Ch 9. 다중 선형 회귀 모형 (1)

다중선형회귀모형

- **다중선형회귀모형으로의 확장**
  - 다중 선형회귀모형
    - 독립변수가 두 개 이상인 선형회귀모형
    - 여러 개의 독립변수를 이용하면 종속변수의 변화를 더 잘 설명할 수 있을 것임.
    - 자료 $((x_{i1}, x_{i2}, \cdots, x_{ik}), Y_i),\ i = 1, \cdots, n$ 에 다음의 관계식이 성립한다고 가정함.

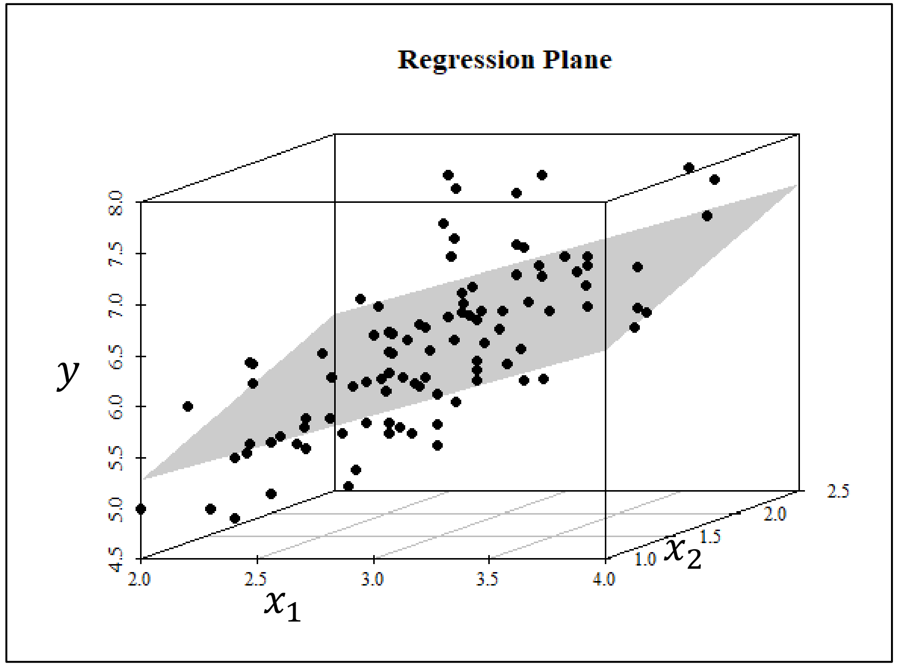
$$Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \quad (i = 1,2,\ldots,n)$$

- 오차항인 $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$는 서로 독립인 확률변수로, $\varepsilon_i \sim N[0, \sigma^2]$: 정규, 등분산, 독립
- 회귀계수 $\alpha, \beta_1, \ldots, \beta_k$, 와 σ^2은 미지인 모수로 상수임.
- $x_{i1}, x_{i2}, \cdots, x_{ik}$는 주어진 상수로 가정함.

- $Y_i \overset{iid}{\sim} Normal[\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik},\ \sigma^2],\ \ i = 1,2,\ldots,n$
  $\rightarrow\ \ E[Y_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$

다중선형회귀모형

- 모형과 회귀 계수에 관한 해석
  - 다중 회귀방정식은 평면($k = 2$) 혹은 초평면 ($k > 2$)을 표현함.
    - $E[Y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$
  - 회귀계수에 관한 해석
    - β_0 (절편, intercept) :
      - $x_1 = x_2 = \cdots = x_k = 0$일 때, $\mathrm{E}[Y]$를 의미.
    - β_j (회귀계수, regression coefficient) :
      - x_j를 제외한 나머지 모든 예측변수들을 상수로 고정시킨 상태에서, x_j의 한 단위 증가에 따른 $E[Y]$의 증분을 의미 ($j = 1, \ldots, k$).



다중선형회귀모형

- **Fidelity Contrafund 펀드에 대한 Fama-French 3 Factor Model 예제**
  - FCNTX 데이터
    - FCNTX 펀드의 월간 초과수익률을 Fama-French의 3 factor(Mkt-RF, SMB, HML)으로 설명하는 다중 선형회귀모형을 설정 (‘FF_FCNTX.csv’ 이용)
      - 기간
        - 1985년 2월부터 2022년 4월까지
      - 변수
        - portfolio : FCNTX 펀드의 수정종가로 구한 월간 수익률
          → pf_excess : FCNTX 월간 초과수익률
        - Mkt-RF (Market - Riskfree) : 시장 리스크 프리미엄, 시장 수익률과 무위험 수익률의 차이
        - SMB (Small Minus Big) : 소형주의 대형주 대비 초과수익률
        - HML (High Minus Low) : 가치주(High B/M)의 성장주(Low B/M) 대비 초과수익률
        - RF : 무위험 수익률

  - 다중 선형 회귀 모형의 설정

$$pf\_excess = \alpha + \beta_1 \cdot (Mkt - RF) + \beta_2 \cdot SMB + \beta_3 \cdot HML + \varepsilon$$

다중선형회귀모형

```
> setwd("E:/DFMBA 경영통계")
> FFdata <- read.csv('FF_FCNTX.csv')
> head( FFdata )
     Date    portfolio Mkt.RF   SMB   HML   RF
1 1985-03  0.008356388  -0.84 -1.07  4.07 0.62
2 1985-04 -0.011049770  -0.96  0.15  3.72 0.72
3 1985-05  0.019553251   5.09 -2.22 -0.94 0.66
4 1985-06  0.006392563   1.27  0.52  0.41 0.55
5 1985-07  0.010889151  -0.74  2.85 -1.60 0.62
6 1985-08  0.018851254  -1.02 -0.31  2.28 0.55
> tail( FFdata )
       Date     portfolio Mkt.RF   SMB   HML   RF
441 2021-11 -0.0009929735  -1.55 -1.36 -0.42 0.00
442 2021-12 -0.0675945301   3.10 -1.60  3.22 0.01
443 2022-01 -0.0019821943  -6.25 -5.93 12.74 0.00
444 2022-02 -0.0609755742  -2.29  2.18  3.09 0.00
445 2022-03  0.0462709939   3.06 -1.61 -1.82 0.00
446 2022-04 -0.1155688753  -9.44 -1.40  6.16 0.00
```

다중선형회귀모형

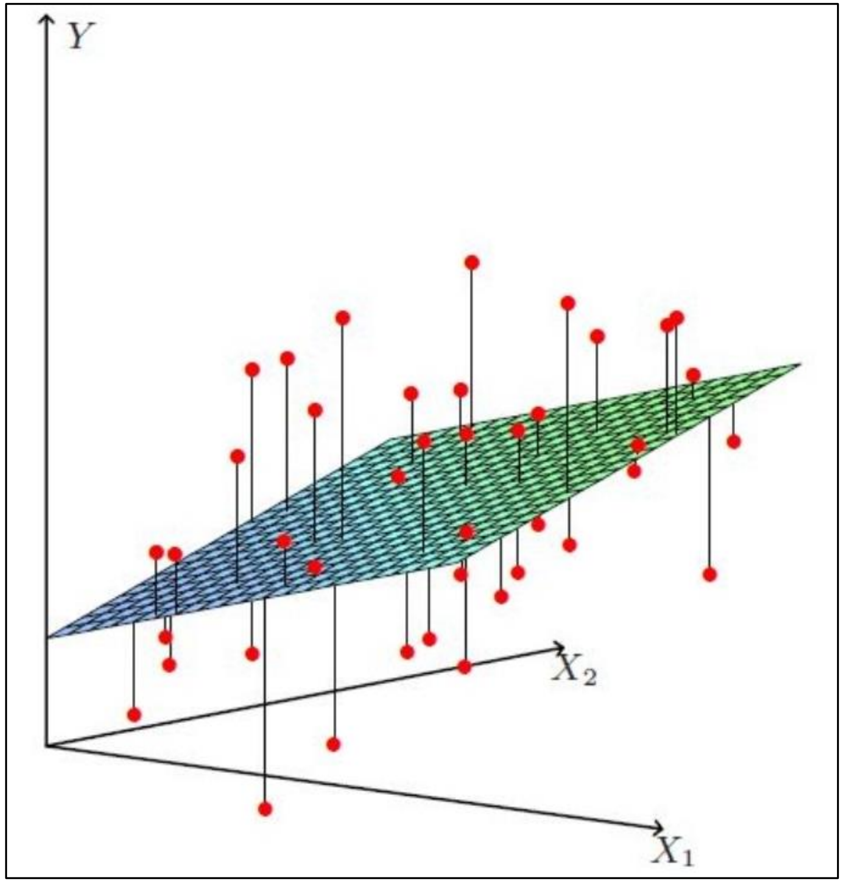
```
> FFdata$pf_excess = FFdata$portfolio - FFdata$RF
> colnames(FFdata) [3] <- "mkt_excess"
> head( FFdata )
     Date    portfolio mkt_excess   SMB   HML   RF  pf_excess
1 1985-03  0.008356388      -0.84 -1.07  4.07 0.62 -0.6116436
2 1985-04 -0.011049770      -0.96  0.15  3.72 0.72 -0.7310498
3 1985-05  0.019553251       5.09 -2.22 -0.94 0.66 -0.6404467
4 1985-06  0.006392563       1.27  0.52  0.41 0.55 -0.5436074
5 1985-07  0.010889151      -0.74  2.85 -1.60 0.62 -0.6091108
6 1985-08  0.018851254      -1.02 -0.31  2.28 0.55 -0.5311487
```

다중선형회귀모형의 추정

- **추정**
  - 회귀계수 $\alpha, \beta_1, \dots, \beta_k$ 의 추정
    - 수직거리 제곱합

      $SS(\alpha, \beta_1, \dots, \beta_k) = \sum_{i=1}^{n}(y_i - \alpha - \beta_1 x_{i1} - \cdots - \beta_k x_{ik})^2$

      이 최소가 되도록 $\alpha, \beta_1, \dots, \beta_k$ 를 추정
    - 최소제곱 추정량 : $\hat{\alpha}, \hat{\beta}_1, \dots, \hat{\beta}_k$



  - y_i의 추정치(predicted values)
    - $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \ldots + \hat{\beta}_k x_{ki}$ $(i = 1, 2, \ldots, n)$
  - 잔차(residuals)
    - $e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}$ $(i = 1, 2, \ldots, n)$

다중선형회귀모형의 추정

- 오차항의 분산 σ^2의 추정

  - 오차에 대응되는 잔차의 변동성을 이용하여 아래와 같이 정의되는 MSE로 추정함.

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

$$MSE = \frac{SSE}{n-k-1} = \frac{\sum_{i=1}^{n} e_i^2}{n-k-1} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k-1} = \frac{\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik})^2}{n-k-1}$$

  - $E[MSE] = \sigma^2$ 임을 보일 수 있음.

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

다중선형회귀모형의 추정

```
> lmfit <- lm( pf_excess ~ mkt_excess + SMB + HML, data=FFdata )
> lmfit

Call:
lm(formula = pf_excess ~ mkt_excess + SMB + HML, data = FFdata)

Coefficients:
(Intercept)   mkt_excess          SMB          HML
 -0.2481499    0.0090902    0.0072790   -0.0006216

> lmfit$coefficients
  (Intercept)    mkt_excess           SMB           HML
-0.2481499107  0.0090902065  0.0072790359 -0.0006215782
```

다중선형회귀모형의 추정

```
> fitresult <- cbind( FFdata$pf_excess, lmfit$fitted.values, lmfit$residuals )
> colnames( fitresult ) <- c("Y", "Yhat", "e")
> fitresult[1:10, ]
            Y       Yhat          e
1  -0.6116436 -0.2661041 -0.3455395
2  -0.7310498 -0.2580969 -0.4729528
3  -0.6404467 -0.2174559 -0.4229908
4  -0.5436074 -0.2330751 -0.3105323
5  -0.6091108 -0.2331369 -0.3759740
6  -0.5311487 -0.2610956 -0.2700531
7  -0.6502207 -0.3017950 -0.3484258
8  -0.5961964 -0.2230648 -0.3731316
9  -0.5571829 -0.1856603 -0.3715225
10 -0.6332772 -0.2156350 -0.4176422
```

다중선형회귀모형의 추정

```
> summary( lmfit )

Call:
lm(formula = pf_excess ~ mkt_excess + SMB + HML, data = FFdata)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.52058 -0.16696  0.01609  0.20982  0.35407 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.2481499  0.0102627 -24.180  < 2e-16 ***
mkt_excess   0.0090902  0.0023346   3.894 0.000114 ***
SMB          0.0072790  0.0034470   2.112 0.035275 *  
HML         -0.0006216  0.0033824  -0.184 0.854280    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2131 on 442 degrees of freedom
Multiple R-squared:  0.05478,	Adjusted R-squared:  0.04837 
F-statistic: 8.539 on 3 and 442 DF,  p-value: 1.596e-05
```

다중선형회귀모형의 추정

```
> predict (lmfit,
+          data.frame( mkt_excess=9.12, SMB=11.5, HML=-6.5))
          1
-0.07749806
```

다중선형회귀모형의 유의성 검정 및 적합도

- **모형의 유의성 검정**
  - 개별 독립변수의 유의성 검정 (t 검정을 활용함)

    개별 독립변수 x_j가 종속변수 Y를 설명하기에 유용한 변수인가에 대한 통계적 추론은 x_j에 대응되는 회귀계수 β_j에 대한 검정을 통해 파악할 수 있음.

    - 가설

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

    - 검정통계량과 표본분포

    귀무가설 H_0 이 사실일 때,

$$T = \frac{\widehat{\beta_j} - 0}{\widehat{S.E.[\widehat{\beta_j}]}} \sim t\,[n-k-1]$$

    - 기각역

    $|T| > t_{\alpha/2,n-k-1}$ 또는 p-value ($P[T > |t_0|] \times 2$) $< \alpha$ 면 귀무가설을 기각

    → x_j는 Y를 설명하는데 유용함.

다중선형회귀모형의 유의성 검정 및 적합도

- 모형의 전반적인 유의성 검정 (F 검정을 활용함)
  - 가설

    $H_0: \beta_j$가 모두 0이다. ($\beta_1 = \beta_2 = \cdots = \beta_k = 0$)

    $H_1: \beta_j$가 모두 0 은 아니다.

  - 검정통계량과 표본분포

    귀무가설 H_0 이 사실일 때,

    $$F = \frac{MSR}{MSE} = \frac{SSR/k}{SSE/(n-k-1)} \sim F[k, n-\mathrm{k}-1]$$

  - 기각역

    $F = \frac{MSR}{MSE} > F_{\alpha,k,n-k-1}$ 면 귀무가설을 기각

다중선형회귀모형의 유의성 검정 및 적합도

- 분산분석표를 이용하여 결과를 정리

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

다중선형회귀모형의 유의성 검정 및 적합도

- **모형의 적합성 검토**
  - 결정계수 R^2 를 적합도 지표로 활용 시 유의할 점
    - 독립변수가 여러 개인 다중회귀모형에서는 결정계수 R^2의 해석에 유의해야 함.
    - 모형에 포함된 독립변수의 수가 많을수록 결정계수 R^2 는 언제나 증가함.

예) Y, X_1, X_2 세 변수에 대한 자료가 주어졌다고 할 때, 아래 두 후보 모형 중 어느 것이 더 적합도가 높은지를 판단함에 있어 R^2 는 그 기준이 될 수 없음.

- 모형 1 : Y 를 X_1 으로만 설명하는 모형

- 모형 2 : Y 를 X_1, X_2 로 설명하는 모형

▶ 언제나 모형2의 R^2 값이 모형1의 R^2 값보다 커지기 때문.

다중선형회귀모형의 유의성 검정 및 적합도

- **수정결정계수 (수정된 $\boldsymbol{R^2}$)**
  - 정의

$$Adjusted \quad R^2 = 1 - \frac{MSE}{MST} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$$

  - 모형에 새로운 독립변수를 추가했을 때, $SST/(n-1)$는 변화가 없으나, $SSE/(n-k-1)$는 추가된 독립변수가 종속변수를 설명하는데 기여하는 바가 큰 경우에만 감소하고, 그렇지 않은 경우는 증가하게 됨.

  - 수정된 R^2는 다중회귀분석에서의 여러 후보모형 간 적합도를 비교하는 지표로 활용될 수 있음.

다중선형회귀모형의 유의성 검정 및 적합도

```
> summary( lmfit )

Call:
lm(formula = pf_excess ~ mkt_excess + SMB + HML, data = FFdata)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.52058 -0.16696  0.01609  0.20982  0.35407 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.2481499  0.0102627 -24.180  < 2e-16 ***
mkt_excess   0.0090902  0.0023346   3.894 0.000114 ***
SMB          0.0072790  0.0034470   2.112 0.035275 *  
HML         -0.0006216  0.0033824  -0.184 0.854280    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2131 on 442 degrees of freedom
Multiple R-squared:  0.05478,	Adjusted R-squared:  0.04837 
F-statistic: 8.539 on 3 and 442 DF,  p-value: 1.596e-05
```