R을 활용한 데이터분석

호서대학교 빅데이터경영공학부 연규필 (kpyeon1@hoseo.edu)

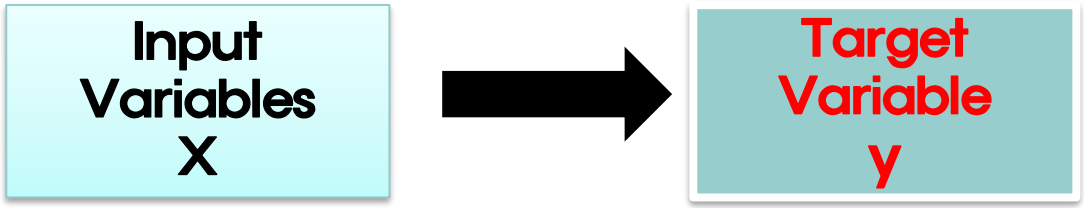
회귀모형

1. 통계적 학습모형
2. 단순회귀분석
3. 다중회귀분석

1. 통계적 학습모형

■ 지도학습모형 (Supervised learning)

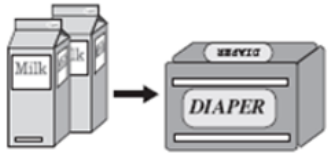
- 목표변수(target variable)가 정해져 있으며, 목표변수를 예측하는 것이 목적인 학습모형

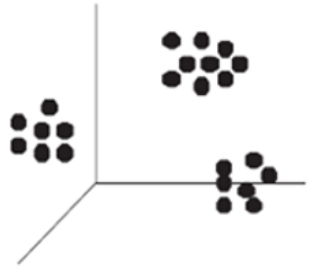


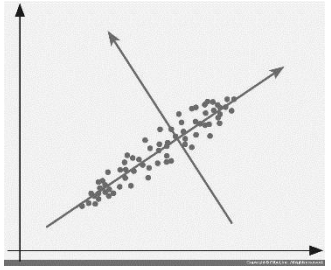
Target function $y = f(x_1, x_2, \cdots, x_p)$

■ 비지도학습모형 (Unsupervised learning)

- 목표변수가 정해져 있지 않음
- 자료 구조를 파악하거나 차원 축소를 위한 목적인 경우가 많음







1. 통계적 학습모형

■ 지도학습모형의 종류

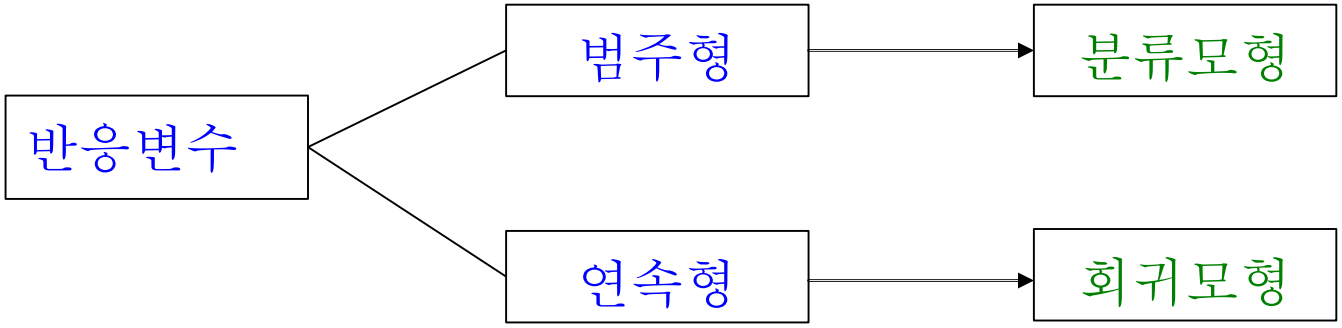
- 판별분석 (Discrimination Analysis)
- 일반화선형모형 (GLM, Generalized Linear Model)
  - 다중선형회귀분석 (Multiple Regression Analysis)
  - 로지스틱 회귀분석 (Logistic Regression)
- 사례기반추론 (Case-Based Reasoning, k-Nearest Neighbor)
- 신경망 (Artificial Neural Network)
- 서포트벡터머신 (Support Vector Machine)
- 앙상블 (Ensemble)

■ 비지도학습모형의 종류

- 연관성규칙발견 (Association Rule Discovery, Market Basket)
- 군집분석 (Heirarchical Clustering, k-Means Clustering)
- 특이치 탐색 (Anomaly Detection, Outlier Detection, Novelty Detection)
- 인자분석(Factor Analysis), 주성분분석(Principal Component Analysis)
- SOM (Self Organizing Map, Kohonen Network)

1. 통계적 학습모형

■ 지도학습모형의 분류



입력변수 / 목표변수 / 예측확률

Obs.	Sex	Age	Region	y	P(y = 1)
1	F	18	A	1	0.75
2	M	25	D	0	0.12
3	F	67	D	1	0.93
4	F	43	B	1	0.53
5	F	28	A	0	0.15
6	M	53	C	0	0.31
7	F	42	A	0	0.12

$$\hat{P}(y=1)=\frac{\exp(a+b_1x_1+b_2x_2+\ldots+b_px_p)}{1+\exp(a+b_1x_1+b_2x_2+\ldots+b_px_p)}$$

(예측모형: 로지스틱 회귀분석)

입력변수 / 목표변수 / 예측값

Obs.	Sex	Age	Region	y	$\hat{y}$
1	F	18	A	125	120
2	M	25	D	35	38
3	F	67	D	150	147
4	F	43	B	45	53
5	F	28	A	13	15
6	M	53	C	38	36
7	F	42	A	20	21

$$\hat{y}=a+b_1x_1+b_2x_2+\ldots+b_px_p$$

(예측모형: 선형 회귀분석)

1. 통계적 학습모형

■ 지도학습모형을 위한 통계적 방법론들

회귀모형

- Multiple Linear Regression
- Regression Tree
- Artificial Neural Network
- Support Vector Regression
- Ensemble

분류모형

- Discriminant Analysis
- Logistic Regression
- Classification Tree
- Artificial Neural Network
- Naïve Bayes
- Support Vector Classification
- Nearest Neighbor
- Ensemble

■ 지도학습모형 구축을 위한 데이터 구분

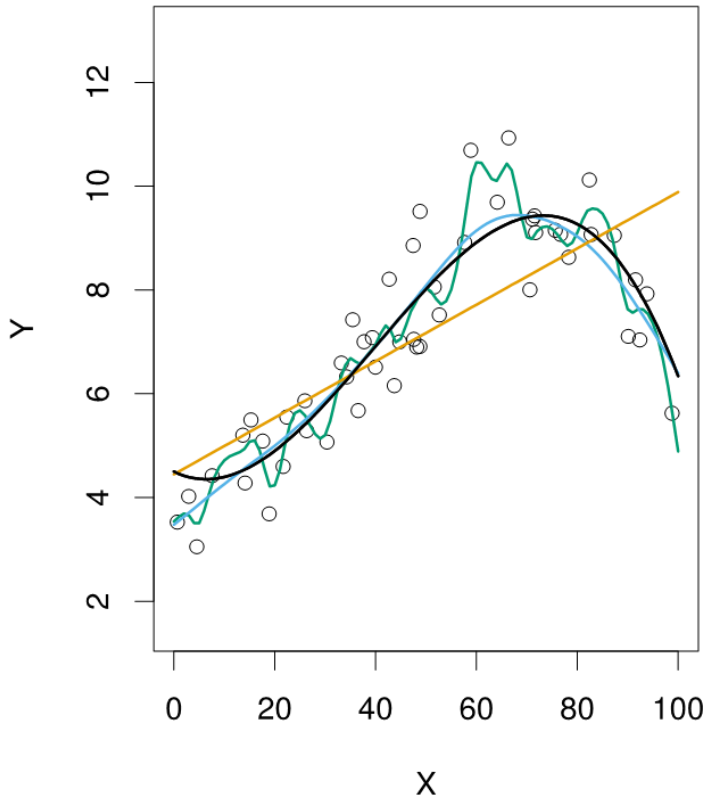
- 분석용 데이터(Training Data):
  - 데이터를 분석(학습)하여 모형을 만드는 데 직접적으로 사용되는 데이터

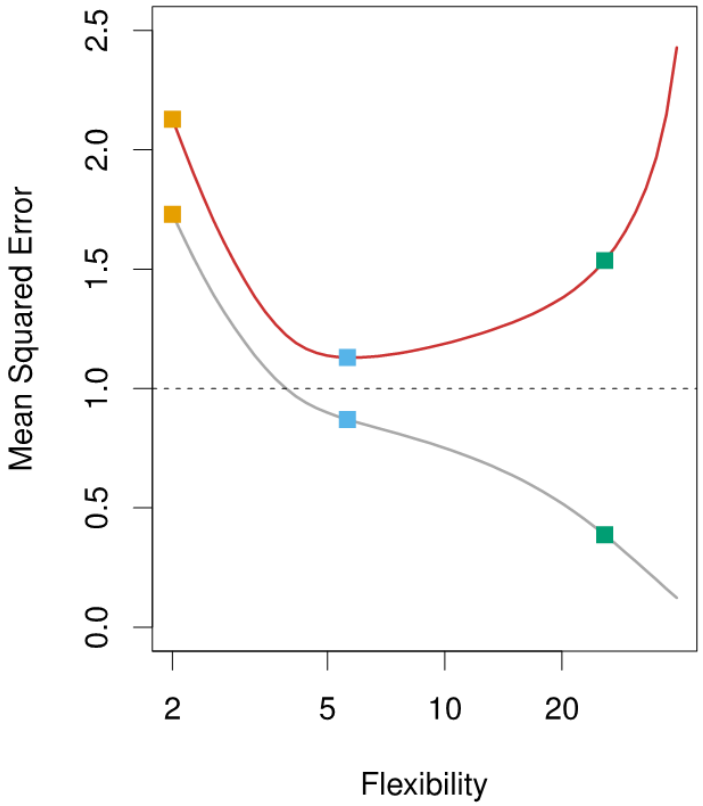
- 평가용 데이터(Validation Data):
  - 모형의 성능을 감독하고 개선하기 위하여 간접적으로 사용되는 데이터
  - 모형선택 또는 최적모형 결정에 사용되는 데이터로서 과적합을 방지하는데 사용

- 검증용 데이터(Test Data):
  - 모형의 생성에 전혀 사용되지 않으며, 일반화의 검토를 위해 남겨 두는 데이터

■ 과적합이란…

- 복잡한 모형일수록 Training data에서의 정확도는 높으나, 새로운 데이터에 대해서는 정확도가 오히려 떨어지는 현상

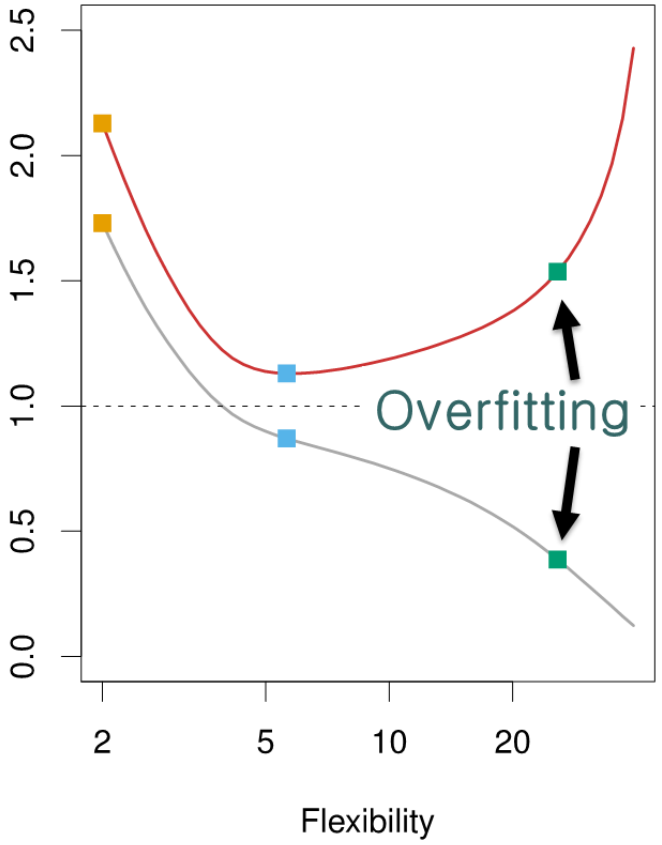




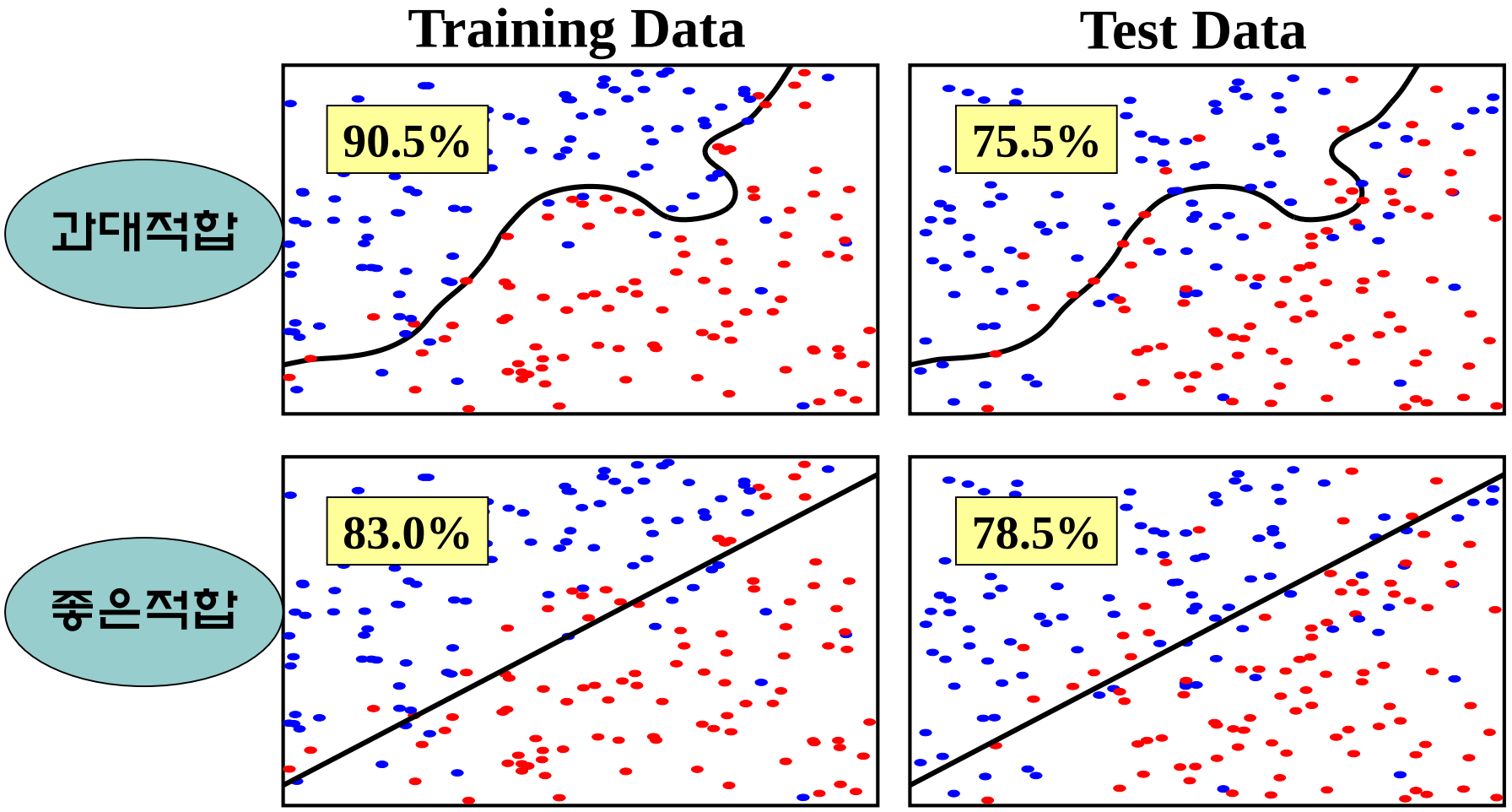
■ 과적합이란…

- 모형의 유연성(flexibility)이 증가함에 따라:
  - Training MSE: 지속적인 감소
  - Test MSE: U-shape

- 데이터와 분류방법에 상관없이 항상 발생하는 현상



■ 분류모형에서의 과적합



■ 단순 회귀분석 (Simple Regression Analysis)

- 단순선형회귀 모형

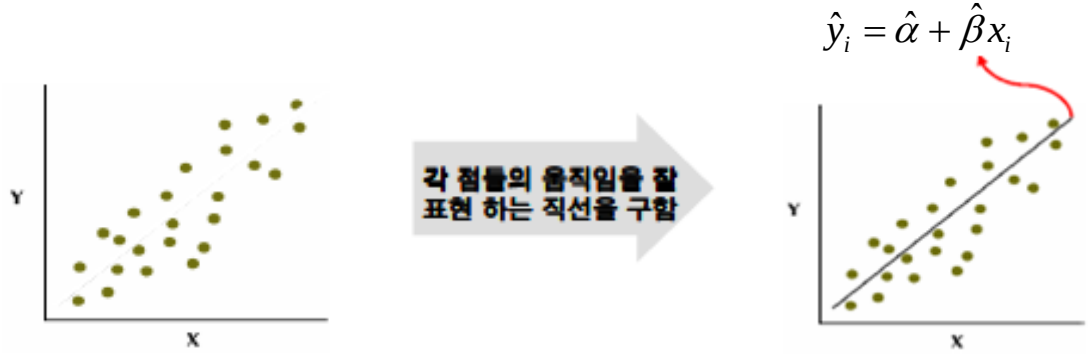
  독립변수의 정해진 값 $x_1, \cdots, x_n$ 에서 측정되는 종속변수 $Y_1, \cdots, Y_n$ 에 대하여 다음의 관계식이 성립한다고 가정하자.

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

  여기에서 $\varepsilon_1, \cdots, \varepsilon_n$은 서로 독립이며 $\varepsilon_i \sim N(0, \sigma^2)$ 이고, α, β, σ^2 은 미지의 모수임

- 단순회귀모형은 $x_1, \cdots, x_n$ 에서 $Y_1, \cdots, Y_n$ 의 관측값 $y_1, \cdots, y_n$ 을 사용하여, x_i에 따른 Y_i 의 기대값 $E(Y_i) = \alpha + \beta x_i$ 를 추정하는 것이다.

- 귀무가설 $H_0 : \beta = 0$ (독립변수는 종속변수에 영향을 주지 않는다)



■ 최소제곱법을 이용한 모수추정

- **최소제곱법(Least Square Method)**

단순회귀모형 $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad (i = 1, 2, \cdots, n)$ 에서 오차의 제곱합

$$S = S(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^{\,2} = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

이 최소가 되도록 β_0 와 β_1 를 추정하는 방법을 최소제곱법이라 하고,

이 때 다음과 같이 얻어지는 추정량 $\hat{\beta}_0$ 과 $\hat{\beta}_1$ 은 **최소제곱추정량(least square estimator)**이라 한다.

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

여기서 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \quad s_{XY} = \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}), \; s_{XX} = \sum_{i=1}^{n} (X_i - \bar{X})^2$

■ 최소제곱법을 이용한 모수추정

- **회귀모형에서의 분산모수 σ^2 에 대한 추정**
  - **잔차(residual)**

    단순회귀모형 $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \ (i = 1, 2, \cdots, n)$ 에서 Y_i와 $E(Y_i) = \beta_0 + \beta_1 X_i$ 의 추정량 $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ 의 차를 잔차라고 하며 다음과 같이 정의한다.

    $$e_i = Y_i - \hat{Y}_i$$

  - **잔차제곱합 (residual sum of squares, RSS)**

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

    $$= S_{YY} - \frac{S_{XY}^{\ 2}}{S_{XX}} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 - \hat{\beta}_1^2 \sum_{i=1}^{n} (X_i - \bar{X}_i)^2$$

  - **σ^2 의 불편추정량**

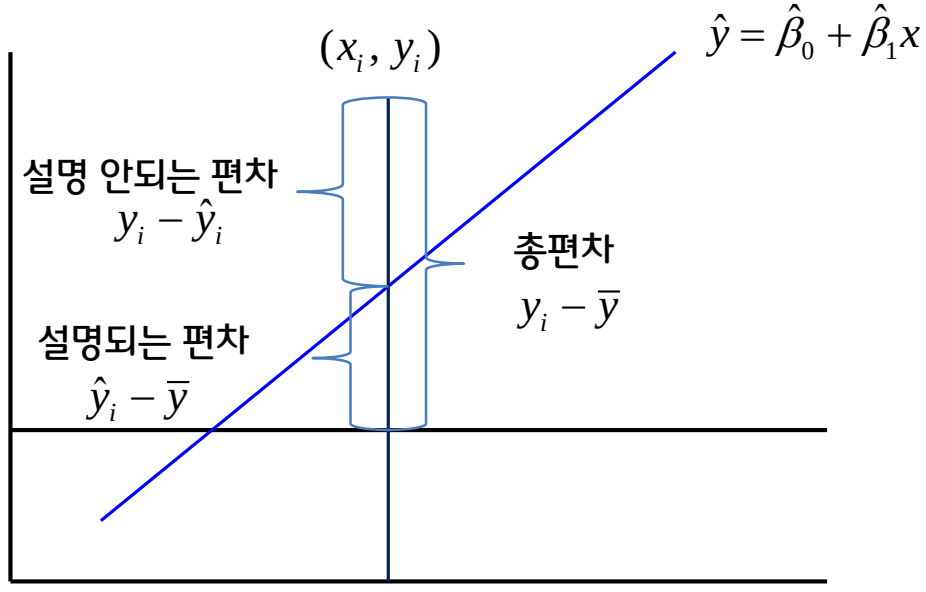
    $$\hat{\sigma}^2 = \frac{RSS}{n-2} = MSE$$, 잔차자유도 = 관측개체수 –모형에서의 모수의 수

■ 제곱합(변동)의 분해

▪ 총변동의 분해

- 단순회귀모형 $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad (i = 1, 2, \cdots, n), \quad \varepsilon_i \sim iid\ N(0, \sigma^2)$

  최소제곱회귀직선이 독립변수가 종속변수를 어느 정도 잘 설명하는지를 알고자 함



■ 제곱합(변동)의 분해

▪ 총변동의 분해

- 총변동(total sum of squares, TSS)은 회귀식에 설명이 되는 변동(sum of squares due to regression, SSR)과 회귀식으로 설명이 안되는 오차변동(sum of squares due to error, SSE)으로 분해된다.

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

전체제곱합(TSS) = 회귀제곱합(SSR) + 오차제곱합(SSE)

- 계산

$$TSS=\sum_{i=1}^{n}(Y_i-\bar{Y})^2=\sum_{i=1}^{n}Y_i^2-\frac{\left(\sum_{i=1}^{n}Y_i\right)^2}{n}$$

$$SSR=\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2=\sum_{i=1}^{n}(\bar{Y}+\hat{\beta}_1(X_i-\bar{X})-\bar{Y})^2=\hat{\beta}_1^2S_{XX}=\frac{S_{XY}^2}{S_{XX}}$$

$$SSE=\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2=TSS-SSR$$

■ 결정계수

- **결정계수(determination coefficient, R^2)**
  - 전체 변동에 대해 추정된 회귀식에 의해 설명되는 정도를 나타내고 다음과 같이 정의된다.

$$R^2 = \frac{SSR}{TSS} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = 1 - \frac{SSE}{TSS}$$

  - $0 \le R^2 \le 1$
  - R^2 이 클수록 추정된 회귀식의 설명력이 높다고 할 수 있다.
  - 단순회귀분석의 경우 상관계수의 제곱은 결정계수가 된다.

$$r^2 = \left(\frac{s_{XY}}{\sqrt{s_{XX}\, s_{YY}}} \right)^2 = \frac{s_{XY}^2}{s_{XX}\, s_{YY}} = \frac{s_{XY}^2}{s_{XX}} \frac{1}{s_{YY}} = \frac{SSR}{SST} = R^2$$

■ 분산분석표

- 분산분석표 (ANOVA table)

요인 (source)	제곱합 (SS)	자유도 (df)	평균제곱합 (MS)	F	p-value
모형	SSR	1	$MSR = SSR/1$	$F = \frac{MSR}{MSE}$	$F(1, n-2) > F$
오차	SSE	$n-2$	$MSE = \frac{SSE}{n-2}$		
계 (total)	TSS	$n-1$			

- 모형의 유의성 검정
  - 가설 $H_0 : Y = \beta_0 + \varepsilon$ $vs.$ $H_1 : Y = \beta_0 + \beta_1 X + \varepsilon$
  - 검정통계량

$$F = \frac{MSR}{MSE} \sim F_{1,n-2} \quad \text{under } H_0$$

  - 기각역

$$F \geq F(1, n-2; \alpha)$$

■ 잔차분석

- 선형모형에서 오차항에 대한 가정

① $E(\varepsilon_i) = 0$

② $Var(\varepsilon_i) = \sigma^2$ (등분산성)

③ ε_i는 서로 독립이다 (독립성).

④ ε_i는 모든 i에 대하여 정규분포를 따른다 (정규성).

- **선형회귀모형의 적합성**
  - 선형회귀모형이 적합한가는 선형회귀모형을 가정하고 자료의 산점도를 그려보거나
  - 설명변수에 대한 잔차산점도 또는 반응변수에 대한 잔차산점도를 그려 보아 짐작할 수 있다.
  - 회귀모형이 타당하고 오차의 등분산성이 성립된다면 잔차산점도에서 잔차들이 0을 중심으로 랜덤하게 나타나야 한다.
    - 잔차산점도가 이차곡선이나 삼차곡선의 형태가 나타난다면 가정된 선형회귀함수는 적절하지 못함

2. 단순회귀분석

■ R에서 회귀모형 표현

- **예제: cars 데이터 (1920년대 측정 데이터)**
  - speed: 자동차의 주행속도, dist : 브레이크를 밟았을 때의 제동 거리

```
> m=lm(dist ~ speed, cars);
> summary(m)                                          # 회귀모형

Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,      Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

2. 단순회귀분석

■ R에서 회귀모형 표현

- 예제: cars 데이터

```
> attributes(m)
$names
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"

$class
[1] "lm “
> coef(m)
(Intercept)       speed
 -17.579095    3.932409
> fitted(m)[1:4]
        1         2         3         4
-1.849460 -1.849460  9.947766  9.947766
> residuals(m)[1:4]
        1         2         3         4
 3.849460 11.849460 -5.947766 12.052234
> fitted(m)[1:4]+residuals(m)[1:4]
 1  2  3  4
 2 10  4 22
> cars$dist[1:4]
[1]  2 10  4 22
```

2. 단순회귀분석

■ R에서 회귀모형 표현

- 예제: cars 데이터

```
> confint(m)
                2.5 %    97.5 %
(Intercept) -31.167850 -3.990340
speed         3.096964  4.767853
> deviance(m)
[1] 11353.52
> predict(m, newdata=data.frame(speed=3))
        1
-5.781869
> coef(m)
(Intercept)       speed
 -17.579095    3.932409

> -17.579095+3.932409*3
[1] -5.781868

> predict(m, newdata=data.frame(speed=3), interval="confidence")
        fit       lwr      upr
1 -5.781869 -17.02659 5.462853
```

2. 단순회귀분석

■ R에서 회귀모형 표현

- 예제: cars 데이터

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

Response: dist
          Df Sum Sq Mean Sq F value   Pr(>F)
speed      1  21186 21185.5  89.567 1.49e-12 ***
Residuals 48  11354   236.5
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> full = lm(dist ~ speed, data=cars); full                    # 적합 모형

Call:
lm(formula = dist ~ speed, data = cars)

Coefficients:
(Intercept)        speed
    -17.579        3.932
```

■ R에서 회귀모형 표현

- 예제: cars 데이터

```
> reduced = lm(dist ~ 1, data=cars); reduced                    # 축소 모형

Call:
lm(formula = dist ~ 1, data = cars)

Coefficients:
(Intercept)
      42.98

> anova(reduced, full)                                          # 모형 비교
Analysis of Variance Table

Model 1: dist ~ 1
Model 2: dist ~ speed
  Res.Df   RSS Df Sum of Sq      F   Pr(>F)
1     49 32539
2     48 11354  1     21186 89.567 1.49e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

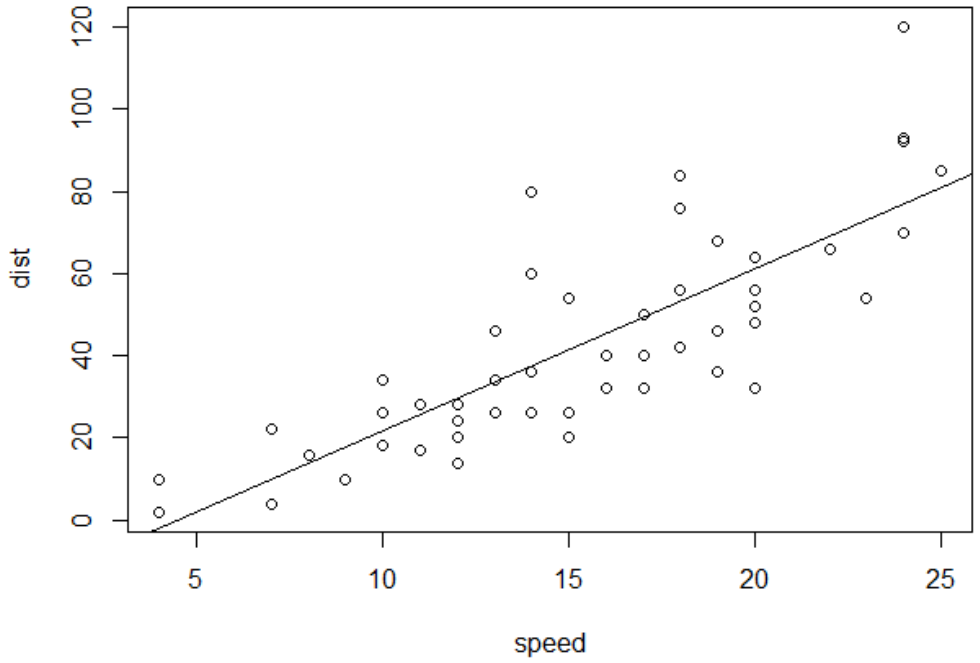
2. 단순회귀분석

■ R에서 회귀모형 표현

- 예제: cars 데이터

```
> par(mfrow=c(1,1))
> with(cars, plot(dist~speed))   #  with(cars, plot(speed, dist))
> abline(coef(m))
```

회귀직선의 시각화

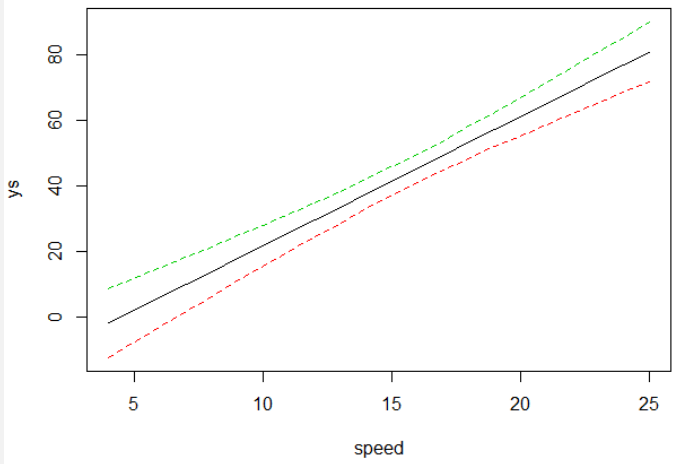


■ R에서 회귀모형 표현

- **예제: cars 데이터**
  - 그래프에 추정값의 신뢰구간을 포함하는 시각화: matplot, matlines 함수 사용

```
> summary(cars$speed)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    4.0    12.0    15.0    15.4    19.0    25.0
> predict(m, newdata=data.frame(speed=seq(4, 25, 0.2)),
          interval="confidence")
           fit         lwr       upr
1   -1.8494599 -12.3295433  8.630624
2   -1.0629781 -11.3914503  9.265494
3   -0.2764964 -10.4538419  9.900849
4    0.5099854  -9.5167401 10.536711
5    1.2964672  -8.5801681 11.173102
6  2.0829489  -7.6441504 11.810048
7  ...

> speed = seq(min(cars$speed), max(cars$speed), .1)
> ys=predict(m, newdata=data.frame(speed=speed),
+    interval="confidence")
> matplot(speed, ys, type='n')
> matlines(speed, ys, lty=c(1,2,2), col=c(1,2,3))
```



■ 다중회귀분석(multiple regression analysis)

- **다중회귀 모형**

  k 개 독립변수의 정해진 값 $x_{1i}, \cdots, x_{ki}\ (i = 1, \cdots, n)$ 에서 측정되는 종속변수 $Y_1, \cdots, Y_n$ 에 대하여 다음의 관계식이 성립한다고 가정하자.

$$Y_i = \alpha + \beta x_{1i} + \beta x_{2i} + \cdots + \beta x_{ki} + \varepsilon_i, \quad (i = 1, 2, \cdots, n)$$

  여기에서 $\varepsilon_1, \cdots, \varepsilon_n$은 서로 독립이며 $\varepsilon_i \sim N(0, \sigma^2)$이고, $\alpha, \beta_1, \beta_2, \cdots, \beta_k, \sigma^2$은 미지의 모수임

- 귀무가설: 독립변수들의 회귀계수들은 0이다.
- 예제
  - 촉진제의 양, 촉진제의 순도, 반응온도 3개의 독립변수가 반응량의 변화에 영향
  - 비만도와 나이가 혈압에 영향을 미칠 때, 이들간의 회귀분석
  - 재료비와 인건비를 통한 생산비용의 예측
  - 학생들의 지능, 성격 점수가 학업성적에 영향을 미치는지 분석

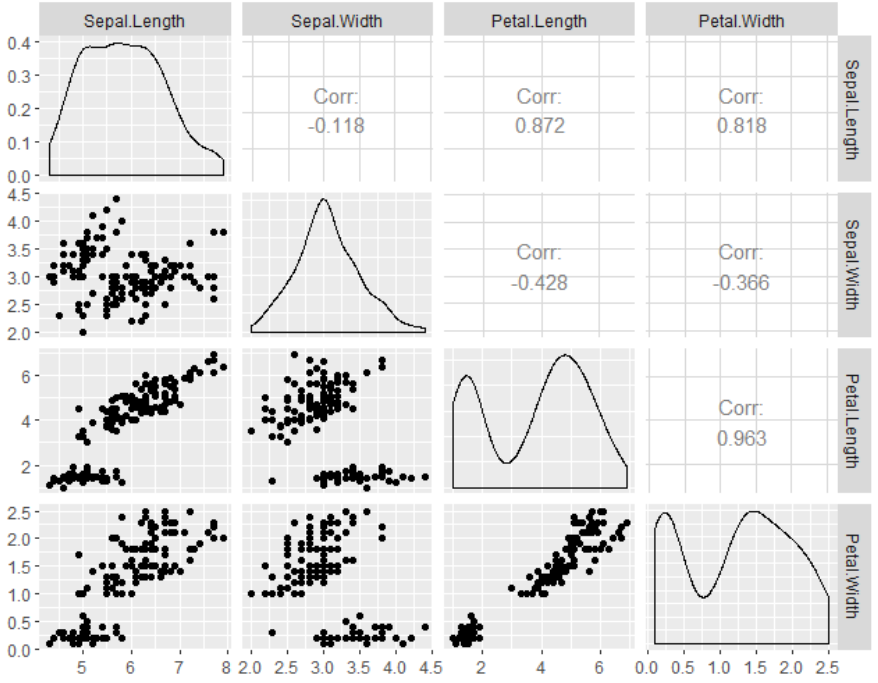
3. 다중회귀분석

■ 다중회귀분석(multiple regression analysis)

▪ 시각적 탐색

- ggpairs 함수: ggplot2, GGally 페키지 사용

```
> library(ggplot2)
> library(GGally)
> ggpairs(iris[, 1:4])
```

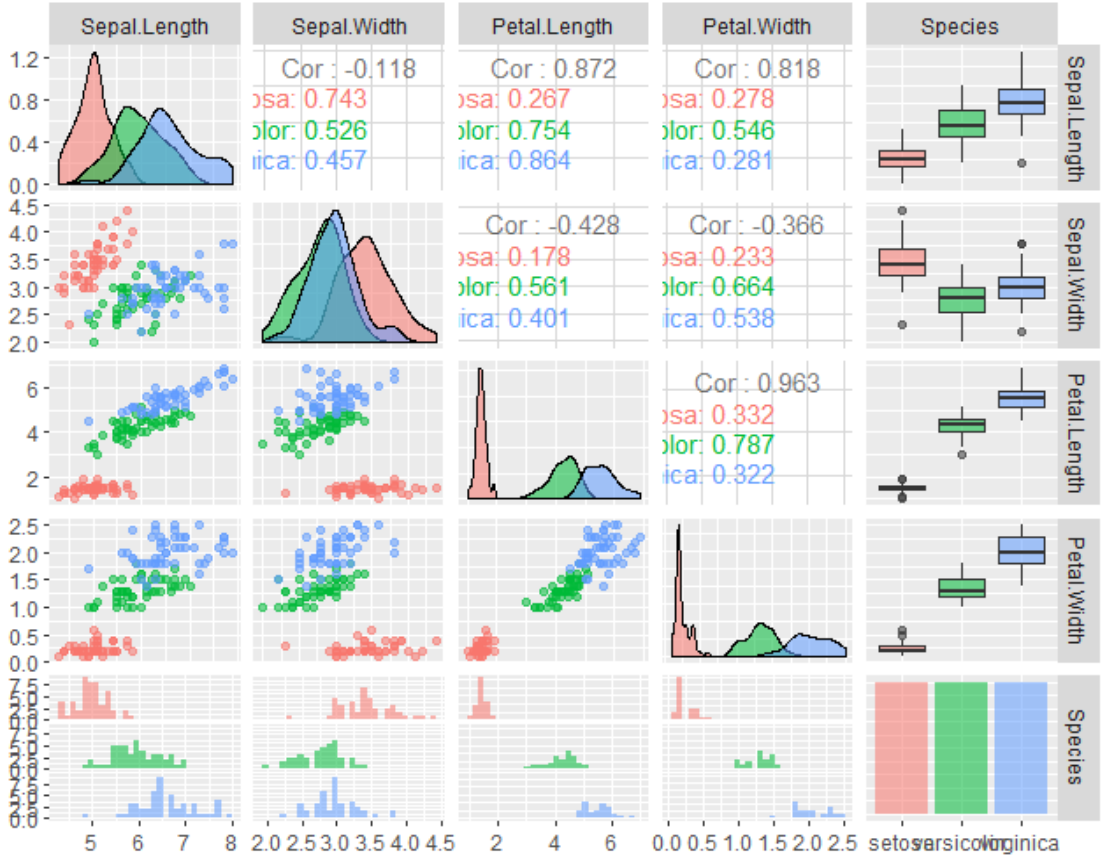


3. 다중회귀분석

■ 다중회귀분석(multiple regression analysis)

▪ 시각적 탐색

```
>ggpairs(iris, aes(colour = Species))
```



■ 다중회귀분석(multiple regression analysis)

▪ 예제: airquality data

- NEW YORK 도시의 153일 동안 오존, 일조량, 기온과 풍속을 기록한 데이터

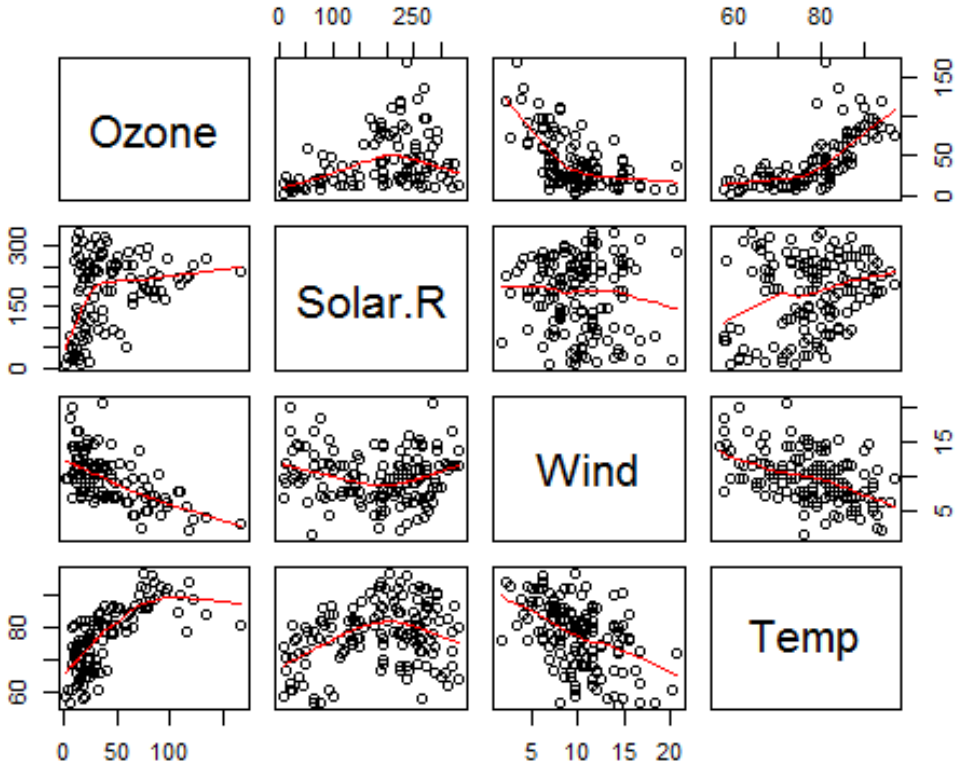
```
> data(airquality)
> head(airquality)
  Ozone Solar.R Wind Temp Month Day
1    41     190  7.4   67     5   1
2    36     118  8.0   72     5   2
3    12     149 12.6   74     5   3
4    18     313 11.5   62     5   4
5    NA      NA 14.3   56     5   5
6    28      NA 14.9   66     5   6
> str(airquality)
'data.frame':           153 obs. of  6 variables:
 $ Ozone  : int  41 36 12 18 NA 28 23 19 8 NA ...
 $ Solar.R: int  190 118 149 313 NA NA 299 99 19 194 ...
 $ Wind   : num  7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6 ...
 $ Temp   : int  67 72 74 62 56 66 65 59 61 69 ...
 $ Month  : int  5 5 5 5 5 5 5 5 5 5 ...
 $ Day    : int  1 2 3 4 5 6 7 8 9 10 ...
```

3. 다중회귀분석

■ 다중회귀분석(multiple regression analysis)

- 예제: airquality data

```
> pairs(airquality[,1:4], panel=panel.smooth)
```



■ 다중회귀분석(multiple regression analysis)

- 예제: airquality data

```
> lm.a = lm(Ozone~Solar.R+Wind+Temp, data=airquality)      # Ozone 회귀모형
> summary(lm.a)
Call:
lm(formula = Ozone ~ Solar.R + Wind + Temp, data = airquality)

Residuals:
    Min      1Q  Median      3Q     Max
-40.485 -14.219  -3.551  10.097  95.619

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -64.34208   23.05472  -2.791  0.00623 **
Solar.R       0.05982    0.02319   2.580  0.01124 *
Wind         -3.33359    0.65441  -5.094 1.52e-06 ***
Temp          1.65209    0.25353   6.516 2.42e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.18 on 107 degrees of freedom
  (42 observations deleted due to missingness)
Multiple R-squared:  0.6059,        Adjusted R-squared:  0.5948
F-statistic: 54.83 on 3 and 107 DF,  p-value: < 2.2e-16
```

Ozone에 대한 회귀식

Ozone=-64.34+0.059 Solar.R-3.333 Wind+1.652Temp

$R^2 = 0.6059$

$\hat{\sigma} = 21.181$

■ 다중회귀분석(multiple regression analysis)

- 예제: airquality data

```
> lm.ab = lm(log(Ozone)~Solar.R+Wind+Temp, data=airquality)   # log(Ozone) 회귀모형
> summary(lm.ab)
Call:
lm(formula = log(Ozone) ~ Solar.R + Wind + Temp, data = airquality)

Residuals:
     Min       1Q   Median       3Q      Max
-2.06193 -0.29970 -0.00231  0.30756  1.23578

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.2621323  0.5535669  -0.474 0.636798
Solar.R      0.0025152  0.0005567   4.518 1.62e-05 ***
Wind        -0.0615625  0.0157130  -3.918 0.000158 ***
Temp         0.0491711  0.0060875   8.077 1.07e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5086 on 107 degrees of freedom
  (42 observations deleted due to missingness)
Multiple R-squared:  0.6644,        Adjusted R-squared:  0.655
F-statistic: 70.62 on 3 and 107 DF,  p-value: < 2.2e-16
```

Log(Ozone)에 대한 회귀식

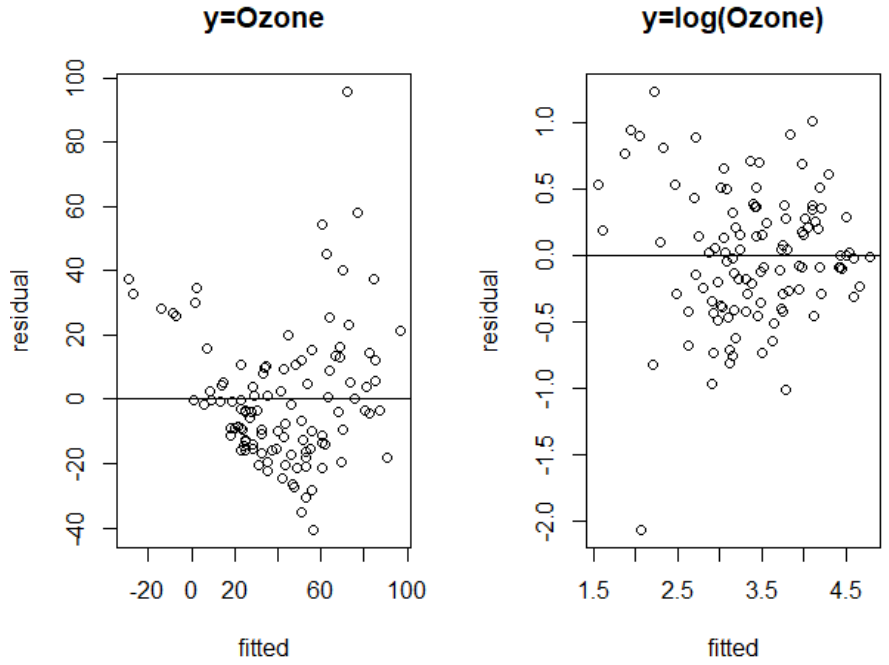
Log(Ozone)=-0.262+0.0025 Solar.R-0.062Wind+0.049Temp

3. 다중회귀분석

■ 다중회귀분석(multiple regression analysis)

- 예제: airquality data

```
> op=par(mfrow=c(1,2))
> plot(fitted(lm.a), residuals(lm.a), xlab='fitted', ylab='residual', main='y=Ozone')
> abline(h=0)
> plot(fitted(lm.ab), residuals(lm.ab), xlab='fitted', ylab='residual', main='y=log(Ozone)')
> abline(h=0)
```



Ozone 모형에 대한 예측값이 커짐에 따라 잔차는 분포가 커지고 있으나,
Log(Ozone) 모형은 예측값이 커짐에 따라 0을 중심으로 분산이 일정하게 유지되고 있음

■ 다중회귀분석(multiple regression analysis)

- 예제: airquality data

```
> shapiro.test(residuals(lm.a))   #  잔차 정규성 검정

        Shapiro-Wilk normality test

data:  residuals(lm.a)
W = 0.91709, p-value = 3.618e-06

> shapiro.test(residuals(lm.ab)) #  잔차 정규성 검정

        Shapiro-Wilk normality test

data:  residuals(lm.ab)
W = 0.97749, p-value = 0.05726
```

Ozone에 대한 회귀모형은 Shapiro-Wilk W=0.91709로 유의수준 0.05에서 잔차가 정규성을 만족하지 않지만 Log(Ozone) 에 대한 회귀모형은 유의수준 0.05에서 잔차가 정규성을 만족하고 있다.

■ 다중회귀분석(multiple regression analysis)

- 예제: airquality data

```
> confint(lm.ab)                    #  log(Ozone) 회귀모형 회귀계수 95% 신뢰영역
                  2.5 %      97.5 %
(Intercept) -1.359514052  0.83524943
Solar.R      0.001411525  0.00361883
Wind        -0.092711591 -0.03041335
Temp         0.037103361  0.06123889
> coef(lm.ab)                       #   parameter coeff
 (Intercept)      Solar.R         Wind         Temp
-0.262132313  0.002515177 -0.061562470  0.049171124

> install.packages("ellipse");  library(ellipse)
> op=par(mfrow=c(1,3))
> plot(ellipse(lm.ab,c(2,3)), type="l")
> points(coef(lm.ab)[2], coef(lm.ab)[3], pch=18)
> abline(v=confint(lm.ab)[2,], lty=2)
> abline(h=confint(lm.ab)[3,], lty=2)
> plot(ellipse(lm.ab,c(2,4)), type="l")
> points(coef(lm.ab)[2], coef(lm.ab)[4], pch=18)
> plot(ellipse(lm.ab,c(3,4)), type="l")
> points(coef(lm.ab)[3], coef(lm.ab)[4], pch=18)
```

ellipse(,c(2,3))은 변수2와 3의 상관관계 표현

점의 종류 18= ◆

■ 다중회귀분석(multiple regression analysis)

- 예제: airquality data

```
> x0 = data.frame(Sola.R=170, Wind=8, Temp=70, Month=0, Day=0)
> x0
  Sola.R Wind Temp Month Day
1    170    8   70     0   0

> predict(lm.ab, newdata=x0, interval="confidence")
       fit      lwr      upr
1 3.114927 2.951955 3.277899

> predict(lm.ab, newdata=x0, interval="prediction")
       fit      lwr      upr
1 3.114927 2.093657 4.136196


>  #  CSV로 데이터 export 함
> data(airquality)
> write.csv(airquality, "D:/data/airquality.csv", na=" ", row.names=TRUE)
```

회귀직선에 대한 신뢰구간

예측값에 대한 신뢰구간

■ 다중공선성

▪ 다중공선성 (multicollinearity)

- 어느 한 변수가 다른 변수들에 의해 선형관계가 성립하면 공선성(collinear)이 존재한다고 함
- 공선성이 존재하는 경우 추정에 문제가 발생하게 되어 공선성 문제를 야기한다.
- 설명변수 중 한 개가 다른 설명변수들의 선형결합으로 표현된다면 $X'X$가 비정칙 (singular) 상황이 되며 다중공선성을 갖게 되고 $(X'X)^{-1}$를 구할 수 없어 회귀계수 β에 대한 최소제곱추정량을 구하는 데 문제가 발생한다.
- 예를 들어, 설명변수가 2개인 다중회귀모형 $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$을 고려하자.

  r_{12}를 X_1과 X_2의 상관계수라 할 때 다음이 성립한다.

$$Var(\hat{\beta}_j) = \sigma^2 \left(\frac{1}{1-r_{12}^2} \right) \left(\frac{1}{S_{X_j X_j}} \right), \quad S_{X_j X_j} = \sum_{l=1}^{n} (X_{jl} - \bar{X}_j)^2$$

  - $Var(\hat{\beta}_j)$은 $r_{12}^2 = 0$ 일 때 최솟값을 가지며 $r_{12}^2 \approx 1$ 이면 $\hat{\beta}_i$ 의 분산은 크게 팽창되어(inflated) 회귀계수 추정값은 매우 불안정하게 된다.

  - 다중공선성이 존재하는 경우 반응변수의 민감도는 설명변수의 위치에 따라 크게 죄우될 수 있으며 관측된 데이터에서 멀어지는 경우 예측값을 불안정하게 만드는 원인이 된다.

■ 다중공선성

- **다중공선성이 판단되는 경우**

① 설명변수들 간의 상관행렬에서 절대값이 큰 상관계수를 갖는 경우

② j번째 설명변수를 반응변수 위치에 놓고 나머지 설명변수들로 회귀모형을 적합했을 때 결정계수 R_j^2이 1에 가까운 경우

③ $X'X$ 의 고유값을 구한 후 크기 순서대로 늘어 놓아 가장 큰 고유값을 λ_1이라 놓는다.

$\lambda_1 > \lambda_2 > \cdots > \lambda_{k+1}$ 을 만족하며 가장 작은 고유값 λ_{k+1}에 대해 조건수(condition number)

$$\phi = \sqrt{\lambda_1 / \lambda_{k+1}}$$

ⅰ) $\phi > 30$ 이면 심각한 다중공선성 존재

ⅱ) $10 < \phi < 30$ 이면 다소 강한 다중공선성 존재

ⅲ) $\phi < 10$ 이면 다중공선성이 거의 없다고 판단한다.

④ 가장 큰 고유값과 각 고유값의 비 $\phi_k = \sqrt{\lambda_1 / \lambda_k}$, $k = 1, 2, \cdots, k+1$를 상태지표(condition index)라 한다.

⑤ 개체 고유값 자체로써 다중공선성을 진단할 수 있다.

- $\lambda_{\min} < 0.05$ 이면 심각한 다중공선성이 있음을 나타낸다.

■ 다중공선성

- **분산팽창인수**
  - 설명변수들 k 개 포함한 다중회귀모형 $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$ 에서 회귀계수 추정량의 분산은

$$Var(\hat{\beta}_j) = \sigma^2 \left(\frac{1}{1 - R_j^2} \right) \left(\frac{1}{S_{X_j X_j}} \right), \quad j = 1, 2, \cdots, k$$

여기서 R_j^2 은 j 번째 설명변수를 반응변수 위치에 놓고 나머지 설명변수들로 회귀모형을 적합했을 때의 결정계수이다. 만약 j 번째 설명변수가 나머지 변수들의 어떤 부분집합에 거의 선형종속이면 R_j^2 은 1에 가깝게 되고 $1/(1-R_j^2)$ 은 커진다.

  - 분산팽창인수 (variance inflation factor)

$$VIF_j = \frac{1}{1 - R_j^2}, \quad j = 1, 2, \cdots, k$$

    - 한 개 또는 그 이상의 VIF 값이 크면 다중공선성이 있음을 나타낸다.
    - $R_j^2 \geq 0.60$ 이상, $VIF_j \geq 2.5$ 일 경우 다중공선성을 확인해야 한다.
    - $VIF_j > 10$ 이면 심각한 다중공선성 문제를 가진다.
  - R: library(car)의 vif 함수를 이용

■ 다중공선성

- **다중공선성의 문제**
  - 다중공선성은 추정식과 예측값을 불안정하게 한다.
  - 설명변수들 간의 상관관계로 인해 회귀계수에 대한 검정이 유의하지 않을 수 있음
    - 추정식이 유의하지 않게 되며 따라서 예측값에 대해 신뢰할 수 없게 된다.
  - 설명변수의 위치에 따라 예측값이 크게 좌우될 수 있으며 관측된 데이터에서 멀어지는 경우 예측값을 불안정하게 만드는 원인이 된다.

- 다중공선성 해결 방법
  - 상관성이 큰 설명변수를 제거한 후 분석한다.
  - 모든 설명변수를 포함시키고자 한다면 능형회귀(ridge regression) 등의 대안 모형을 모색한다.
  - 상관관계가 높은 설명변수들끼리 선형결합하는 주성분회귀(principal components regression) 방법을 사용할 수 있다.

■ 변수선택

- 유의한 설명변수를 삽입하거나 유의하지 않은 설명변수들을 제거함으로써 유의한 설명변수 만으로 회귀모형을 설정하는 방법
- 전진적 선택(forward selection), 후진적 선택(backward selection), 단계별 선택(stepwise selection) 접근 방법이 있으며 단계별 선택 접근법이 가장 많이 사용된다.
- **단계별 변수 선택(stepwise selection) 방법**
  - ① 설명변수 중 반응변수와 상관관계가 가장 크고 설명력이 유의한 설명변수를 먼저 선택한다.
  - ② 이미 선택된 설명변수를 제외한 나머지 설명변수를 하나씩 삽입하여 유의한 설명변수를 선택한다.
  - ③ 새로 선택된 설명변수가 먼저 선택되었다고 가정하고, 이미 삽입되어 있던 설명변수의 유의성에 대한 F-검정을 한다. 이때 유의하지 않은 설병변수는 제외된다.
  - ④ 이와 같은 과정을 반복하면서 추가된 모형이 전 단계의 모형보다 유의하지 않으면 선택 과정을 멈추고 최종모형으로 결정한다.
  - 단계별 변수선택 방법으로 결정된 모형은 모형 선택을 위한 다른 기준으로 점검해 볼 때 최적 모형이 아닐 수 있다.
  - 단계별 변수 선택 방법에서 설명변수들의 순서가 정해지는 방법은 데이터 기반 통계량에 의한 방법이며 실질적인 관계를 가지는 변수가 제외될 수 도 있다.

■ 모형 선택

- **stepAIC 함수:** AIC를 사용해 최적의 모형을 선택함. package: MASS
- step 함수도 동일한 포멧으로 동일한 결과를 줌

함수		Arguments
stepAIC(object, scope, scale = 0, direction = c("both", "backward", "forward"), trace = 1, keep = NULL, steps = 1000, use.start = FALSE, k = 2, ...)	object	an object representing a model of an appropriate class. This is used as the initial model in the stepwise search.
	scope	defines the range of models examined in the stepwise search. This should be either a single formula, or a list containing components upper and lower, both formulae. See the details for how to specify the formulae and how they are used.
	scale	used in the definition of the AIC statistic for selecting the models, currently only for lm and aov models (see extractAIC for details).
	direction	the mode of stepwise search, can be one of "both", "backward", or "forward", with a default of "both". If the scope argument is missing the default for direction is"backward".
	trace	if positive, information is printed during the running of stepAIC. Larger values may give more information on the fitting process.
	keep	a filter function whose input is a fitted model object and the associated AIC statistic, and whose output is arbitrary. Typically keep will select a subset of the components of the object and return them. The default is not to keep anything.
	steps	the maximum number of steps to be considered. The default is 1000 (essentially as many as required). It is typically used to stop the process early.
	use.start	if true the updated fits are done starting at the linear predictor for the currently selected model. This may speed up the iterative calculations for glm (and other fits), but it can also slow them down. **Not used** in R.

■ 모형 선택

- 예제 : longley data
  - 1947년~1962년 연간 7개의 경제변수들에 대한 데이터

```
> data(longley)
> attach(longley)
> str(longley)
'data.frame':          16 obs. of  7 variables:
 $ GNP.deflator: num  83 88.5 88.2 89.5 96.2 ...
 $ GNP         : num  234 259 258 285 329 ...
 $ Unemployed  : num  236 232 368 335 210 ...
 $ Armed.Forces: num  159 146 162 165 310 ...
 $ Population  : num  108 109 110 111 112 ...
 $ Year        : int  1947 1948 1949 1950 1951 1952 1953 1954 1955 1956 ...
 $ Employed    : num  60.3 61.1 60.2 61.2 63.2 ...
> y = GNP
> x1 = Unemployed; x2=Population; x3=Armed.Forces
> mydata = data.frame(y,x1,x2,x3)
> fit= lm(y ~ x1+x2+x3, data=mydata)
```

■ 모형 선택

- 예제 : longley data

```
> summary(fit)     # show results

Call:
lm(formula = y ~ x1 + x2 + x3, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-14.525  -6.989   1.574   5.657  13.434

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.352e+03  5.160e+01 -26.195 5.86e-12 ***
x1          -1.142e-01  4.109e-02  -2.780   0.0167 *
x2           1.498e+01  5.834e-01  25.677 7.42e-12 ***
x3           6.480e-02  4.308e-02   1.504   0.1584
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.384 on 12 degrees of freedom
Multiple R-squared:  0.9943,         Adjusted R-squared:  0.9929
F-statistic: 698.8 on 3 and 12 DF,  p-value: 9.943e-14
```

Y=GNP 에 대한 회귀식

Y = -1352-0.014x1 + 14.98x2+0.065x3

$R^2 = 0.994$

$\hat{\sigma} = \sqrt{MSE} = 8.384$

■ 모형 선택

- 예제 : longley data

```
> install.packages("perturb")
> library(perturb)
> cor(mydata[,2:4])
           x1        x2         x3
x1  1.0000000 0.6865515 -0.1774206
x2  0.6865515 1.0000000  0.3644163
x3 -0.1774206 0.3644163  1.0000000

> colldiag(fit,center=TRUE, add.intercept=FALSE)    # 다중공선성 검정
Condition
Index   Variance Decomposition Proportions
          x1    x2    x3
1  1.000 0.078 0.089 0.014
2  1.224 0.054 0.005 0.358
3  3.495 0.868 0.907 0.628

> vif(fit)
      x1       x2       x3
3.146686 3.514335 1.918225
```

■ 모형 선택

- 예제 : longley data

```
# Stepwise Regression
> library(MASS)
> fit= lm(y ~ x1+x2+x3, data=mydata)
> step = stepAIC(fit, direction = "both")
Start:  AIC=71.44
y ~ x1 + x2 + x3

       Df Sum of Sq   RSS     AIC
<none>               843  71.438
- x3    1       159  1002  72.202
- x1    1       543  1387  77.392
- x2    1     46338 47181 133.827
```

AIC로 보면 Model1이 선택됨
⇒작은 모형이 선호됨으로 Model2를 선택하는 것이 바람직함 (x1과 x2 변수 포함 모형)

■ 모형 선택

- 예제 : longley data

```
> attributes(step)
$names
 [1] "coefficients"  "residuals"     "effects"       "rank"          "fitted.values"
 [6] "assign"        "qr"            "df.residual"   "xlevels"       "call"
[11] "terms"         "model"         "formula"       "anova"

$class
[1] "lm “
> step$anova        #display results
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
y ~ x1 + x2 + x3

Final Model:
y ~ x1 + x2 + x3


  Step Df Deviance Resid. Df Resid. Dev      AIC
1                        12   843.4129 71.43789
```