제8강: 다중회귀분석 및 일반회귀분석

금융 통계 및 시계열 분석

TRADE INFORMATIX

2014년 2월 4일

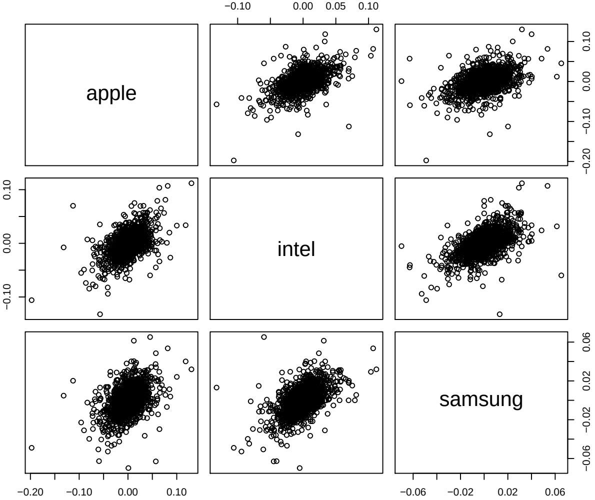
목차

다중회귀분석 (Multiple Linear Regression) 의 예 1

예제 : apple/intel 주가를 이용한 삼성전자 주가 예측

```
> library(quantmod)
> d1 <- getSymbols("NASDAQ:AAPL", src="google", auto.assign=FALSE)
> d2 <- getSymbols("NASDAQ:INTC", src="google", auto.assign=FALSE)
> d3 <- getSymbols("KRX:005930", src="google", auto.assign=FALSE)
> r1 <- lag(ROC(d1[,4]))
> r2 <- lag(ROC(d2[,4]))
> r3 <- log(d3[,1]/lag(d3[,4]))
> r <- as.data.frame(merge(r1,r2,r3))
> names(r) <- c("apple", "intel", "samsung")


> head(r)

                  apple        intel      samsung
2007-01-02           NA           NA           NA
2007-01-03           NA           NA  0.003194891
2007-01-04           NA           NA  0.000000000
2007-01-05  0.021952965  0.039504173  0.001646091
2007-01-08 -0.007146653 -0.003312045 -0.005054771
2007-01-09  0.004926118 -0.004274526  0.008554372
```

apple/intel vs. 삼성전자

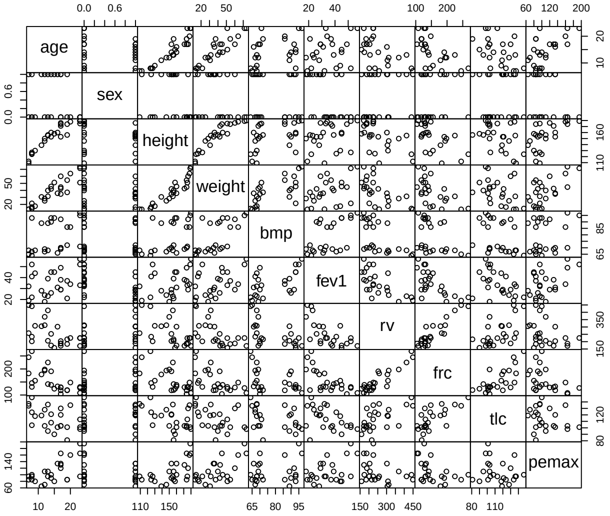


다중회귀분석 (Multiple Linear Regression) 의 예 2

예제 : Cystic fibrosis (낭포성섬유종) 환자

```
> library(ISwR)
> head(cystfibr, 20)

   age sex height weight bmp fev1  rv frc tlc pemax
1    7   0    109   13.1  68   32 258 183 137    95
2    7   1    112   12.9  65   19 449 245 134    85
3    8   0    124   14.1  64   22 441 268 147   100
4    8   1    125   16.2  67   41 234 146 124    85
5    8   0    127   21.5  93   52 202 131 104    95
6    9   0    130   17.5  68   44 308 155 118    80
7   11   1    139   30.7  89   28 305 179 119    65
8   12   1    150   28.4  69   18 369 198 103   110
9   12   0    146   25.1  67   24 312 194 128    70
10  13   1    155   31.5  68   23 413 225 136    95
11  13   0    156   39.9  89   39 206 142  95   110
12  14   1    153   42.1  90   26 253 191 121    90
13  14   0    160   45.6  93   45 174 139 108   100
14  15   1    158   51.2  93   45 158 124  90    80
15  16   1    160   35.9  66   31 302 133 101   134
16  17   1    153   34.8  70   29 204 118 120   134
17  17   0    174   44.7  70   49 187 104 103   165
18  17   1    176   60.1  92   29 188 129 130   120
19  17   0    171   42.6  69   38 172 130 103   130
20  19   1    156   37.2  72   21 216 119  81    85
```

Cystic fibrosis (낭포성섬유종) 환자



다중선형회귀분석 (Multiple Linear Regression)

다중선형회귀분석

- ❑ 반응변수 y의 기대값 μ를 복수의 설명변수 x의 선형 조합으로 설명하려는 시도

$$y \sim N(\mu, \sigma) = N(b_0 + b_1 x_1 + \cdots + b_p x_p, \sigma) \tag{1}$$

OLS (Ordinary Least Squares) Solution

❑ 선형대수방정식

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e} \tag{2}$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} x_{p,1} & \cdots & x_{1,1} & 1 \\ x_{p,2} & \cdots & x_{1,2} & 1 \\ \vdots & & \vdots & \vdots \\ x_{p,n} & \cdots & x_{1,n} & 1 \end{pmatrix}, \quad \beta = \begin{pmatrix} b_p \\ \vdots \\ b_1 \\ b_0 \end{pmatrix}, \quad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}, \tag{3}$$

❑ 오차 제곱의 합을 최소화

$$\hat{\beta} = \arg\min(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) \tag{4}$$

❑ 계수 추정치 $\hat{\beta}$

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{5}$$

Multiple Linear Regression in R

lm

- ❏ `lm(formula, data)`
  - ▶ `formula` : `response  factor 1 + factor 2`
  - ▶ `data` : 자료가 dataframe인 경우 dataframe 이름

```
> m <- lm(samsung ~ apple + intel, data=r)
> summary(m)

Call:
lm(formula = samsung ~ apple + intel, data = r)

Residuals:
      Min        1Q    Median        3Q       Max
-0.068899 -0.006176  0.000049  0.005895  0.075985

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0003728  0.0002526   1.476     0.14
apple       0.1094247  0.0129196   8.470   <2e-16 ***
intel       0.2699130  0.0147360  18.317   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01045 on 1713 degrees of freedom
  (126 observations deleted due to missingness)
Multiple R-squared: 0.3111, Adjusted R-squared: 0.3103
F-statistic: 386.8 on 2 and 1713 DF,  p-value: < 2.2e-16
```

Confidence Interval & Prediction Interval in R

lm

- `predict(model, newdata, interval, level=0.95)`
  - `model` : `lm`명령으로 계산한 모델 오브젝트
  - `newdata` : column name이 모델 변수이름인 데이터프레임
  - `interval` : `'confidence'`, `'prediction'`
  - `level` : $1 - \alpha$

```
> newdata <- data.frame(apple=0.02, intel=0.01)
> predict(m, newdata, interval="confidence")
          fit         lwr         upr
1 0.005260457 0.00461616 0.005904753
> predict(m, newdata, interval="prediction")
          fit         lwr         upr
1 0.005260457 -0.01524662 0.02576753
```

Relationships in Sum of Squares

- Total Sum of Squares (Total Variations)

$$\text{TSS} = \sum (y_i - \bar{y})^2 \tag{6}$$

- Residual Sum of Squares (Unexplained Variations)

$$\text{RSS} = \sum (y_i - \hat{y}_i)^2 \tag{7}$$

- Regression Sum of Squares (Exxplained variations)

$$\text{RegSS} = \sum (\hat{y}_i - \bar{y})^2 \tag{8}$$

- Total Variation = Explained Variation + Unexplained Variation

$$\text{TSS} = \text{RegSS} + \text{RSS} \tag{9}$$

수정결정계수 (modified coefficient of determination)

- 결정계수 : 추정된 선형회귀모형이 실제 자료를 설명할 수 있는 능력의 척도

$$R^2 = \frac{\text{RegSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

- 수정결정계수 : 팩터수 증가에 따른 자동적인 결정계수 증가 방지

$$R^2_{\text{adj}} = 1 - \frac{\text{RSS}/(n-p-1)}{\text{TSS}/(n-1)}$$

- F-test : 다음 test-statistics는 자유도 $(p, n-p-1)$의 F 분포

$$F = \frac{\text{RegSS}/p}{\text{RSS}/(n-p-1)}$$

$$H_0 : b_1 = 0 \ \text{ against } \ H_a : b_1 \neq 0$$

ANOVA for Multiple Linear Regression in R

lm

- ❑ `anova(model)`
  - ▶ `model` : `lm`명령으로 계산한 모델 오브젝트

```
> anova(m)
Analysis of Variance Table

Response: samsung
            Df   Sum Sq  Mean Sq F value    Pr(>F)    
apple        1 0.047855 0.047855  438.18 < 2.2e-16 ***
intel        1 0.036640 0.036640  335.50 < 2.2e-16 ***
Residuals 1713 0.187079 0.000109                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

다중공선성 (Multicollinearity)

다중공선성 (Multicollinearity)

- ❏ 설명변수들 간에 강한 상관관계가 있는 경우
  - ▶ 회귀분석의 기본 가정을 무시한 결과
  - ▶ 특정 데이터 샘플에 대한 설명력은 강하지만 데이터 샘플이 달라지면 회귀분석에 의한 모형 계수가 크게 변함

```
> cor(r, use="complete.obs")

            apple     intel   samsung
apple   1.0000000 0.5163146 0.4197757
intel   0.5163146 1.0000000 0.5313022
samsung 0.4197757 0.5313022 1.0000000

> lm(samsung ~ apple + intel, data=r[1:900,])

Call:
lm(formula = samsung ~ apple + intel, data = r[1:900, ])

Coefficients:
(Intercept)        apple        intel
  0.0001591    0.1118423    0.2660966

> lm(samsung ~ apple + intel, data=r[-(1:900),])

Call:
lm(formula = samsung ~ apple + intel, data = r[-(1:900), ])

Coefficients:
(Intercept)        apple        intel
  0.0005761    0.1070460    0.2771079
```

Multiple Linear Regression in R : Example 2

```
> m1 <- lm(pemax~age+sex+height+weight+bmp+fev1+rv+frc+tlc, data=cystfibr)
> summary(m1)

Call:
lm(formula = pemax ~ age + sex + height + weight + bmp + fev1 +
    rv + frc + tlc, data = cystfibr)

Residuals:
    Min      1Q  Median      3Q     Max
-37.338 -11.532   1.081  13.386  33.405

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 176.0582   225.8912   0.779    0.448
age          -2.5420     4.8017  -0.529    0.604
sex          -3.7368    15.4598  -0.242    0.812
height       -0.4463     0.9034  -0.494    0.628
weight        2.9928     2.0080   1.490    0.157
bmp          -1.7449     1.1552  -1.510    0.152
fev1          1.0807     1.0809   1.000    0.333
rv            0.1970     0.1962   1.004    0.331
frc          -0.3084     0.4924  -0.626    0.540
tlc           0.1886     0.4997   0.377    0.711

Residual standard error: 25.47 on 15 degrees of freedom
Multiple R-squared: 0.6373, Adjusted R-squared: 0.4197
F-statistic: 2.929 on 9 and 15 DF,  p-value: 0.03195
```

Dummy Variable

- Dummy Variable
  - 설명변수가 category값인 경우 숫자 0, 1로 치환
  - Single Dummy Variable의 경우 ANOVA 분석
- Analysis of Covariance
  - Multiple Regression에서 Dummy Variable이 있는 경우

$$y = b_0 + b_1 x + b_2 d + e$$

  - interaction 항을 이용하여 Slope와 intercept가 다른 두개의 모형으로 표현

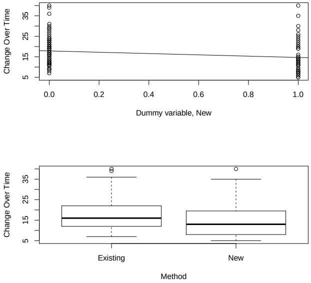
$$\begin{aligned} y &= b_0 + b_1 x + b_2 d + b_3 (d \cdot x) + e \\ &= \begin{cases} b_0 + b_1 x + e & \text{if } d = 0 \\ (b_0 + b_2) + (b_1 + b_3) x + e & \text{if } d = 1 \end{cases} \end{aligned}$$

Single Dummy Variable 예

```
> url <- "http://www.stat.tamu.edu/~sheather/book/docs/datasets/changeover_times.txt"
> changeover_times <- read.table(url, header=TRUE)
> head(changeover_times)

    Method Changeover New
1 Existing         19   0
2 Existing         24   0
3 Existing         39   0
4 Existing         12   0
5 Existing         29   0
6 Existing         19   0

> m2 <- lm(Changeover~New, data=changeover_times)
> summary(m2)

Call:
lm(formula = Changeover ~ New, data = changeover_times)

Residuals:
    Min      1Q  Median      3Q     Max
-10.861  -5.861  -1.861   4.312  25.312

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.8611     0.8905  20.058   <2e-16 ***
New          -3.1736     1.4080  -2.254    0.026 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.556 on 118 degrees of freedom
Multiple R-squared: 0.04128, Adjusted R-squared: 0.03315
F-statistic: 5.081 on 1 and 118 DF,  p-value: 0.02604
```

Single Dummy Variable 예

```
> attach(changeover_times)
> par(mfrow=c(2,1))
> plot(New, Changeover,
+      xlab="Dummy variable, New",
+      ylab="Change Over Time")
> abline(lsfit(New, Changeover))
> boxplot(Changeover ~ Method,
+         xlab="Method",
+         ylab="Change Over Time")
> detach(changeover_times)
```

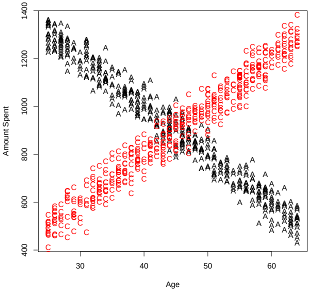


Analysis of Covariance 예

```
> url <- "http://www.stat.tamu.edu/~sheather/book/docs/datasets/travel.txt"
> travel <- read.table(url, header=TRUE)
> head(travel)

  Amount Age Segment C
1    997  44       A 0
2    997  43       A 0
3    951  41       A 0
4    649  59       A 0
5   1265  25       A 0
6   1059  38       A 0

> attach(travel)
> mfull <- lm(Amount ~ Age + C + C:Age)
> summary(mfull)

Call:
lm(formula = Amount ~ Age + C + C:Age)

Residuals:
     Min       1Q   Median       3Q      Max
-143.298  -30.541   -0.034   31.108  130.743

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1814.5445     8.6011   211.0   <2e-16 ***
Age           -20.3175     0.1878  -108.2   <2e-16 ***
C           -1821.2337    12.5736  -144.8   <2e-16 ***
Age:C          40.4461     0.2724   148.5   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 47.63 on 921 degrees of freedom
Multiple R-squared: 0.9601, Adjusted R-squared: 0.9599
F-statistic:  7379 on 3 and 921 DF,  p-value: < 2.2e-16

> detach(travel)
```

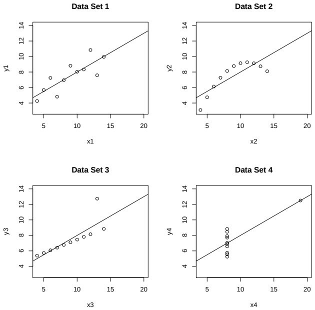
Analysis of Covariance 예

```
> attach(travel)
> par(mfrow=c(1,1))
> plot(Age[C==0], Amount[C==0],
+      pch=c("A"), col=c("black"),
+      xlab="Age",
+      ylab="Amount Spent")
> points(Age[C==1], Amount[C==1],
+        pch=c("C"), col=c("red"))
> detach(travel)
```



선형회귀의 문제점

```
> url <-
+   paste("http://www.stat.tamu.edu",
+         "/~sheather/book/docs/",
+         "datasets/anscombe.txt",
+         sep="")
> anscombe <- read.table(url,
+                        header=TRUE)
> attach(anscombe)
> par(mfrow=c(2,2))
> xlim <- c(4,20); ylim <- c(3,14)
> plot(x1,y1,xlim=xlim,ylim=ylim,
+      main="Data Set 1")
> abline(lsfit(x1,y1))
> plot(x2,y2,xlim=xlim,ylim=ylim,
+      main="Data Set 2")
> abline(lsfit(x2,y2))
> plot(x3,y3,xlim=xlim,ylim=ylim,
+      main="Data Set 3")
> abline(lsfit(x3,y3))
> plot(x4,y4,xlim=xlim,ylim=ylim,
+      main="Data Set 4")
> abline(lsfit(x4,y4))
> detach(anscombe)
```



선형회귀의 문제점

```
> attach(anscombe)
> summary(m1 <- lm(y1~x1))

Call:
lm(formula = y1 ~ x1)

Residuals:
     Min       1Q   Median       3Q      Max
-1.92127 -0.45577 -0.04136  0.70941  1.83882

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0001     1.1247   2.667  0.02573 *
x1            0.5001     0.1179   4.241  0.00217 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.237 on 9 degrees of freedom
Multiple R-squared: 0.6665, Adjusted R-squared: 0.6295
F-statistic: 17.99 on 1 and 9 DF,  p-value: 0.00217

> summary(m2 <- lm(y2~x2))

Call:
lm(formula = y2 ~ x2)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9009 -0.7609  0.1291  0.9491  1.2691

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.001      1.125   2.667  0.02576 *
x2             0.500      0.118   4.239  0.00218 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.237 on 9 degrees of freedom
Multiple R-squared: 0.6662, Adjusted R-squared: 0.6292
F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002179

> detach(anscombe)
```

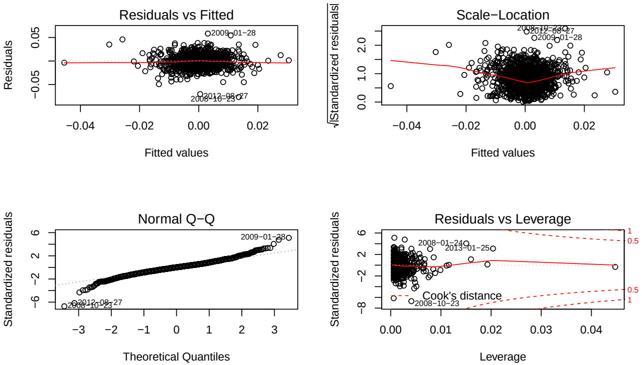
```
> attach(anscombe)
> summary(m3 <- lm(y3~x3))

Call:
lm(formula = y3 ~ x3)

Residuals:
    Min      1Q  Median      3Q     Max
-1.1586 -0.6146 -0.2303  0.1540  3.2411

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0025     1.1245   2.670  0.02562 *
x3            0.4997     0.1179   4.239  0.00218 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.236 on 9 degrees of freedom
Multiple R-squared: 0.6663, Adjusted R-squared: 0.6292
F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002176

> summary(m4 <- lm(y4~x4))

Call:
lm(formula = y4 ~ x4)

Residuals:
   Min     1Q Median     3Q    Max
-1.751 -0.831  0.000  0.809  1.839

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0017     1.1239   2.671  0.02559 *
x4            0.4999     0.1178   4.243  0.00216 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.236 on 9 degrees of freedom
Multiple R-squared: 0.6667, Adjusted R-squared: 0.6297
F-statistic:    18 on 1 and 9 DF,  p-value: 0.002165

> detach(anscombe)
```

선형회귀 결과진단 (Diagnostics)

- 표준잔차 (Standardized Residuals)
  - 올바른 모형인 경우 표준잔차는 평균이 0인 Normal 분포
  - 올바른 모형인 경우 표준잔차의 분산은 fitted value와 상관없이 상수
  - Log-likelihood
- Leverage Points
  - 어떤 샘플 포인트가 분석결과에 가장 큰 영향력을 미치는지 파악
- Outliers
  - 어떤 샘플 포인트 가장 설명이 되지 않는지를 표시
- Added-Variable Plot
  - 특정 팩터를 제외하고 분석한 회귀분석 잔차를 그 팩터로 회귀분석

- ❑ plot(model)
  - ▶ model : lm()의 결과로 나온 모델 오브젝트

```
> layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
> plot(lm(samsung ~ apple, data=r))
```



표준잔차 (Standardized Residuals) 분석

- 잔차 (Residuals)
  - 실제 종속변수값과 모델 종속변수값의 차이

$$e_i = y_i - \hat{y}_i$$

- 표준잔차 (Standardized Residuals)
  - 잔차를 잔차 표준 편차로 정규화

$$r_i = \frac{e_i}{\mathrm{Var}(e_i)}$$

- Residuals vs Fitted
  - Fitted values 값에 따른 잔차의 평균과 분산 값 표시
  - Fitted values 값에 따른 평균이나 분산값의 변화가 적으면 적합
- Scale vs Location
  - Scale : 표준잔차의 제곱근
  - 잔차의 부호를 생략하고 크기만 절대적 비교

표준잔차 (Standardized Residuals) 분포 분석

- ❏ QQ plot
  - ► 표준잔차의 Normality를 눈으로 확인
- ❏ Sharpiro-Wilk test
  - ► 표준잔차의 Normality를 수치적으로 테스트
- ❏ Log-Likelihood
  - ► 잔차가 동일한 normal 분포로부터 나왔을 경우의 Log-Likelihood 값을 계산
  - ► 두 개의 다른 모델 중 선택하는 경우 Log-Likelihood가 높은 모델 선택

Leverage Points

- Leverage Point
  - 모형 예측 결과와 크게 영향력을 미치는 샘플 포인트

$$\hat{\mathbf{y}} = \mathbf{X}\beta = \left(\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)\mathbf{y} = \mathbf{H}\mathbf{y}$$

  - hat matrix $\hat{\mathbf{y}}$의 (i,j) 번째 원소를 $h_{i,j}$ 라고 하면

$$\hat{y}_i \quad = \quad h_{i,i}y_i + \sum_{j \neq i} h_{i,j}y_j$$

$$h_{i,j} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{y})}{\sum_k (x_k - \bar{x})^2}$$

  - 다른 샘플 포인트와 멀리 떨어져 있을 수록 leverage 증가
  - `http://www.rob-mcculloch.org/teachingApplets/Leverage/index.html`

Outliers

❑ 평균 레버리지

$$\text{Average}(h_{i,i}) = \frac{2}{n}$$

❑ Rule of thumb for finding high leverage points

$$h_{ii} > 2 \cdot \text{Average}(h_{i,j}) = \frac{4}{n}$$

❑ 아웃라이어 (Outlier)

- ▶ 모형 예측 결과와 크게 다른 값을 가지는 샘플 포인트
- ▶ 일반적으로 (rule of thumb), 표준잔차의 크기가 2보다 크면 아웃라이어

❑ Bad Leverage Point

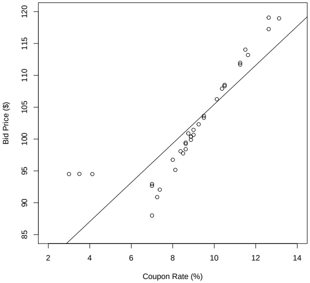
- ▶ Outlier인 Leverage Point

Leverage & Outlier 예

```
> url <- "http://www.stat.tamu.edu/~sheather/book/docs/datasets/bonds.txt"
> bonds <- read.table(url, header=TRUE)
> head(bonds)

  Case CouponRate BidPrice
1    1      7.000    92.94
2    2      9.000   101.44
3    3      7.000    92.66
4    4      4.125    94.50
5    5     13.125   118.94
6    6      8.000    96.75

> m1 <- lm(BidPrice~CouponRate, data=bonds)
> summary(m1)

Call:
lm(formula = BidPrice ~ CouponRate, data = bonds)

Residuals:
   Min     1Q Median     3Q    Max
-8.249 -2.470 -0.838  2.550 10.515

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  74.7866     2.8267  26.458  < 2e-16 ***
CouponRate    3.0661     0.3068   9.994 1.64e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.175 on 33 degrees of freedom
Multiple R-squared: 0.7516, Adjusted R-squared: 0.7441
F-statistic: 99.87 on 1 and 33 DF,  p-value: 1.645e-11
```

Leverage & Outlier 예

```
> attach(bonds)
> par(mfrow=c(1,1))
> plot(CouponRate, BidPrice,
+      xlab="Coupon Rate (%)",
+      ylab="Bid Price ($)",
+      xlim=c(2,14),
+      ylim=c(85,120))
> abline(lsfit(CouponRate,BidPrice))
> detach(bonds)
```

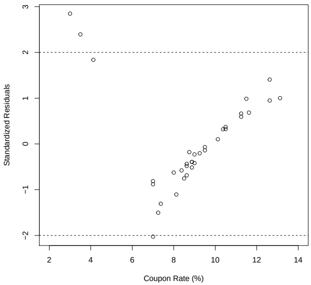


Leverage & Outlier 예

```
> attach(bonds)
> leverage1 <- hatvalues(m1)
> StanRes1 <- rstandard(m1)
> residual1 <- m1$residuals
> lt <- cbind(Case, CouponRate, BidPrice,
+             round(leverage1,3), round(residual1,3), round(StanRes1,3))
> lt[c(1:6, 10:15, 33:35),]

   Case CouponRate BidPrice
1     1      7.000    92.94 0.049 -3.309 -0.812
2     2      9.000   101.44 0.029 -0.941 -0.229
3     3      7.000    92.66 0.049 -3.589 -0.881
4     4      4.125    94.50 0.153  7.066  1.838
5     5     13.125   118.94 0.124  3.911  1.001
6     6      8.000    96.75 0.033 -2.565 -0.625
10   10     10.125   106.25 0.036  0.419  0.102
11   11     11.625   113.19 0.068  2.760  0.685
12   12      8.625    99.44 0.029 -1.792 -0.435
13   13      3.000    94.50 0.218 10.515  2.848
14   14     10.500   108.31 0.042  1.329  0.325
15   15     11.250   111.69 0.058  2.410  0.595
33   33      9.250   102.31 0.029 -0.838 -0.204
34   34      7.000    88.00 0.049 -8.249 -2.025
35   35      3.500    94.53 0.187  9.012  2.394

> detach(bonds)
```

Leverage & Outlier 예

```
> attach(bonds)
> plot(CouponRate,StanRes1,
+      xlab="Coupon Rate (%)",
+      ylab="Standardized Residuals",
+      xlim=c(2,14))
> abline(h=2,lty=2)
> abline(h=-2,lty=2)
> # identify(CouponRate,StanRes1,Case)
> detach(bonds)
```

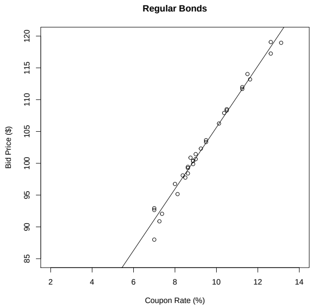


Leverage & Outlier 예

```
> summary((m2 <- update(m1, subset=(1:35)[-c(4,13,35)])))
Call:
lm(formula = BidPrice ~ CouponRate, data = bonds, subset = (1:35)[-c(4,
    13, 35)])

Residuals:
    Min      1Q  Median      3Q     Max
-3.1301 -0.3789  0.2240  0.4576  1.8099

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  57.2932     1.0358   55.31   <2e-16 ***
CouponRate    4.8338     0.1082   44.67   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.024 on 30 degrees of freedom
Multiple R-squared: 0.9852, Adjusted R-squared: 0.9847
F-statistic:  1996 on 1 and 30 DF,  p-value: < 2.2e-16
```

Leverage & Outlier 예

```
> attach(bonds)
> plot(CouponRate[-c(4,13,35)],
+      BidPrice[-c(4,13,35)],
+      main="Regular Bonds",
+      xlab="Coupon Rate (%)",
+      ylab="Bid Price ($)",
+      xlim=c(2,14),
+      ylim=c(85,120))
> abline(m2)
> detach(bonds)
```



일반선형회귀 (GLM: Generalized Linear Models)

❏ 반응변수 분포가 정상분포가 아닌 경우도 사용가능

$$f(y;\mu) = \exp\left(\left(\mu - \gamma(\mu)\right)/(\phi/A) + \tau(y,\phi/A)\right) \tag{10}$$

❏ 반응변수 분포의 평균이 설명변수의 단순 선형 함수가 아니라 일반함수

$$\mu = m(b_0 + b_1 x_1 + \cdots + b_p x_p) \tag{11}$$

❏ link function $\eta = m^{-1}$: 분포평균 추정함수 m 의 역함수

로지스틱 회귀 (Logistic Regression)

- 반응변수 분포가 Binomial 분포이고 성공확률 p 가 단일 팩터 x에 의존하는 경우, m번 시도에 대한 성공횟수 Y 의 분포는

$$Y|x \sim \text{Binomial}(m, p) \tag{12}$$

- 성공확률 p를 다음과 같은 logit 함수를 이용하여 모형화하면 GLM 사용 가능

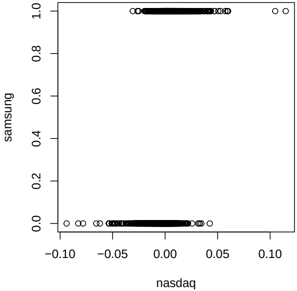
$$p = \frac{1}{1 + \exp(-(b_0 + b_1 x))} \tag{13}$$

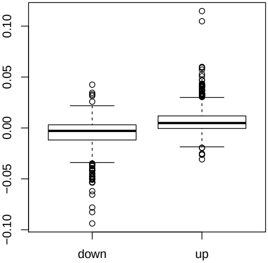
로지스틱 회귀의 예

예제 : NASDAQ을 이용한 삼성전자 상승/하락 예측

```
> library(quantmod)
> d1 <- getSymbols("NASDAQ:QQQ", src="google", auto.assign=FALSE)
> d2 <- getSymbols("KRX:005930", src="google", auto.assign=FALSE)
> r1 <- lag(ROC(d1[,4]))
> r2 <- log(d2[,1]/lag(d2[,4])) > 0
> r <- as.data.frame(merge(r1,r2))
> names(r) <- c("nasdaq", "samsung")


> head(r, 10)

                 nasdaq samsung
2007-01-02           NA      NA
2007-01-03           NA       1
2007-01-04           NA       0
2007-01-05  0.0187863486       1
2007-01-08 -0.0047776226       0
2007-01-09  0.0006839166       1
2007-01-10  0.0050011470       0
2007-01-11  0.0117224066       1
2007-01-12  0.0102565002       1
2007-01-15           NA       1
```

NASDAQ vs. 삼성전자 상승/하락





Maximum Likelihood Estimation

- log-likelihood
  - 실제 데이터 샘플이 나올 확률 함수의 log 값

$$logL = \log \prod_{i=1}^{n} P(Y = y_i | x + i) = \log \prod_{i=1}^{n} \binom{n}{k} p(x_i)^{y_i} (1 - p(x_i))^{1-y_i} \tag{14}$$

- MLE (Maximum Likelihood Estimation)
  - log-likelihood 를 최대화하는 $p(x)$ 의 계수를 찾는 방법
  - 반복적인 (iterative) 비선형 최적화를 사용

Deviance

- ❏ 선형회귀분석에서의 잔차 (residuals) 에 해당하는 개념
- ❏ 비선형최적화를 통해 찾아낸 모델 M 과 saturated 모델 S 의 log-likelihood의 차이
- ❏ saturated 모델 : 단일 샘플에 대해 계산된 모델

$$
\begin{aligned}
G^2 &= 2(logL_S - logL_M) \\
&= 2\sum_{i=1}^{n}(y_i \log y_i + (1-y_i)\log(1-y_i)) - \\
& 2\sum_{i=1}^{n}(y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i))
\end{aligned}
$$

- ❏ logistic 문제에 한해 saturated 모델의 log-likelihood 는 0

$$G^2 = 2\sum_{i=1}^{n}(y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)) \tag{15}$$

- ❏ deviance 는 approximately chi-squared 분포

Logistic Regression in R

glm

- `glm(formula, family, data)`
  - `family` : logistic의 경우 `binomial(link="logit")`

```
> m <- glm(samsung ~ nasdaq, binomial(link="logit"), data=r)
> summary(m)

Call:
glm(formula = samsung ~ nasdaq, family = binomial(link = "logit"),
    data = r)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.72504  -1.04762   0.01183   1.02517   2.38039

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.05633    0.05397  -1.044    0.297
nasdaq      88.08439    5.73218  15.367   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2378.9  on 1715  degrees of freedom
Residual deviance: 2006.5  on 1714  degrees of freedom
  (126 observations deleted due to missingness)
AIC: 2010.5

Number of Fisher Scoring iterations: 5
```

로지스틱 분석 결과

```
> plot(samsung ~ nasdaq, data=r)
> curve(predict(m,data.frame(nasdaq=x), type="resp"), add=TRUE)
> points(m$model$nasdaq, m$fitted, col="red", add=TRUE)
```