}{n \sum (x_i - \bar{x})^2} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ \sum x_i & n \end{pmatrix}$$

$$Var(\hat{\beta}_0) = \frac{\sum x_i^2}{n \sum (x_i - \bar{x})^2} \sigma^2$$

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\sum x_i}{n \sum (x_i - \bar{x})^2} \sigma^2 = \frac{-\bar{x}}{\sum (x_i - \bar{x})^2} \sigma^2$$

자주 쓰는 수식을 단순화하기 위함

$$y_i = \hat{y}_i + \hat{u}_i \quad (2.32)$$

- y_i : measured (from one sample)
- \hat{y}_i : predicted part of y_i : calculated from x_i and OLS estimates (from all sample data)
- \hat{u}_i : unpredicted part of y_i

Definition (Total Sum of Squares (**TSS**, Explained Sum of Squares (**RSS**), Sum of Squared Residuals (**SSE** 2.20))

$$\text{TSS} := \sum_i (y_i - \bar{y})^2 \quad \text{RSS} := \sum_i (\hat{y}_i - \bar{y})^2 \quad \text{SSE} := \sum_i \hat{u}_i^2 \quad (2.33 - 2.35)$$

회귀계수 $\hat{\beta}_j$ 의 의미

Regression Model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$$

오차항
error term

Estimated Model

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik} + \hat{\epsilon}_i$$

잔차
residual

- x_j 를 제외한 모든 다른 변수 $x(-j)$ 들이 모두 고정되어 있다는 전제 하에 x_j 가 1 증가할 때 y 값이 증가하는 양

회귀계수 $\hat{\beta}_j$ 의 의미: Examples

$$\widehat{KOSDAQ}_i = 16.7979 + 0.0272 * NASDAQ_i.$$

- 나스닥지수 1 증가시 코스닥은 0.0272 증가

$$\widehat{\text{소비}} = \hat{\beta}_0 + \hat{\beta}_1 \text{소득} + \hat{\beta}_2 \text{학력}$$

- 학력이 고정된 상태에서 소득 1 증가는 소비를 $\hat{\beta}_1$ 만큼 증가시킴

회귀계수 β_j 검정: 모분산을 아는 경우

$X^{(jj)} := (j, j)th$ element of $(X'X)^{-1}$

- 추가적 가정 필요 (A6)
 - 오차항은 정규분포를 따름
 - 평균은 A4, 분산은 A5에 의해 규정됨
 - $\hat{\beta}_j$ 의 분포를 알고 나머지 변수들은 모두 알 수 있는 값이므로 β_j 에 대한 검정이 가능함

$$\epsilon_i \sim iidN(0, \sigma^2)$$

$$\hat{\beta} \sim N(\beta, (X'X)^{-1}\sigma^2)$$

$$\hat{\beta}_j \sim N(\beta_j, X^{(jj)}\sigma^2)$$

$$H_0 : \beta_j = \beta_{0j}$$

$$H_a : \beta_j \neq \beta_{0j}$$

$$\frac{\hat{\beta}_j - \beta_{0j}}{\sqrt{X^{(jj)}\sigma^2}} \sim N(0, 1)$$

0인 경우
가 대부분

회귀계수 β_j 검정: 모분산을 모르는 경우

$$\hat{\sigma}^2 = s_Y^2 = \frac{1}{n - (k + 1)} \sum_i^n (y_i - \hat{y}_i^2) = \frac{\text{SSE}}{n - (k + 1)}$$

- 대부분의 경우 모분산을 알지 못 함 \Rightarrow 추정해야 함
- 모분산의 추정량은 y 의 표본분산
 - 이를 위해 $k+1$ 개의 $\hat{\beta}_j$
 $(j=0, 1, 2, \dots, k)$ 자유도가 사용됨
- Good news: ϵ_i 가 표준정규분포를 따르지 않아도 CLT에 의해 표본수가 많으면 검정통계량은 표준정규분포를 따름 \Rightarrow t검정 가능

$$\text{SSE} := \sum_i^n \epsilon_i^2 \sim \chi^2(df = n)$$

$$\frac{\hat{\beta}_j - \beta_{0j}}{\sqrt{X^{(jj)}\hat{\sigma}^2}} \sim t(df = n - (k + 1))$$

$SE(\hat{\beta}_j)$

$$\frac{\hat{\beta}_j - \beta_{0j}}{\sqrt{X^{(jj)}\sigma^2}} \xrightarrow{d} N(0, 1)$$

$SD(\hat{\beta}_j)$

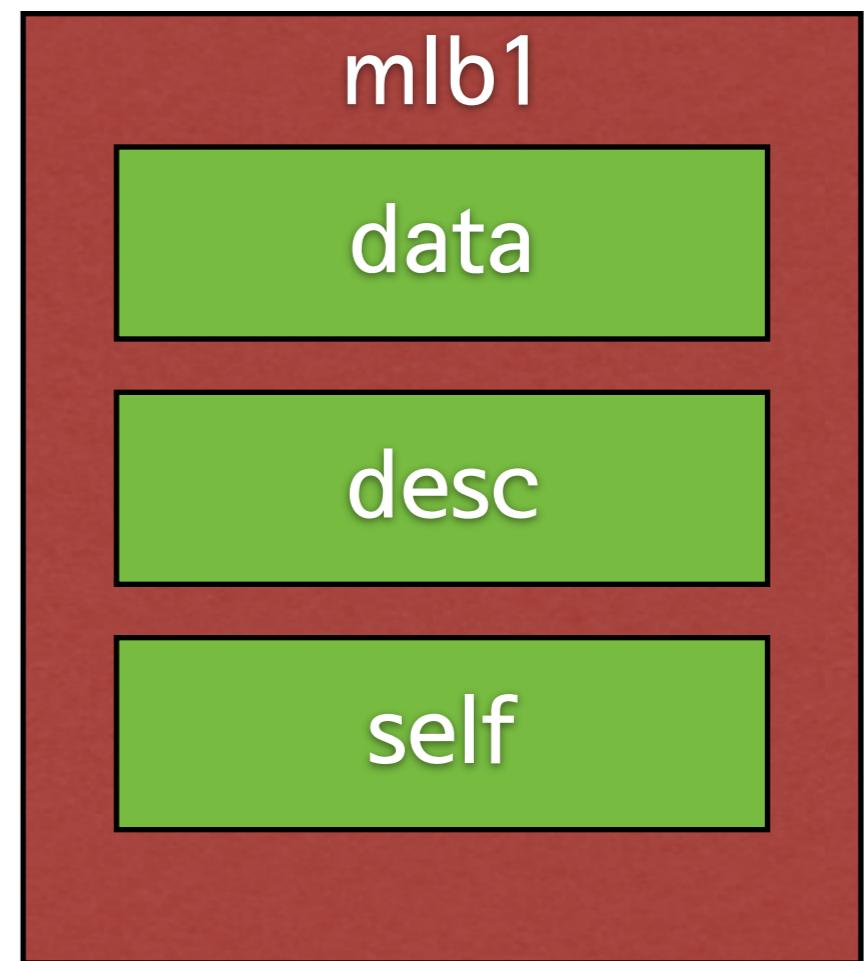
선형회귀분석 연습

Exercise for import data

- target: mlb1.Rdata
- load("FOLDER/LOCATION/mlb1.Rdata")
- ls() #list all user defined objects in current environment
- print(data) # print the content of table data
- head(data) # print first 6 obs of table data
- tail(data) # print last 6 obs of table data

Structure of mlb1

- dataset mlb1 consists of three tables:
 - data
 - desc
 - self



ceosal1.RData

$$salary = \beta_0 + \beta_1 roe + u.$$

- setwd("FOLDER/FOR/YOURDATA")
- load("ceosal1.RData")
- model<-lm(salary~roe, data=data)
- summary(model)
- plot(salary~roe, data=data)
- abline(model)

$$\widehat{salary} = 963.191 + 18.501 roe$$

[2.26]

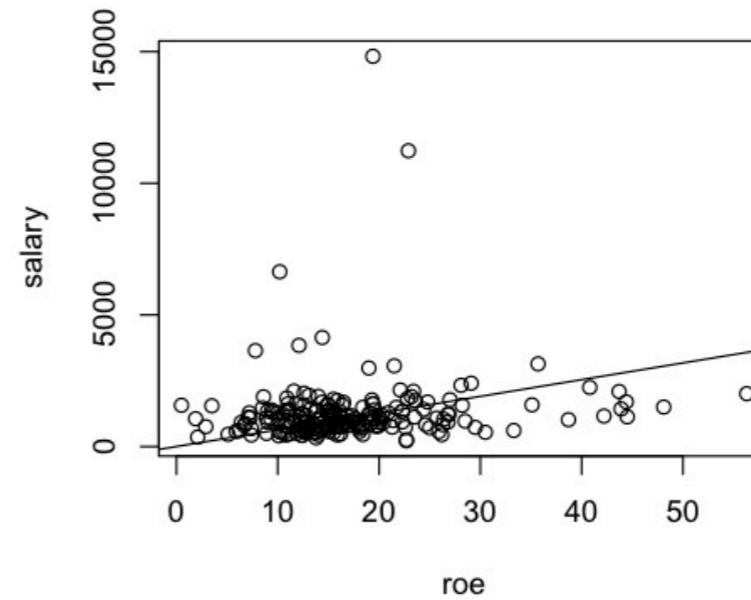
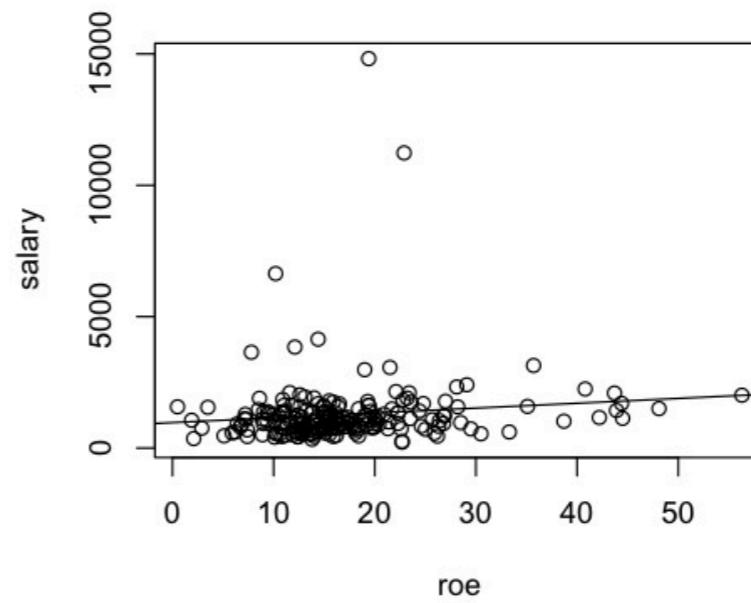
$n = 209,$

Manual Calculation

- $\text{beta1} = \text{cov}(x, y) / \text{var}(x)$
- $\text{beta1} <- \text{cov}(\text{data}\$roe, \text{data}\$salary) / \text{var}(\text{data}\$roe)$
- print(beta1)
- $\text{df}=207 = 209 - (1+1)$

Model with no Constant

- SLR for fixing $\beta_0=0$ (no y intercept)
 - `model0<-lm(salary~roe -1, data=data)`
 - `plot(salary~roe -1, data=data)`
 - `abline(model0)`
 - df: 207 → 208



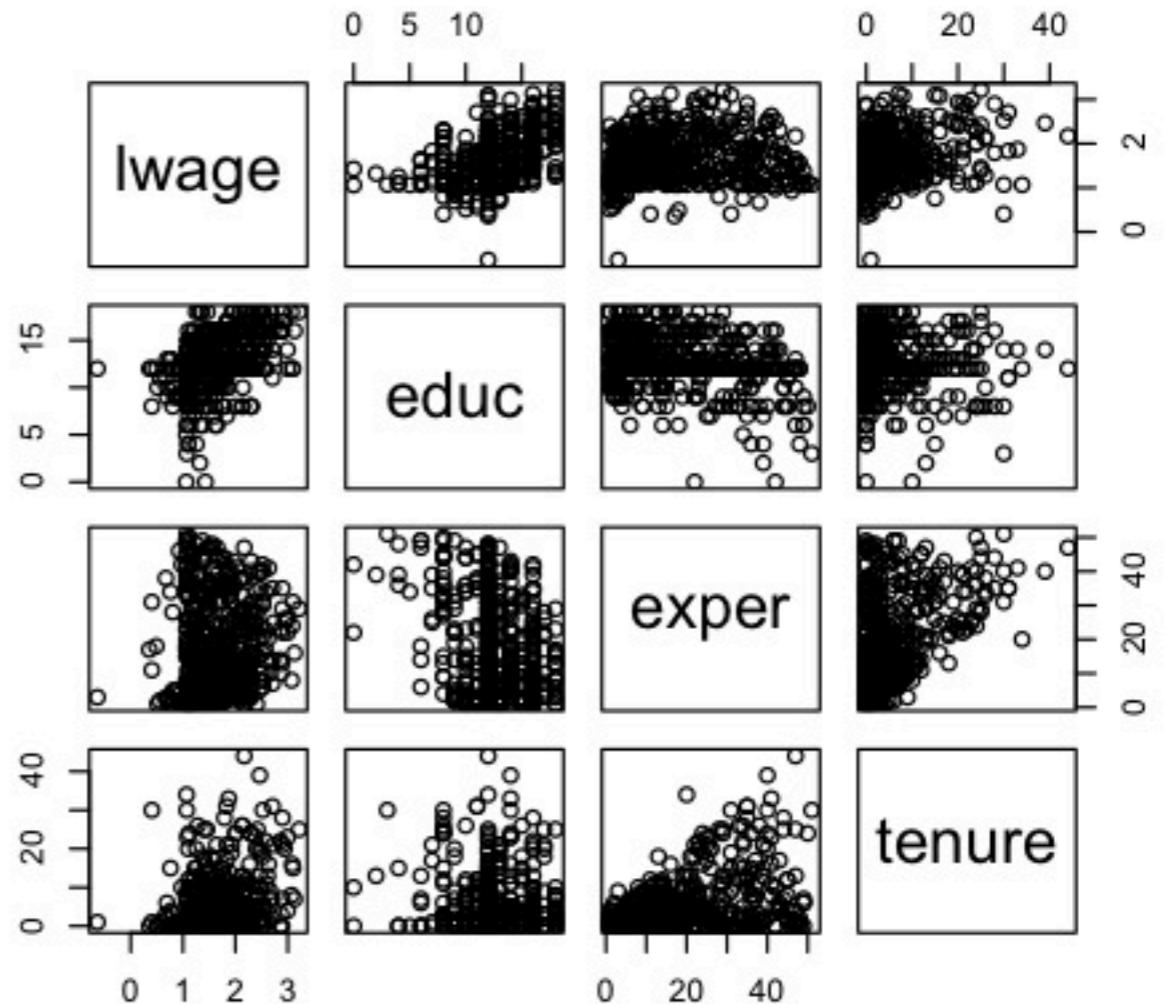
Multiple Regression

$$\widehat{\log(wage)} = .284 + .092 \text{ educ} + .0041 \text{ exper} + .022 \text{ tenure}$$
$$(.104) \quad (.007) \quad (.0017) \quad (.003)$$
$$n = 526, R^2 = .316,$$

- Example 4.1
- rm(list=ls())
- load("wage1.RData")
- model <-
lm(log(wage)~educ+exper+tenure,data=data)
- summary(model)

Scatter plots

- `print(desc)`
- `pairs(data[c(22,2,3,4)])`



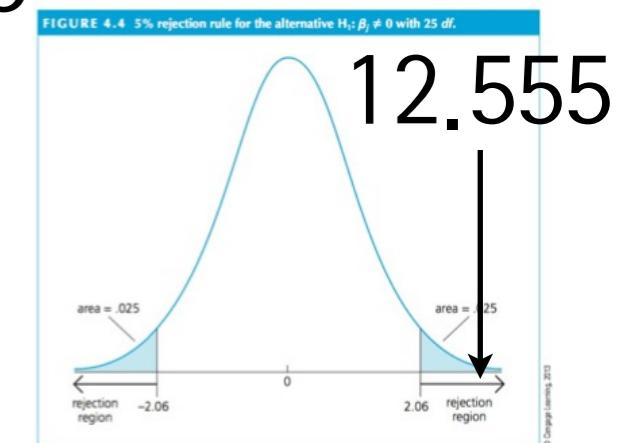
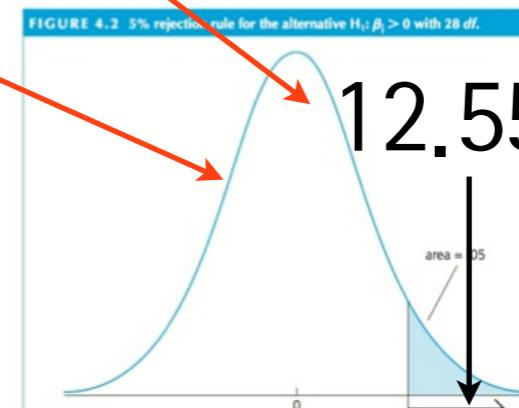
Inference: t-test

- If you can get t-statistic and its distribution for your H_0 , then you can test H_0
- $H_0: \text{exper}=0$
- $H_1: \text{exper} \neq 0$ (already displayed
`summary(model)` ==> two tail test)
- $H_1': \text{exper}>0$ ==> one tail test
 - Rejection area expands

$H_0: \text{exper}=0$

- try calculate t-statistics manually
 - $0.092029/0.007330 \approx 12.555 \sim t_{522}$
 - p-value for t distribution for df 522 = 526-(3+1)
 - n=526
 - k=3
 - critical value for significance level 0.95 (one-tail)
 - $\text{qt}(0.95, \text{df}=522)=1.647778$
 - for two-tail: $\text{qt}(0.975, \text{df}=522)=1.964519$
 - We can reject H_0 in both H_1

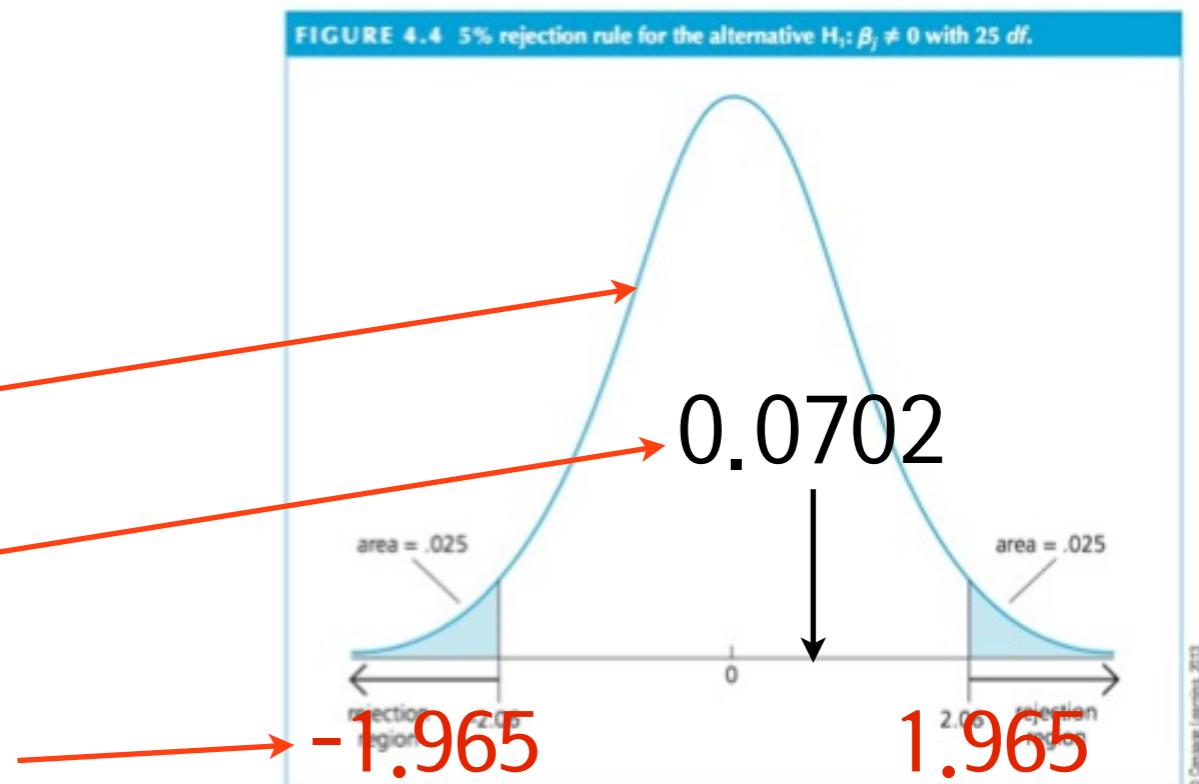
$$t_{\hat{\beta}_j} := \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$



$$H_0: \beta_j = a_j$$

$$H_1: \text{not } H_0$$

- $H_0: \text{exper}=0.004$
- Regress
- get proper t-statistics for H_0
 - $t=(\hat{\text{exper}} - 0.004)/\text{se}(\hat{\text{exper}}) \sim t_{522}$
 - $(0.004121-0.004)/0.001723 = 0.07022635$
 - $\text{qt}(0.975, df=522)=1.964519 > t$
statistics
 - We cannot reject above H_0



qt versus pt

- **pt(t-statistics, df)**

- returns 1- (p-value)
- ex) $\text{pt}(1.965, \text{df}=522) \rightarrow 0.975$
- ex2) $1-\text{pt}(0.0702, \text{df}=522) \rightarrow 0.472$
- ex3)
 $\text{pt}(0.0702, \text{df}=522, \text{lower.tail=FALSE}) \rightarrow 0.472$
- p-value for previous t-statistic is $0.472 > 0.05 \Rightarrow \text{reject } H_0$

- **qt(probability, df)**

- returns critical value for argument
- $\text{qt}(0.975, \text{df}=522) \rightarrow 1.965$

TDist {stats}

R Documentation

The Student t Distribution

Description

Density, distribution function, quantile function and random generation for the t distribution with `df` degrees of freedom (and optional non-centrality parameter `ncp`).

Usage

```
dt(x, df, ncp, log = FALSE)
pt(q, df, ncp, lower.tail = TRUE, log.p = FALSE)
qt(p, df, ncp, lower.tail = TRUE, log.p = FALSE)
rt(n, df, ncp)
```

Arguments

- `x, q` vector of quantiles.
- `p` vector of probabilities.
- `n` number of observations. If `length(n) > 1`, the length is taken to be the number required.
- `df` degrees of freedom (> 0 , maybe non-integer). `df = Inf` is allowed.
- `ncp` non-centrality parameter *delta*; currently except for `rt()`, only for $\text{abs}(\text{ncp}) \leq 37.62$. If omitted, use the central t distribution.
- `log, log.p` logical; if TRUE, probabilities `p` are given as $\log(p)$.
- `lower.tail` logical; if TRUE (default), probabilities are $P[X \leq x]$, otherwise, $P[X > x]$.

H0: $\text{beta_exper} = \text{beta_educ}$,
H1: not H0

$$\log(wage) = \beta_0 + \beta_{\text{educ}} \text{educ} + \beta_{\text{exper}} \text{exper} + \beta_{\text{tenure}} \text{tenure} + u$$

$$\log(wage) = \beta_0 + \beta_{\text{educ}} \text{educ} + (\theta + \beta_{\text{educ}}) \text{exper} + \beta_{\text{tenure}} \text{tenure} + u$$

$$\log(wage) = \beta_0 + \theta \text{exper} + \beta_{\text{educ}}(\text{educ} + \text{exper}) + \beta_{\text{tenure}} \text{tenure} + u$$

- $\text{theta} := \text{beta_exper} - \text{beta_educ} \implies \text{beta_exper} = \text{theta} + \text{beta_educ}$
- then model can be modified
- H0: $\text{theta} = 0$, H1: $\text{theta} \neq 0$
- model should be adjusted
- you should regress adjusted model

$$\log(wage) = \beta_0 + \beta_{educ}educ + \beta_{exper}exper + \beta_{tenure}tenure + u$$

$$\log(wage) = \beta_0 + \beta_{educ}educ + (\theta + \beta_{educ})exper + \beta_{tenure}tenure + u$$

$$\log(wage) = \beta_0 + \theta exper + \beta_{educ}(educ + exper) + \beta_{tenure}tenure + u$$

- $x \leftarrow \text{data\$educ} + \text{data\$exper}$
- `adjusted_model <- lm(log(wage) ~ exper + x + tenure, data = data)`
- t-statistic for our H_0
 - $t = -0.092029 / 0.007330 = 12.55512 \sim t(df=522)$
 - $c = \pm 1.965 \Rightarrow \text{reject } H_0$

meaning of t-statistics

- numerator: distance from hypothetical value
- denominator: standard error of distance from hypothetical value
- meaning of t:
standardized value of distance between estimator and hypothetical value
 - large $|t|$ (i.e., far from 0 means rare situation under H_0)

$$t_{\hat{\beta}_j} := \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

Model of College GPA

Example (Ex4.3:College GPA model)

$$colGPA = \beta_0 + \beta_1 hsGPA + \beta_2 ACT + \beta_3 skipped$$

. reg colGPA hsGPA ACT skipped				n		
Source	SS	df	MS	Number of obs = 141		
Model	4.53313314	3	1.51104438	F(3, 137) =	13.92	
Residual	14.8729663	137	.108561798	Prob > F =	0.0000	
	df=141-3-1			R-squared =	0.2336	
Total	19.4060994	140	.138614996	Adj R-squared =	0.2168	
				Root MSE =	.32949	
k=3						
y colGPA		beta	se(beta)	t-statistics	95% CI	
		Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
					p-value	
x1	hsGPA	.4118162	.0936742	4.40	0.000	.2265819 .5970505
x2	ACT	.0147202	.0105649	1.39	0.166	-.0061711 .0356115
x3	skipped	-.0831131	.0259985	-3.20	0.002	-.1345234 -.0317028
beta0	_cons	1.389554	.3315535	4.19	0.000	.7339295 2.045178
=beta/se(beta)						

Figure : Regression result of data GPA1.raw

- two-sided $c=1.96$ (sig. level 5%), 2.58 (sig. level 1%)

Reporting Regression Results

- Mandatory
 - dependent variable x_j
 - estimated OLS coef. $\hat{\beta}_j$
 - standard error $se(\hat{\beta}_j)$: in parentheses
 - more general form (especially when H_0 is not like $\hat{\beta} = 0$)
 - R^2
 - # of obs. (n)
- Optional
 - SSR
 - F -statistics for overall significance of this regression
 - df
 - t -statistics
 - CI
 - p -value
- for example, see Figure 8

더빈왓슨 통계량과 잔차플롯

금융자료의 회귀분석을 위한 전제조건

- 모든 회귀분석은 A1 - A5 가 충족되어야 함
- 금융자료의 경우 시간과 관계가 깊음
 - 시계열 자료
- 시계열 자료는 미래가 과거에 의존하는 특성상 잔차가 독립이 아닐 가능성이 높음
 - 이 경우 A5 조건 (잔차간 독립)을 위배 ⇒ 회귀분석 결과를 신뢰할 수 없음
- 잔차의 상관관계를 검정할 필요가 있음

자기상관

Autocorrelation

- 시간과 관계있는 변수의 경우 가장 가까운 과거값이 영향을 미칠 가능성이 가장 높은 경우가 일반적
 - 영향이 시간 간격을 둘 경우에는 시간간격 만큼의 과거값이 더 큰 영향을 미칠 수도 있음
- 자기상관을 검토하기 위한 방법
 - 더빈왓슨(DW) 통계량 - 수치적 방법
 - 잔차플롯 - 시각적 방법

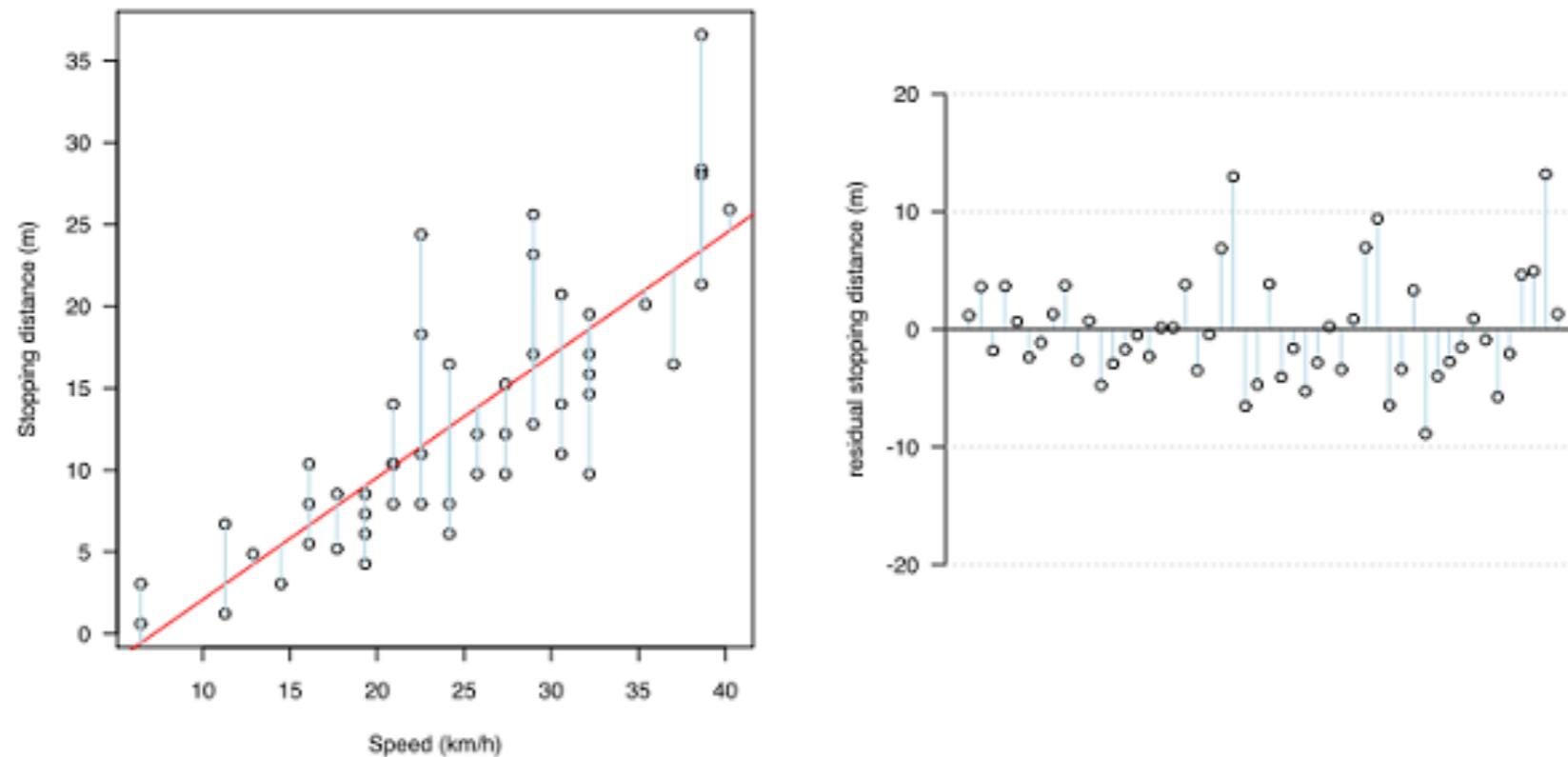
Durbin-Watson (DW) Statistics

- 무상관의 경우 $DW \approx 2$
- 양의 상관 $DW \approx 0$
- 음의 상관 $DW \approx 4$
- 1차 자기상관만을 검토한다
는 한계가 있음

$$DW := \frac{\sum_{t=2}^T (\hat{\epsilon}_t - \hat{\epsilon}_{t-1})^2}{\sum_{t=1}^T \hat{\epsilon}_t^2}$$
$$\approx 2 - 2 \frac{\sum_t \hat{\epsilon}_t \hat{\epsilon}_{t-1}}{\sum_t \hat{\epsilon}_t^2}$$

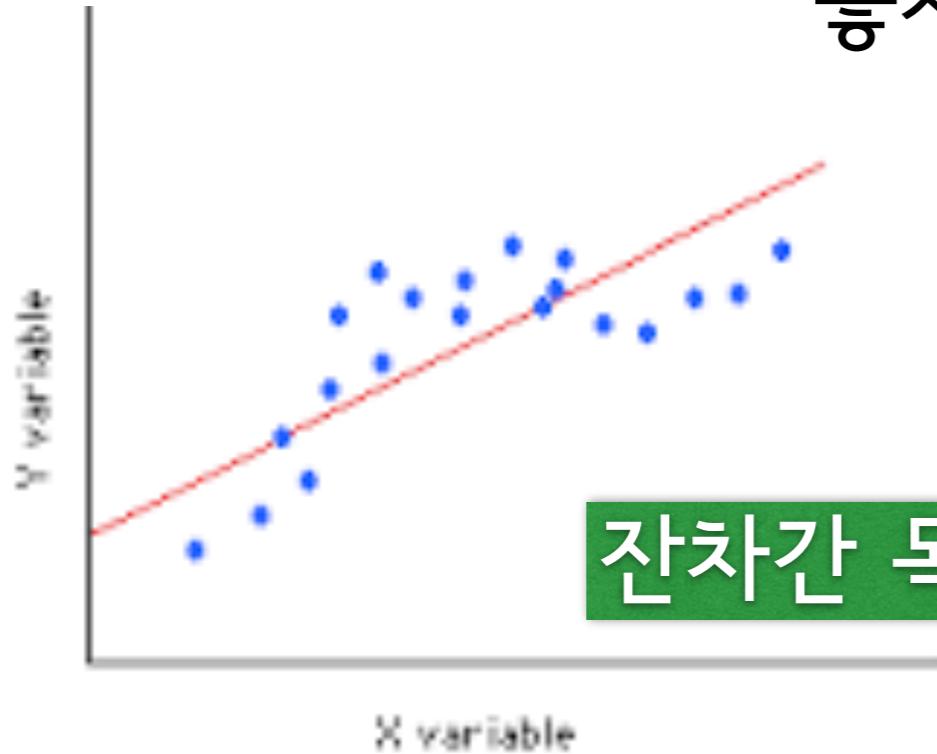
잔차플롯 Residual Plot

- 종속변수 예측치(\hat{y})를 가로축에, 잔차($\hat{\epsilon}$)을 세로축에 두고 그린 산점도
- ϵ_t 가 서로 독립이라면 $\hat{\epsilon}$ 은 아무런 패턴을 보이지 않아야 함 (White Noise)

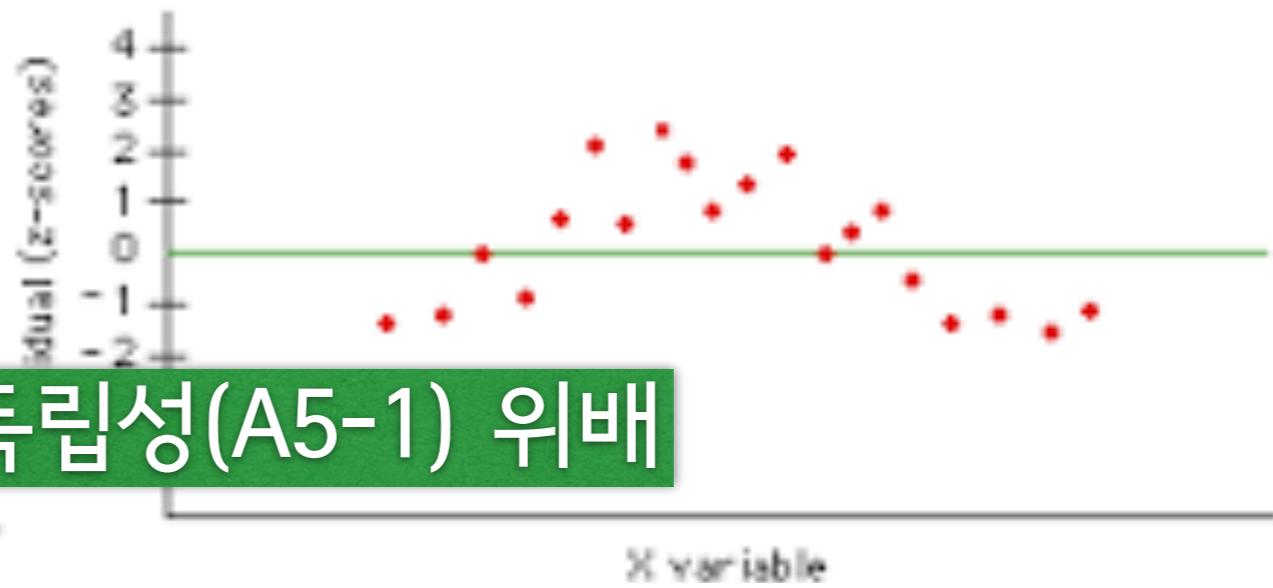


Residual Plots: Examples

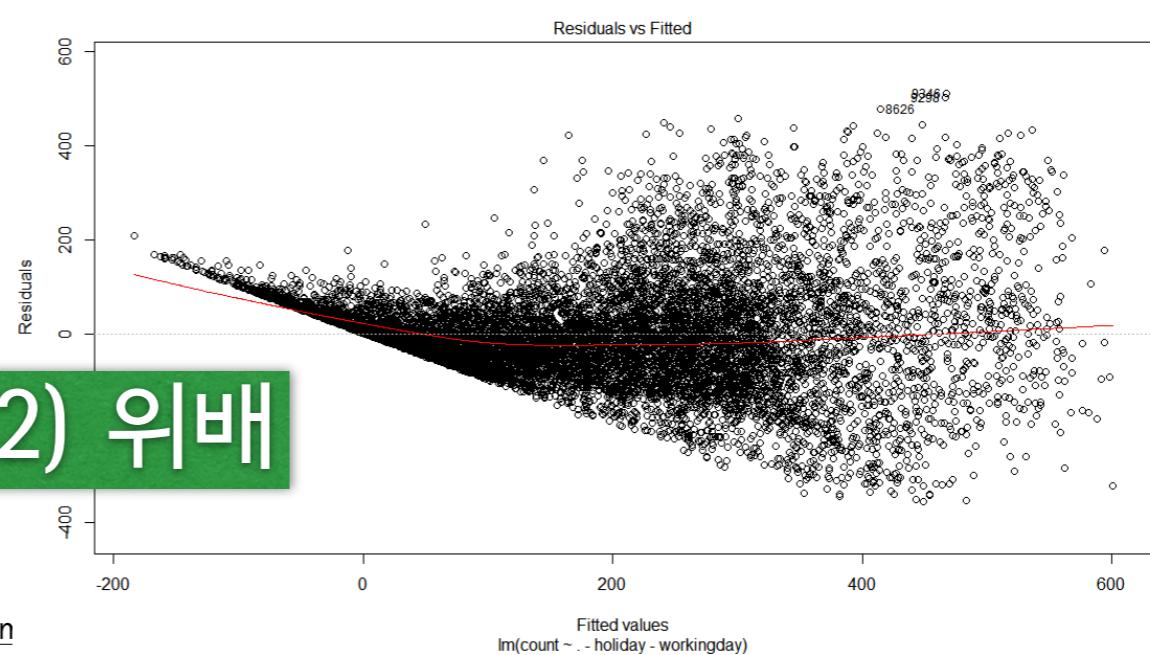
좋지 않은 사례



잔차간 독립성(A5-1) 위배



등분산성(A5-2) 위배



자기상관, 혹은 A5 조건을 충족하지 못할 경우

- 충족할 경우 \Rightarrow OLS Analysis
- 충족하지 못할 경우 \Rightarrow OLS는 좋은 선택이 아님
 \Rightarrow 다른 더 나은 추정법 사용
 - 상급 수업에서 다룰 내용

회귀계수 $\hat{\beta}$ 의 CI

- 통계량의 분포로부터 유의수준에 상응하는 critical value 찾기 $\Rightarrow c$ (t 분포표로부터 구할 수 있음)
- $\hat{\beta}_j - c \times se(\hat{\beta}_j) \leq CI \leq \hat{\beta}_j + c \times se(\hat{\beta}_j)$
- Ex8.7, 8.8

Confidence Interval (CI)

Definition ($Z\%$ Confidence Interval (CI) for β_j)

Let c be the critical value with $Z\%$ significant level,
then $\underline{Z\% \text{ CI for } \beta_j} \in (\underline{\beta}_j, \bar{\beta}_j)$

$$\underline{\beta}_j := \hat{\beta}_j - c \cdot se(\hat{\beta}_j)$$

$$\bar{\beta}_j := \hat{\beta}_j + c \cdot se(\hat{\beta}_j)$$

- $x \in (a, b) := a < x < b$ (interval)
- Meaning
 - $\Pr(\beta_j \in CI) = \Pr(\underline{\beta}_j < \beta_j < \bar{\beta}_j) = Z\%$
 - \therefore This does NOT guarantee $\beta \in CI$ from our sample

Inference Procedure with CI

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

STEP 1 Gather data $\xrightarrow{OLS} \hat{\beta}, se(\hat{\beta})$

STEP 2 Check $(n, k) \xrightarrow{t-distribution} c$

STEP 3 Get CI: significant range for β_j

STEP 4 $a_j \in CI$: can NOT reject H_0 , otherwise can reject H_0

변동분해, 결정계수

SST, SSE, SSR

$$y_i = \hat{y}_i + \hat{u}_i \quad (2.32)$$

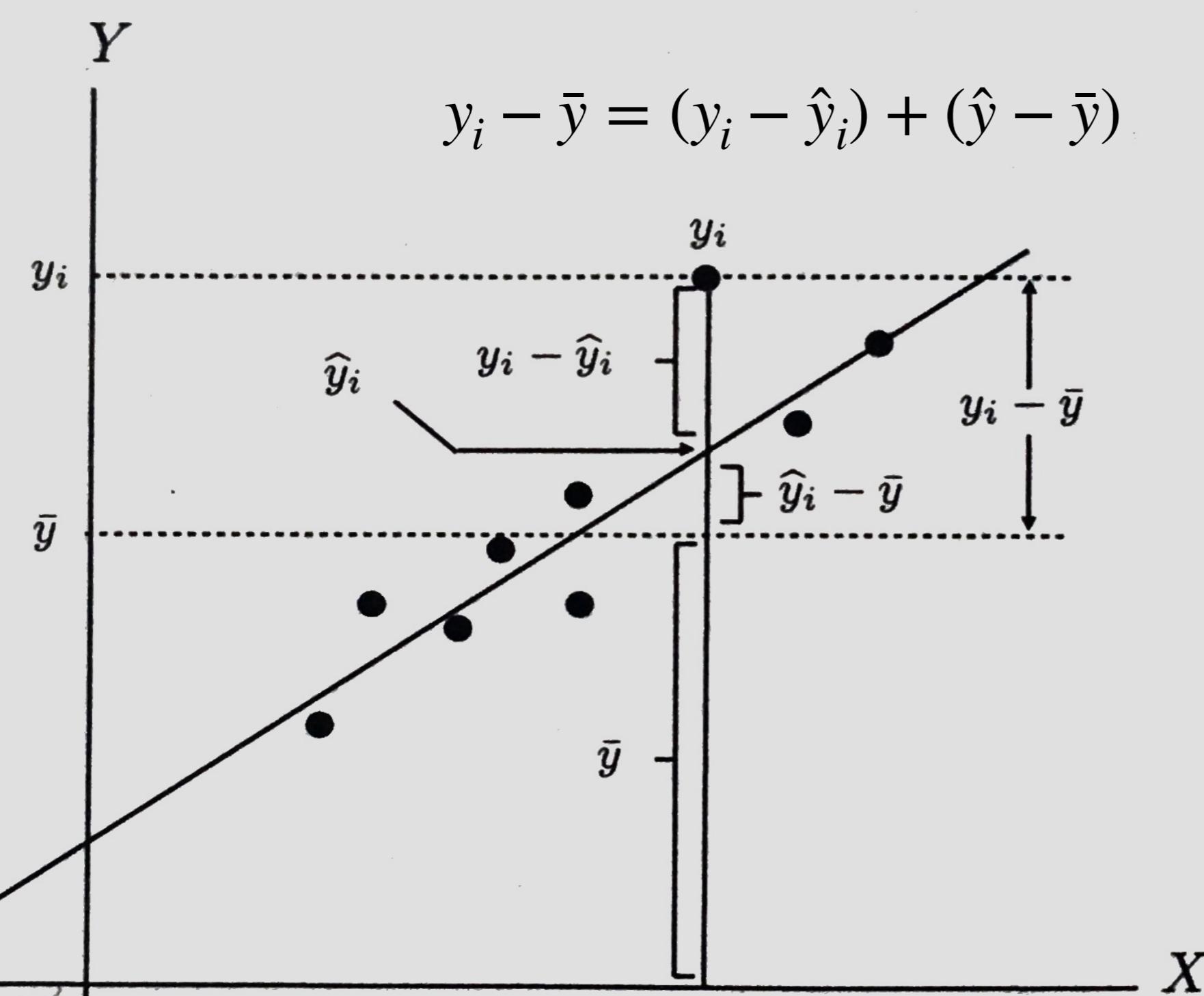
- y_i : measured (from one sample)
- \hat{y}_i : predicted part of y_i : calculated from x_i and OLS estimates (from all sample data)
- \hat{u}_i : unpredicted part of y_i

Definition (Total Sum of Squares (**TSS**, Explained Sum of Squares (**RSS**), Sum of Squared Residuals (**SSE** 2.20))

$$\text{TSS} := \sum_i (y_i - \bar{y})^2 \quad \text{RSS} := \sum_i (\hat{y}_i - \bar{y})^2 \quad \text{SSE} := \sum_i \hat{u}_i^2 \quad (2.33 - 2.35)$$

Definition (Total Sum of Squares (TSS), Explained Sum of Squares (RSS, Sum of Squared Residuals (SSE) 2.20))

$$\text{TSS} := \sum_i (y_i - \bar{y})^2 \quad \text{RSS} := \sum_i (\hat{y}_i - \bar{y})^2 \quad \text{SSE} := \sum_i \hat{u}_i^2 \quad (2.33 - 2.35)$$



Goodness of Fit: R^2

$$TSS = RSS + SSE \quad (2.36)$$

- Total Variation = Explained Variation + Unexplained Variation

Definition (R-squared)

$$R^2 := \frac{RSS}{TSS} = 1 - \frac{SSE}{TSS} \quad (2.38)$$

- Meaning: Explained Variation/Total Variation: Fraction of explained variation by OLS estimate.
- Goodness of Fit
 - Use carefully: even OLS with low R^2 can be meaningful and high R^2 can be useless

수정결정계수 Adjusted R²

$$\text{Adjusted } R^2 := 1 - \frac{n-1}{n-k}(1 - R^2)$$

- R^2 이 1에 가까울 수록 Good News
 - 0에 가까울수록 Bad News
 - 하지만, 설명력이 거의 없는 독립변수를 추가할 경우 R^2 이 증가하는 성질이 존재 \Rightarrow Adjusted R^2
- 수정결정계수 (Adjusted R^2)
 - k: 독립변수의 갯수
 - 독립변수 추가로 이 값이 줄어들면 설명력이 적은 것으로 간주 \Rightarrow 추가하지 않음

F-검정

- 개별 설명변수는 각각 귀무가설을 기각하지 못하더라도 전체 모형은 귀무가설을 기각할 수 있는 경우가 있음
- Q: 최소한 하나의 회귀계수는 0이 아닌가?
 - Ha: 최소한 하나의 회귀계수는 0이 아니다
- $H_0 : \beta_j = 0, \quad \forall j = 1, \dots, k$
- 모형 전체의 설명력은 RSS/SSE로 표현 가능
 - RSS: 모형으로 설명되는 변동
 - SSE: 모형으로 설명 못하는 변동
- RSS/SSE가 클 경우: 이 모형은 설명력이 있다

RSS/SSE에 대한 검정

$$\text{TSS} := \sum_i (y_i - \bar{y})^2 \quad \text{RSS} := \sum_i (\hat{y}_i - \bar{y})^2 \quad \text{SSE} := \sum_i \hat{u}_i^2$$

$$TSS = RSS + SSE$$

$$TSS/\sigma^2 \sim \chi^2(df = n - 1)$$

$$RSS/\sigma^2 \sim \chi^2(df = k)$$

$$SSE/\sigma^2 \sim \chi^2(df = n - k - 1)$$

$$\frac{RSS/k}{SSE/(n - k - 1)} \sim F(k, n - k - 1)$$

분산분석표 ANOVA table

Regression Model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$$

	df	SS	MS	F
회귀	k	RSS	RSS/k	$\frac{RSS/k}{SSE/(n - k - 1)}$
잔차	n-k-1	SSE	SSE/(n-k-1)	
총합	n-1	TSS		

Regression Table (STATA)

ANOVA Table

Source	degree of freedom		
	SS	df	MS
Model	4.53313314	3	1.51104438
Residual	14.8729663	137	.108561798
	$df = 141 - 3 - 1$		
Total	19.4060994	140	.138614996

k=3

y colGPA	Coef.	Std. Err.	t	P> t	95% CI	
					p-value	[95% Conf. Interval]
x1 hsGPA	.4118162	.0936742	4.40	0.000	.2265819	.5970505
x2 ACT	.0147202	.0105649	1.39	0.166	-.0061711	.0356115
x3 skipped	-.0831131	.0259985	-3.20	0.002	-.1345234	-.0317028
beta0_cons	1.389554	.3315535	4.19	0.000	.7339295	2.045178

F-test

n	Number of obs =	141
F(3, 137) =	13.92	
Prob > F =	0.0000	
R-squared =	0.2336	
Adj R-squared =	0.2168	
Root MSE =	.32949	

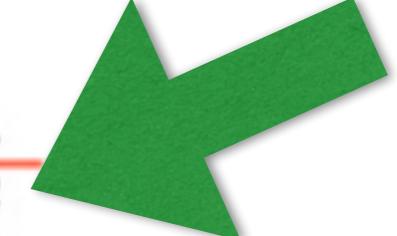


Figure : Regression result of data GPA1.raw

예측과 신뢰구간

Prediction Model

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik} = \mathbf{x}'_i \hat{\boldsymbol{\beta}}$$

- x_0 값을 알고 있지만 y_0 값은 모르는 경우
 - 모분산도 모르는 경우를 상정
(대부분이 이 경우에 해당)
 - 모형으로 예측
 - 점추정: Estimated Model에 x_0 을 대입하면 됨:
$$\hat{y}_0 = \mathbf{x}'_0 \hat{\boldsymbol{\beta}}$$
 - 구간추정: y_0 에 대한 신뢰구간을 구하면 됨

y₀에 대한 CI

Prediction Model

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik} = \mathbf{x}'_i \hat{\boldsymbol{\beta}}$$

Regression Model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i = \mathbf{x}'_i \boldsymbol{\beta} + \epsilon_i$$

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

True Value: 상수

$$\begin{aligned} Var(\hat{y}_0 - y_0) &= Var(\mathbf{x}'_0 (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \epsilon_0) \\ &= Var(\mathbf{x}'_0 \hat{\boldsymbol{\beta}}) + Var(\epsilon_0) \\ &= \mathbf{x}'_0 (X'X)^{-1} \mathbf{x}_0 \sigma^2 + \sigma^2 \\ &= (1 + \mathbf{x}'_0 (X'X)^{-1} \mathbf{x}_0) \sigma^2 \end{aligned}$$

$$\therefore \hat{y}_0 - y_0 \sim N(0, (1 + \mathbf{x}'_0 (X'X)^{-1} \mathbf{x}_0) \sigma^2)$$

CI 구하기 (계속)

$$\hat{y}_0 - y_0 \sim N(0, (1 + \mathbf{x}'_0 (X'X)^{-1} \mathbf{x}_0) \sigma^2)$$

$$\frac{\hat{y}_0 - y_0}{\sqrt{(1 + \mathbf{x}'_0 (X'X)^{-1} \mathbf{x}_0) \hat{\sigma}^2}} \sim t(df = n - k - 1)$$

y₀에 대한 1- α 신뢰구간

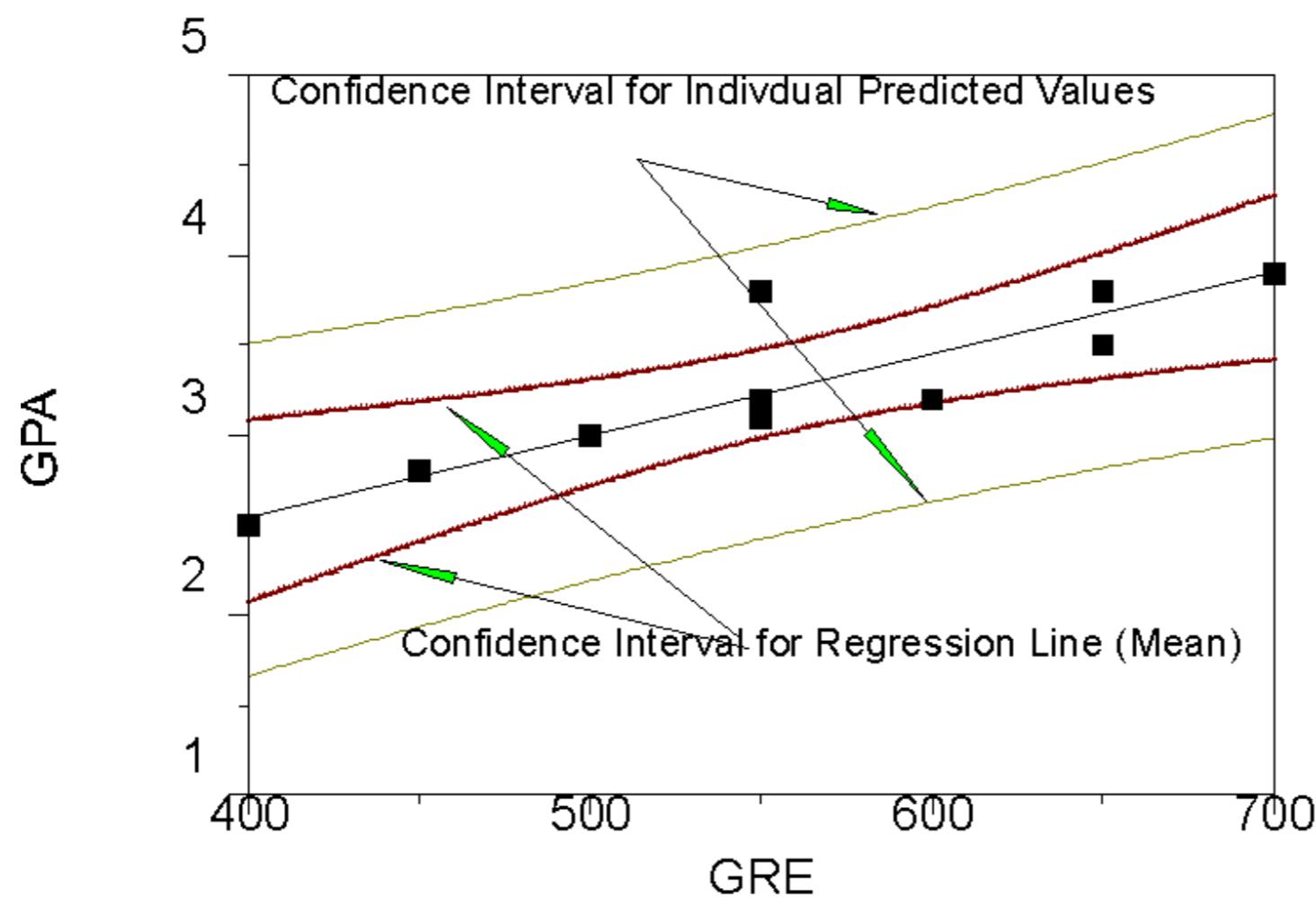
$$\hat{y}_0 - t_{\alpha/2, n-k-1} \sqrt{(1 + \mathbf{x}'_0 (X'X)^{-1} \mathbf{x}_0) \hat{\sigma}^2} \leq y_0 \leq \hat{y}_0 + t_{\alpha/2, n-k-1} \sqrt{(1 + \mathbf{x}'_0 (X'X)^{-1} \mathbf{x}_0) \hat{\sigma}^2}$$

시각화

y_0 에 대한 $1-\alpha$ 신뢰구간

$$\hat{y}_0 - t_{\alpha/2, n-k-1} \sqrt{(1 + \mathbf{x}'_0 (X'X)^{-1} \mathbf{x}_0) \hat{\sigma}^2} \leq y_0 \leq \hat{y}_0 + t_{\alpha/2, n-k-1} \sqrt{(1 + \mathbf{x}'_0 (X'X)^{-1} \mathbf{x}_0) \hat{\sigma}^2}$$

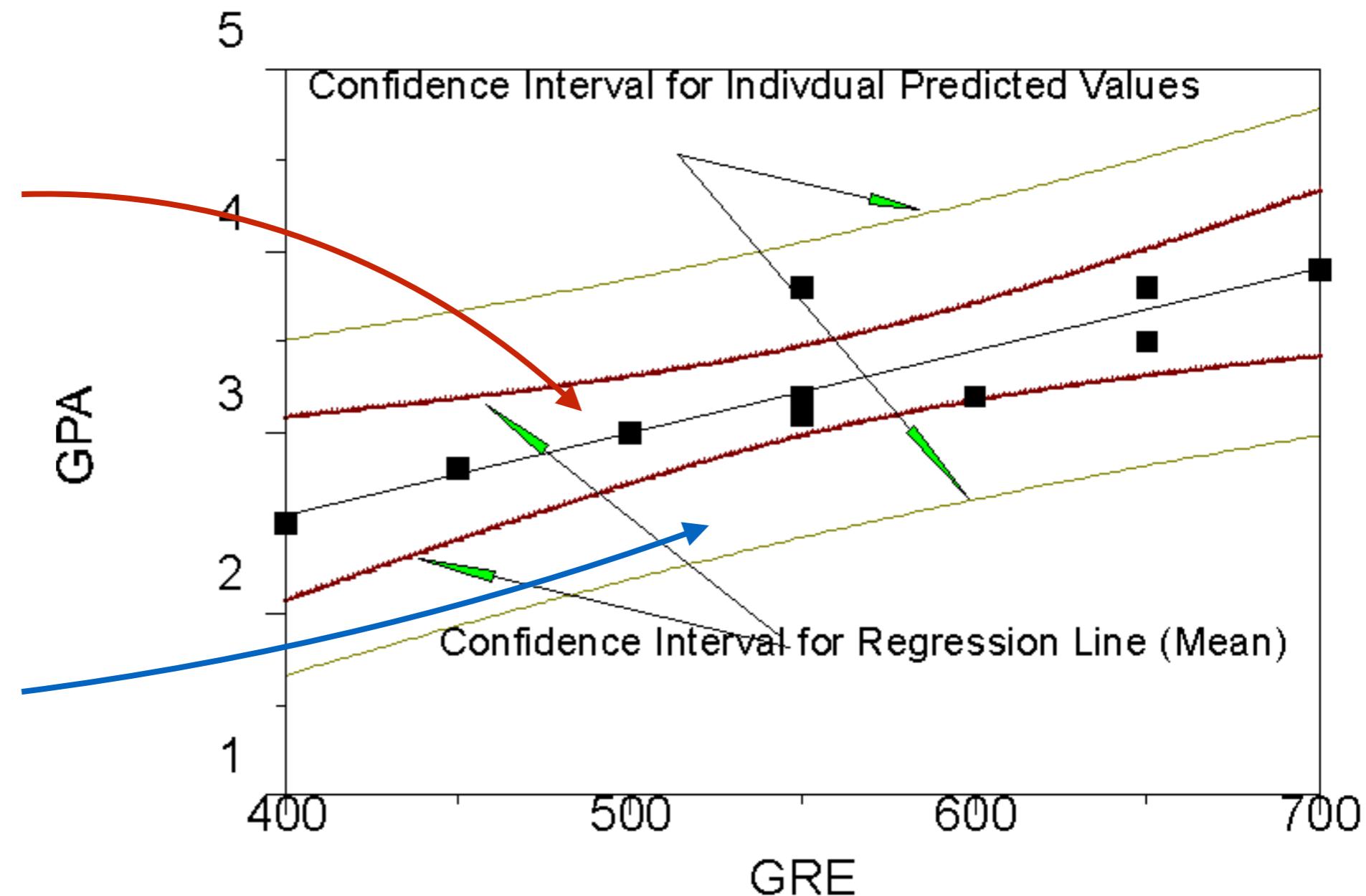
Data from Partial Correlation Example



Prediction Interval

Data from Partial Correlation Example

- $E(y_0)$ 에 대한 신뢰구간
 - regression line의 신뢰 구간
- y_0 에 대한 신뢰구간
 - 개별값에 대한 신뢰구간
 - 더 넓음
 - Prediction Interval

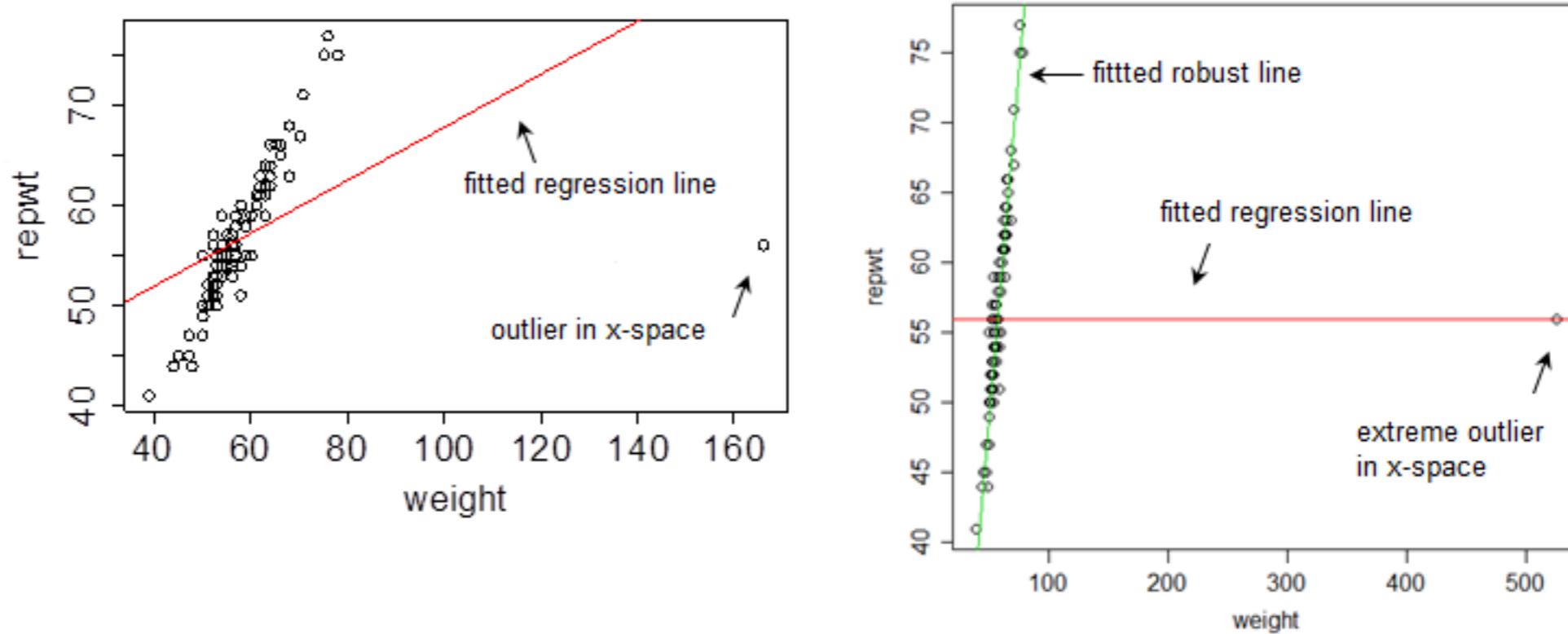


다중공선성

Multicollinearity

- 설명변수간 완벽하진 않더라도 선형관계가 강하게 나타날 경우 표준오차를 커지게 만듦
 - 넓은 CI가 도출되어 설명력이 약해짐
- 대책: 선형관계가 강한 변수들 중 하나를 제거
 - 일반적으로 상관계수 절대값이 0.9보다 크면 둘 중 하나를 제거함

이상치 Outlier



- 이상치는 회귀식을 부정확하게 만듦
- 대책: outlier 제거

회귀분석의 응용

시장모형

Market Model

$$R_{it} = \alpha_i + \beta_i R_{mt} + \epsilon_{it}$$

Regression Model

- 특정개별 주식의 수익률과 시장 포트폴리오 수익률 간의 선형관계를 나타내는 수익률 모형
- Notation
 - R_{it}: 개별주식 i의 시점 t에서의 수익률
 - R_{mt}: 시점 t에서의 시장 포트폴리오의 수익률
 - 현실적으로 구할 수 없는 값 \Rightarrow 종합주가지수 수익률을 사용

$$\hat{R}_{it} = \hat{\alpha}_i + \hat{\beta}_i R_{mt}$$

Prediction Model:
SCL(Security Characteristic Line)

증권특성선

해석

$$\hat{R}_{it} = \hat{\alpha}_i + \hat{\beta}_i R_{mt}$$

Prediction Model:
SCL(Security Characteristic Line)

- hat α_i
 - 시장포트폴리오 수익률이 0일때 개별증권 i의 수익률
- hat β_i
 - 시장포트폴리오 수익률이 1 증가할 때 개별증권 수익률의 증가량
 - >1 : 시장보다 민감하게 반응 \Rightarrow 공격적 자산
 - <1 : 시장보다 둔감하게 반응 \Rightarrow 방어적 자산

자본자산가격결정모형

CAPM

- CAPM: Capital Asset Pricing Model
- 효율적 분산투자의 원리에 따라 행동하는 경우 균형시장에서 자본자산의 위험과 기대수익률 간의 균형 관계를 설명하고자 하는 이론
 - 자본시장선 Capital Market Line (CML)
 - 증권시장선 Security Market Line (SML)
 - CAPM이라고도 부름
- 본 절에서는 간단한 아이디어만 소개

자본시장선 (CML)

- 포트폴리오를 ω 의 위험자산과 $1-\omega$ 의 무위험자산으로 구성한 상황
 - R_m : 위험자산 수익률 (확률변수)
 - R_f : 무위험자산 수익률 (상수)
- 일련의 CAPM 가정 하에서 다음의 결과를 도출할 수 있음:

$$E(R_p) = \omega E(R_m) + (1 - \omega)R_f$$

$$\sigma_p = \omega \sigma_m$$

$$\Rightarrow E(R_p) = R_f + \left(\frac{E(R_m) - R_f}{\sigma_m} \right) \sigma_p$$

CML

해석

CML

$$E(R_p) = R_f + \left(\frac{E(R_m) - R_f}{\sigma_m} \right) \sigma_p$$

절편

기울기

효율적 포트폴리오의
시간가치

총위험 한 단위에 대한
자본시장에서의 리스크
프리미엄

위험균형가격
위험의 시장가격

CAPM

- CML은 개별증권, 혹은 비효율적 포트폴리오와 같은 non systemic risk를 가지고 있는 경우의 기대수익률과 위험간의 관계를 다루지 못함
- CAPM:
 - 균형시장하에서 효율적 포트폴리오를 포함한 모든 위험자산의 기대수익률과 체계적 위험간의 관계를 설명하는 모형

$$\begin{aligned} E(R_i) &= R_f + \left(\frac{E(R_m) - R_f}{\sigma_m} \right) \beta_i \sigma_m \\ &= R_f + (E(R_m) - R_f) \beta_i \end{aligned}$$

시장 포트폴리오와 개별 증권간의 관계

시장 포트폴리오의 초과수익률

과제3

- 자신의 학번 끝자리 3자리수를 종목번호로 하는 기업의 3년치 주식(종속변수)과 KOSDAQ(설명변수)에 대한 단순선형회귀분석 시행
 - 필수적으로 넣어야 하는 것
 - Data import ⇒ Scatter Plot ⇒ Regression Model ⇒ Regression Line + 90% CI 시각화
 - 옵션: 더빈왓슨 통계량, 잔차플롯 (추가시 가산점 10%)
 - Rmd file 형태로 LMS 과제란에 제출할 것
 - 그 자체로 데이터 수집부터 시각화까지 완결적이어야 완전한 점수를 받을 수 있음 (직접 컴파일하여 확인)
 - 사용한 패키지 명시할 것
 - 기한: 2018.12.20 23:59

기말시험 공시

- 12.14 (금)
- 수업시간, 장소 동일
- 연습문제를 충분히 풀어볼 것
- 시험장에 미리 준비한 A4 1page 짜리 노트 한 장을 가지고 시험을 볼 수 있음 (옵션)
 - 공식 등 답안 작성에 도움이 될 것으로 예상하는 내용을 준비해 올 것

한학기동안
수고하셨습니다!

