8. 단순 선형 회귀 모형

회귀분석 소개

- **회귀분석 소개**
  - 회귀분석
    - 독립변수와 종속변수 간의 함수적인 관련성을 규명하기 위하여 어떤 수학적 모형을 가정하고, 이 모형을 측정된 변수들의 자료로부터 통계적으로 추정 및 검정을 하거나 추정된 모형을 예측에 활용하는 분석방법
    - $y = f(x)$의 함수 관계가 있을 때,
      - x를 설명변수(explanatory variable), 예측변수 (predictor) 또는 독립변수(independent variable)
      - y를 반응변수(response variable) 또는 종속변수(dependent variable)

회귀분석 소개

- 회귀분석 적용의 예

종속변수	독립변수
매출액	광고비, 품질, 가격
시장점유율	연구개발비, 품질, 가격
1인당 저축액	소득액, 소비성향, 부양가족수
임금	학력, 경력, 나이, 성별
주식가격	금리, 부동산가격, 통화량, 경상수지
다음 한 주간의 주가상승률	오늘까지의 주식가격

- 회귀분석의 목적
  - 자료의 탐색, 요약, 정리
  - 모형의 설정
  - 모수의 추정
  - 반응값의 추정

단순선형회귀모형의 정의

- **단순선형회귀모형의 정의**

  독립변수의 정해진 값 $x_1, \ldots, x_n$에서 측정되는 종속변수 $Y_1, \ldots, Y_n$에 대하여 다음의 관계식이 성립한다고 가정함.

$$Y_i = \alpha + \beta x_i + \varepsilon_i \quad (i = 1,2, \ldots, n)$$

  - α, β 는 회귀계수로 상수임. α는 회귀선의 절편, β는 회귀선의 기울기를 나타냄.

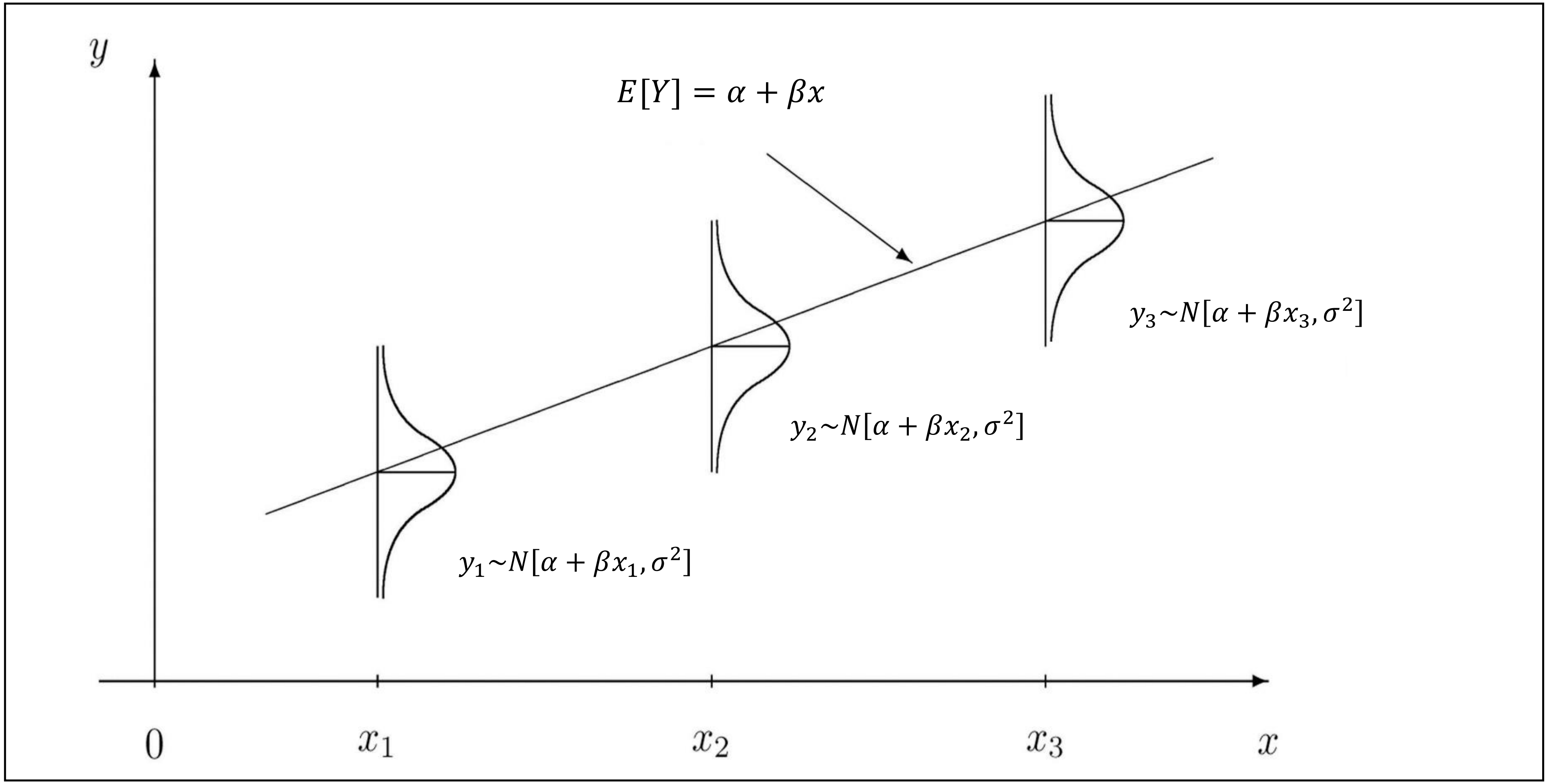
  - 오차항은 $\varepsilon_i \sim N[0, \sigma^2]$인 확률변수

  - $Y_i = \alpha + \beta x_i + \varepsilon_i$ 이므로,
    $Y_i \sim N[\alpha + \beta x_i, \sigma^2]$인 확률변수임.

  - $E[Y_i] = \alpha + \beta x_i$: 주어진 x_i에서의 Y_i의 기대값을 $\alpha + \beta x$의 선형함수로 표현함.

단순선형회귀모형의 정의

- 모형에 관한 가정 요약



단순선형회귀모형의 정의

- 삼성전자의 일별 수익률에 대한 시장 모형 예제

  'reg_data1.csv' 데이터는 2022년 1월 3일부터 5월 24일 까지의 일별 삼성전자 주가와 kospi 지수를 수집한 것이다. 이 자료를 이용하여 삼성전자의 일간 수익률(Y)와 kospi 지수의 일간 수익률(X) 간의 관계를 설명하는 시장모형(market model),을 추정하여라.

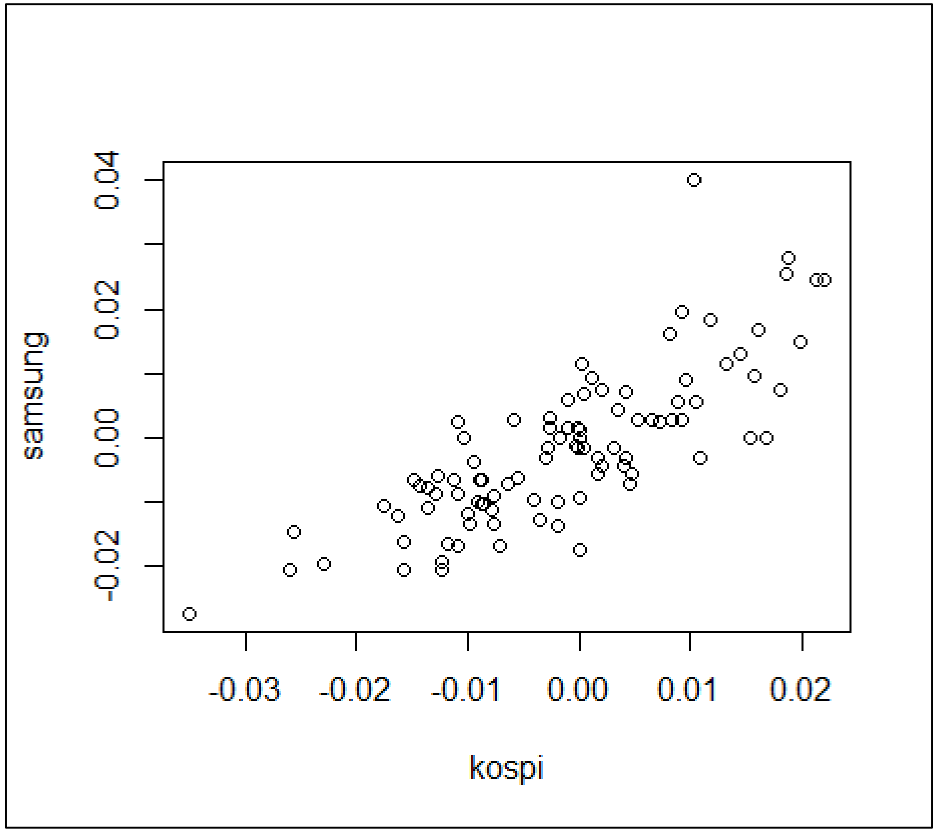
```
> rm(list=ls())
> setwd("E:/DFMBA 경영통계")
> mktdata = read.csv('reg_data1.csv')
> head( mktdata )
        Date samsung   kospi
1 2022-01-03   78600 2988.77
2 2022-01-04   78700 2989.24
3 2022-01-05   77400 2953.97
4 2022-01-06   76900 2920.53
5 2022-01-07   78300 2954.89
6 2022-01-10   78000 2926.72
```

단순선형회귀모형의 정의

```
> nr <- nrow( mktdata )
> mktdata$lagkospi <- c(NA, mktdata$kospi[1:(nr-1)])
> mktdata$lagsamsung <- c(NA, mktdata$samsung[1:(nr-1)])
> mktdata$rtrnkospi <- (mktdata$kospi - mktdata$lagkospi)/mktdata$lagkospi
> mktdata$rtrnsamsung <- (mktdata$samsung - mktdata$lagsamsung)/mktdata$lagsamsung
> head( mktdata )
        Date samsung   kospi lagkospi lagsamsung     rtrnkospi  rtrnsamsung
1 2022-01-03   78600 2988.77       NA         NA            NA           NA
2 2022-01-04   78700 2989.24  2988.77      78600  0.0001572553  0.001272265
3 2022-01-05   77400 2953.97  2989.24      78700 -0.0117989857 -0.016518424
4 2022-01-06   76900 2920.53  2953.97      77400 -0.0113203587 -0.006459948
5 2022-01-07   78300 2954.89  2920.53      76900  0.0117649879  0.018205462
6 2022-01-10   78000 2926.72  2954.89      78300 -0.0095333498 -0.003831418
```

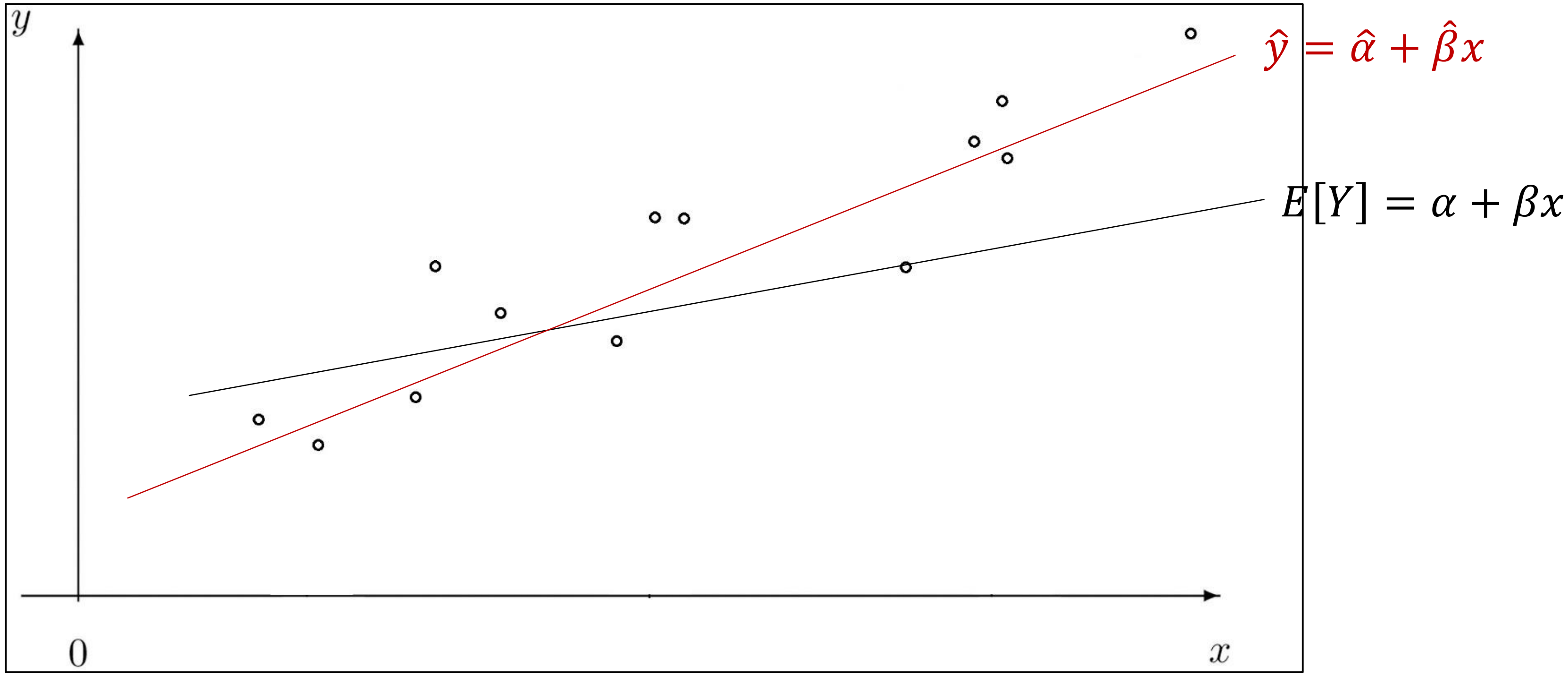
단순선형회귀모형의 정의

```
> plot( mktdata$rtrnkospi, mktdata$rtrnsamsung, xlab="kospi", ylab="samsung" )
```



단순선형회귀모형에서 모수의 추정

- **단순선형회귀모형의 추정**
  - 모형이 포함한 미지의 모수 α, β, σ^2 를 추정하기 위하여 각 독립변수 x_i 에 대응하는 종속변수 y_i 로 짝지어진 n 개의 표본 관찰치 (x_i, y_i)가 주어짐.



단순선형회귀모형에서 모수의 추정

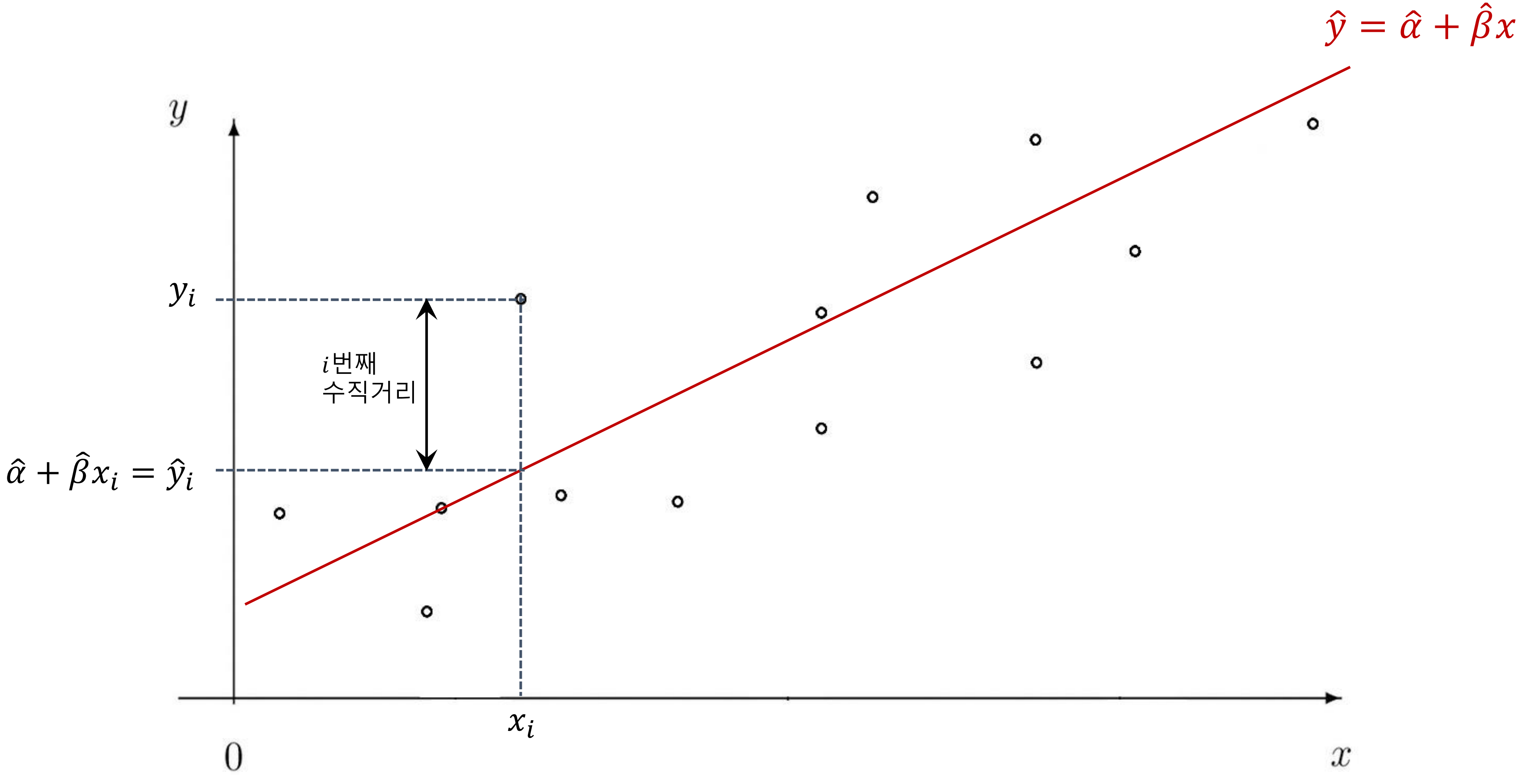
- **최소제곱추정법(Least Square Estimation)**을 이용한 회귀계수 α와 β의 추정 방법

  - 단순회귀모형 $Y_i = \alpha + \beta x_i + \varepsilon_i$에서 관찰된 자료와 회귀직선 간의 수직거리 제곱합

$$SS(\hat{\alpha}, \hat{\beta}) = \sum_{i=1}^{n} \left(y_i - \hat{\alpha} - \hat{\beta} x_i\right)^2$$

이 최소가 되도록 α와 β를 추정하는 방법을 최소제곱법 이라고 하고,
이 때 얻어지는 추정량 $\hat{\alpha}$, $\hat{\beta}$ 은 최소제곱초정량(least square estimators)이라고 함.

단순선형회귀모형에서 모수의 추정



단순선형회귀모형에서 모수의 추정

- 최소제곱추정량 $\hat{\alpha}$, $\hat{\beta}$ 의 도출
  - 선형방정식계

$$\begin{cases} \dfrac{\partial SS}{\partial \hat{\alpha}} = -2\sum_{i=1}^{n}\left(y_i - \hat{\alpha} - \hat{\beta}x_i\right) = 0 \\ \dfrac{\partial SS}{\partial \hat{\beta}} = -2\sum_{i=1}^{n} x_i\left(y_i - \hat{\alpha} - \hat{\beta}x_i\right) = 0 \end{cases}$$

을 풀어주면, 최소제곱 추정량 $\hat{\alpha}$, $\hat{\beta}$은 다음 식으로 정리됨.

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}\ , \qquad \hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

(단, $\bar{x}$는 x_i의 평균, $\bar{y}$는 y_i의 평균)

단순선형회귀모형에서 모수의 추정

- y_i의 추정치(predicted values) : $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i \quad (i = 1,2,\dots,n)$

- 잔차(residuals) : $e_i = y_i - \hat{y}_i = y_i - \hat{\alpha} - \hat{\beta} x_i \quad (i = 1,2,\dots,n)$

단순선형회귀모형에서 모수의 추정

- 오차항의 분산 σ^2의 추정

  - 오차에 대응되는 잔차의 변동성을 이용하여 아래와 같이 정의되는 MSE 로 추정함.

$$SSE = \sum_{i=1}^{n} e_i^2$$

$$MSE = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n} \left(y_i - \hat{\alpha} - \hat{\beta} x_i\right)^2}{n-2}$$

  - $E[MSE] = \sigma^2$ 임을 보일 수 있음.

  - σ^2 의 추정량은 $\hat{\sigma}^2 = MSE$를 이용함.

단순선형회귀모형에서 모수의 추정

- 삼성전자의 일별 수익률에 대한 시장 모형 예제(계속)

```
lm(
  formula,          # 종속변수 ~ 독립변수 형태로 지정한 모형식
  data,             # 모형식을 적용할 데이터. 데이터프레임 형태
  ... )
```

```
> m <- lm(rtrnsamsung ~ rtrnkospi, mktdata)
> m

Call:
lm(formula = rtrnsamsung ~ rtrnkospi, data = mktdata)

Coefficients:
(Intercept)    rtrnkospi
 -0.0005156    0.8496451
```

단순선형회귀모형에서 모수의 추정

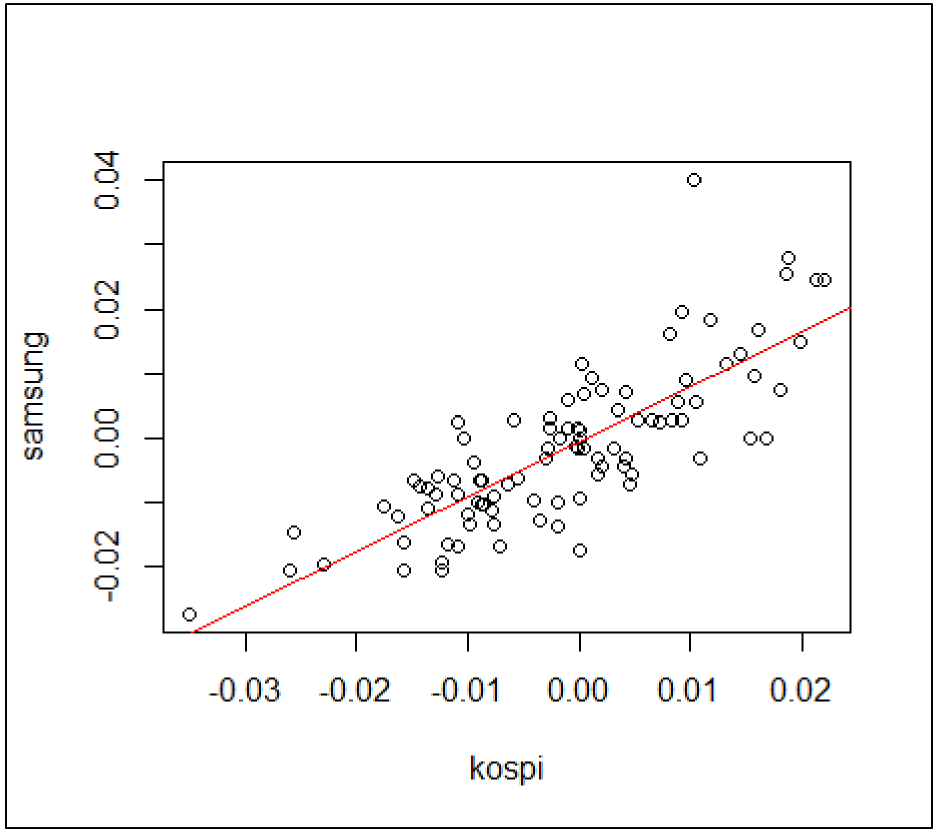
```
Coef(
  object,        # 선형회귀모델 클래스 lm의 인스턴스
  ... )
# 또는
Object$coef
```

회귀계수 추정값 : $\hat{\alpha}$, $\hat{\beta}$

```
> coef( m )
  (Intercept)     rtrnkospi
-0.0005156329  0.8496451224
> m$coefficients
  (Intercept)     rtrnkospi
-0.0005156329  0.8496451224
```

단순선형회귀모형에서 모수의 추정

```
> plot( mktdata$rtrnkospi, mktdata$rtrnsamsung, xlab="kospi", ylab="samsung" )
> abline( coef(m), col='red' )
```



단순선형회귀모형에서 모수의 추정

```
residuals(
  object,        # 선형회귀모델 클래스 lm의 인스턴스
  ... )
# 또는
object$residuals
```

잔차 : $e_i,\ i = 1, \ldots, n$

```
> residuals( m )
            2             3 ...                    95            96
 1.654286e-03 -5.977841e-03 ...         -3.559303e-03 -6.780783e-03
```

```
fitted(
  object,        # 선형회귀모델 클래스 lm의 인스턴스
  ... )
# 또는
object$fitted
```

종속변수 추정값 : $\hat{y}_i,\ i = 1, \ldots, n$

```
> fitted( m )
            2             3   ...           95            96
-3.820217e-04 -1.054058e-02   ... 2.088715e-03 -1.383777e-02
```

단순선형회귀모형에서 모수의 추정

```
Predict(
  object,  # 선형회귀모델 클래스 lm의 인스턴스
  newdata, # 예측을 수행할 새로운 x 데이터 (data.frame 형식)
  ... )
```

```
> predict( m, newdata=data.frame( rtrnkospi = 0.012 ) )
          1
0.009680109
```

단순선형회귀모형에서의 검정

- **추정량의 성질**
  - 최소제곱 추정량 $\hat{\alpha}$, $\hat{\beta}$은 다음의 분포를 가짐

$$\hat{\beta} \sim Normal\left[\beta,\ \frac{\sigma^2}{S_{xx}}\right]$$
$$\hat{\alpha} \sim Normal\left[\alpha,\ \sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)\right]$$

  - 오차항의 분산 추정량 $\hat{\sigma}^2(=MSE)$은 다음의 분포를 가짐

$$\frac{SSE}{\sigma^2} = \frac{(n-2)MSE}{\sigma^2} = \frac{(n-2)\hat{\sigma}^2}{\sigma^2} = \frac{\sum_{i=1}^{n}\left(y_i-\hat{\alpha}-\hat{\beta}\,x_i\right)^2}{\sigma^2} \sim \chi^2\,[n-2]$$

  - $\hat{\alpha}$, $\hat{\beta}$과 $\hat{\sigma}^2$은 서로 독립임.

단순선형회귀모형에서의 검정

- **모형의 유의성 검정**

  독립변수 x가 종속변수 Y를 설명하기에 유용한 변수인가에 대한 통계적 추론은 회귀계수 β에 대한 검정을 통해 파악할 수 있음.

  - **t 검정**
    - 가설

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

    - 검정통계량과 표본분포

      귀무가설 H_0 이 사실일 때,

$$T = \frac{\hat{\beta}}{\widehat{S.E.[\hat{\beta}]}} \sim t\,[n-2]$$

단순선형회귀모형에서의 검정

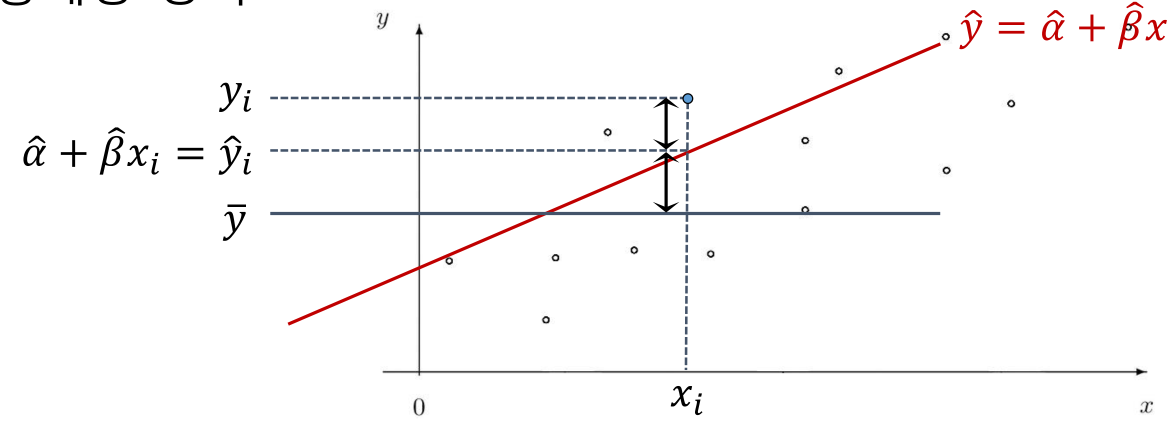
- 유의수준 α에서의 의사결정
  - 기각역 :$|T| = \left|\frac{\hat{\beta}-0}{\widehat{S.E.[\hat{\beta}]}}\right| > t_{\alpha/2,n-2}$
  - 유의확률(p-value) : $p - value\left(= P(T > |t_0|)\right) < \alpha$
  - H_0를 기각하는 경우에는 독립변수 x가 종속변수 Y를 설명하기에 유용한 변수라고 해석할 수 있음.

단순선형회귀모형에서의 검정

- **F 검정**
  - 가설

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

  - 검정통계량 정의



- y_i의 변동을 추정된 회귀모형으로 설명되는 변동과 설명되지 않는 모형으로 분할

제곱합 : $\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

SST (y_i의 변동) = SSR (모형으로 설명되는 변동) + SSE (모형으로 설명되지 않는 변동)

자유도 : $(n-1) = (1) + (n-2)$

단순선형회귀모형에서의 검정

- 평균제곱합의 정의 및 성질
  - 평균회귀제곱합 : $MSR = \frac{SSR}{1}$
    - $E[MSR] = \sigma^2 \quad , H_0 : \beta = 0$
      $E[MSR] > \sigma^2 \quad , H_1 : \beta \neq 0$
  - 평균오차제곱합 : $MSE = \frac{SSE}{n-2}$
    - $E[MSE] = E\left[\frac{SSE}{n-2}\right] = \sigma^2$

- 귀무가설 H_0이 사실일 때, $MSR \approx MSE$이고, 대립가설 H_1이 사실일 때 $MSR \gg MSE$
  ► 검정통계량을 $\frac{MSR}{MSE}$로 정의함.

- 검정통계량 값이 클수록 귀무가설 H_0에 대한 더 강한 반증이 됨. (→ **오른꼬리 검정**)

단순선형회귀모형에서의 검정

- 검정통계량의 표본분포

  귀무가설 H_0 이 사실일 때,

$$F = \frac{MSR}{MSE} = \frac{SS\,R/1}{SS\,E/(n-2)} \sim F[1, n-2]$$

- $\frac{SSR}{\sigma^2} = \frac{1 \cdot MSR}{\sigma^2} \sim \chi^2[1], \quad under\ H_0$
- $\frac{SSE}{\sigma^2} = \frac{MSE \cdot (n-2)}{\sigma^2} \sim \chi^2[n-2]$
- $\frac{SSR}{\sigma^2}$ 과 $\frac{SSE}{\sigma^2}$ 는 서로 독립

단순선형회귀모형에서의 검정

- 유의수준 α에서의 의사결정
  - 기각역 : $F = \frac{MSR}{MSE} > F_{\alpha,1,n-2}$
  - 유의확률(p-value) : $p - value\left(= P(F > f_0)\right) < \alpha$
  - H_0를 기각하는 경우에는 독립변수 x가 종속변수 Y를 설명하기에 유용한 변수라고 해석할 수 있음.

단순선형회귀모형에서의 검정

- 분산분석표를 이용하여 결과를 정리

변동의 정의	SS 통계량	자유도	MS 통계량	검정통계량
회귀모형	SSR	1	MSR	F
오차	SSE	$n-2$	MSE	
전체	SST	$n-1$		

단순선형회귀모형의 적합도

- **모형의 적합성 검토**
  - 결정계수 R^2
    - R^2의 정의

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

    - R^2의 성질 및 해석
      - $SST = SSR + SSE$이므로 항상 0과 1 사이의 값을 가짐 $(0 \le R^2 \le 1)$.
      - y_i의 변동 가운데 추정된 회귀모형으로 통해 설명되는 변동의 비중을 의미함.
      - 0에 가까울 수록 추정된 모형의 설명력이 떨어지는 것으로,
      - 1에 가까울수록 추정된 모형이 y_i의 변동을 완벽하게 설명하는 것으로 해석할 수 있음.

단순선형회귀모형의 적합도

```
summary(
  object,  # 선형회귀모델 클래스 lm의 인스턴스
  ... )
```

모형의 평가 (요약통계량, 유의성 검정, 적합도)

```
> summary( m )

Call:
lm(formula = rtrnsamsung ~ rtrnkospi, data = mktdata)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.017020 -0.004461  0.000007  0.004310  0.031861 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.0005156  0.0007511  -0.686    0.494    
rtrnkospi    0.8496451  0.0658754  12.898   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.007268 on 93 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.6414, Adjusted R-squared:  0.6376 
F-statistic: 166.4 on 1 and 93 DF,  p-value: < 2.2e-16
```

모형의 가정에 대한 검토

- **잔차분석**

회귀 모형에서의 가정이 적절한 것인가에 대한 평가

- $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$에 대한 적정성을 평가하는 것으로 흔히 다음 세가지 사안을 고려함.

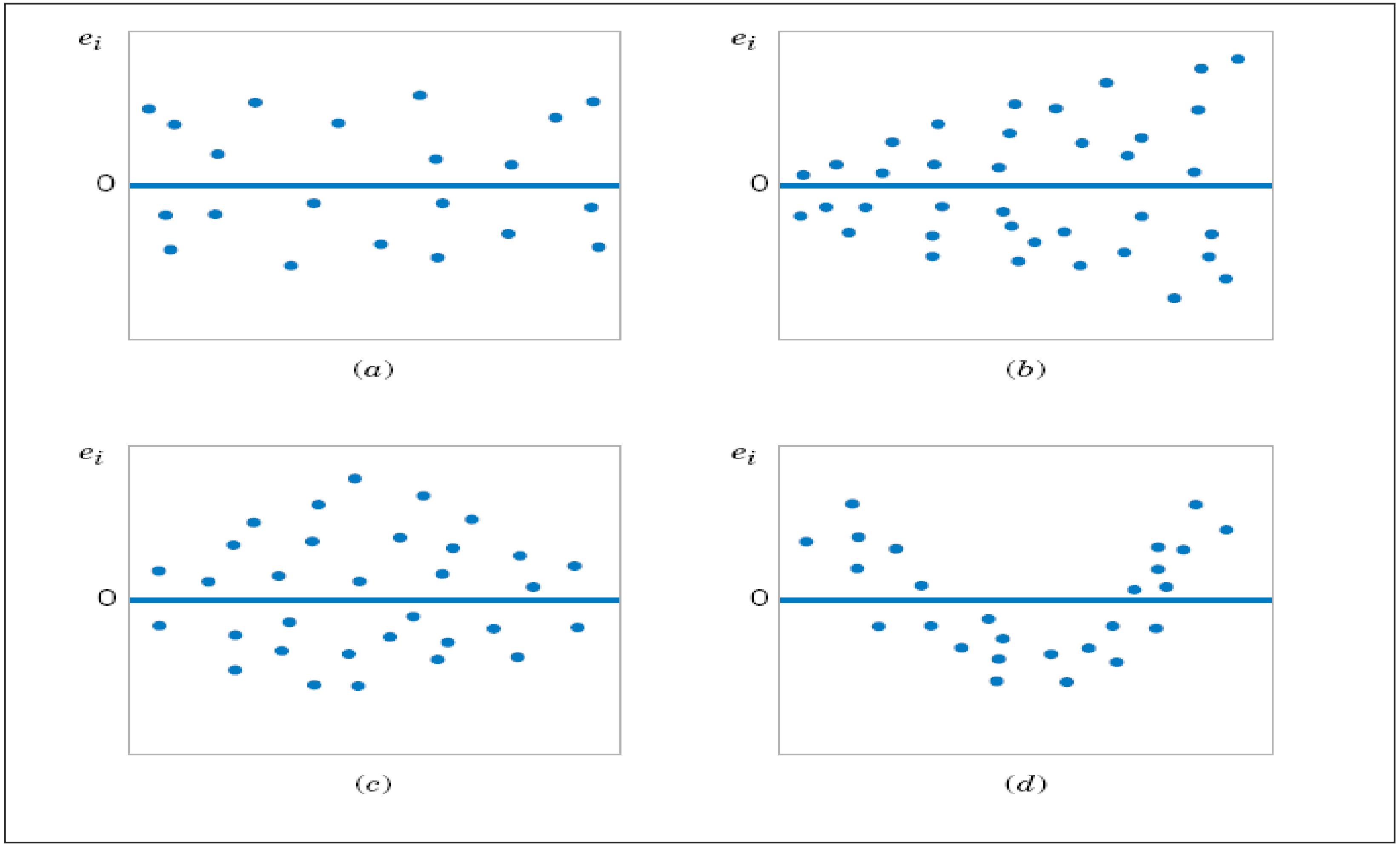
1) 오차의 정규성
2) 오차의 등분산성
3) 오차의 독립성

- 오차는 확률변수로 관찰되지 않는 값이므로, 각 오차에 대응되는 잔차를 관찰한 뒤 잔차들의 분포를 통해 오차에 대한 가정의 적정성을 확인 할 수 있음.

- 각 가정별로, 검정을 통한 방법과 그래프를 통한 시각적인 확인 방법이 가능하며, 시각적 방법을 이용할 경우 다음의 그래프를 이용함.
  - **등분산성 및 독립성** : 잔차산점도
  - **정규성** : 히스토그램 또는 QQplot

모형의 가정에 대한 검토

- 잔차 산점도의 패턴



모형의 가정에 대한 검토

```
plot(
 object,  # 선형회귀모델 클래스 lm의 인스턴스
 which, # 도출할 그래프의 종류. 1~6의 6가지가 가능
        # 1: 잔차산점도, 2: 정규 QQ plot에 해당함.
  ... )
```

```
> plot(m, which=1)
> plot(m, which=2)
```

모형 진단을 위한 그래프