	단위별 학습내용 (Week9)
wk9-1	다중회귀분석 I
wk9-2	다중회귀분석 II
wk9-3	데이터마이닝과 분류
wk9-4	학습데이터와 검증데이터

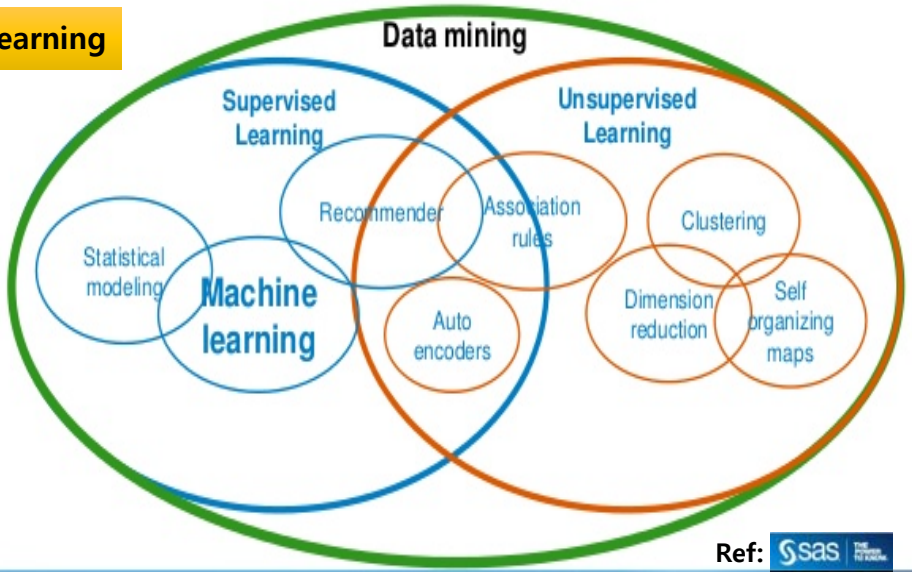
Wk9-1 : 데이터마이닝과 예측
- 다중회귀분석 I -

1. 데이터마이닝 기법

모형화	특징	내용	적용기법
예측	• 타겟변수 값이 주어지는 경우 (supervised learning) • 변수간의 관계	주어진 데이터를 기반으로 모델을 만든 후, y값을 예측 (y=continuous value)	• 다중회귀분석 • 주성분 회귀분석 • 부분최소자승법 • 신경망
분류		학습표본을 기반으로 분류규칙을 생성. 분류규칙의 성능을 검증하기 위해 실제범주와 추정된 범주를 비교 (y=0/1 혹은 다범주)	• 로지스틱 회귀모형 • 의사결정나무 • 선형판별분석 • 서포트벡터머신
군집	• 타겟변수 값이 없는 경우 (unsupervised learning) • 개체간의 관계	주어진 데이터 (X변수들)의 속성으로 군집화	• 계층형 군집 분석 • K-MEANS
연관규칙		연관성있는 변수관계 도출 (동시 발생 빈도 분석)	• 연관규칙 분석

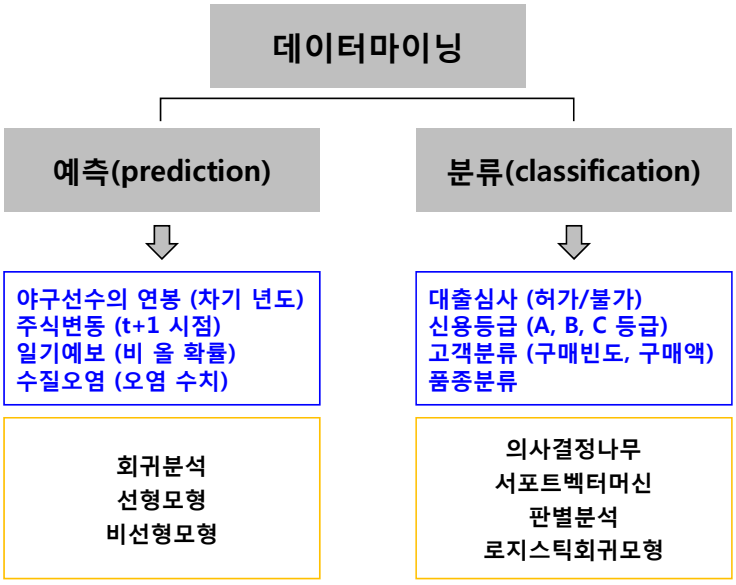
1. 데이터마이닝

- 데이터마이닝, 통계모델, 기계학습, 인공지능..
- **Supervised learning, Unsupervised learning**



Ref: sas

1. 데이터마이닝 기법



2. 다중회귀분석

- **다중회귀모형(multiple regression)**
  - 종속변수 Y를 설명하는 데 k개의 독립변수 $X_1, \dots, X_k$ 있을때 다중회귀모형은 다음과 같이 정의

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots \beta_k X_{ki} + \varepsilon_i, \qquad i = 1,2, \dots, n$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

회귀계수 β_k 의 해석 : 다른 독립변수들이 일정할 때 X_k의 한단위 변화에 따른 평균변화량

2. 다중회귀분석

• autompg 데이터

```
# lec9_1_MLR.r
# Multiple Regression
# stepwise method

# set working directory
setwd("D:/tempstore/moocr/wk9")

# autompg data
car<-read.csv("autompg.csv")
head(car)
str(car)
attach(car)
```

```
> head(car)
  mpg cyl disp hp   wt accler year origin                  carname
1  18   8  307 17 3504   12.0   70      1 chevrolet chevelle malibu
2  15   8  350 35 3693   11.5   70      1         buick skylark 320
3  18   8  318 29 3436   11.0   70      1        plymouth satellite
4  16   8  304 29 3433   12.0   70      1             amc rebel sst
5  17   8  302 24 3449   10.5   70      1               ford torino
6  15   8  429 42 4341   10.0   70      1          ford galaxie 500
```

Y

종속변수: mpg (연비)

X

독립변수: displacement (배기량)
horsepower (마력)
weight (무게)
acceleration (가속)

2. 다중회귀분석

• 다중회귀모형 : lm(y변수~x1+x2+x3, data=)

1st model : 전체변수를 모두 포함한 회귀모형

```
# multiple regression : 1st full model
r1<-lm(mpg ~ disp+hp+wt+accler, data=car)
summary(r1)
```

```
Call:
lm(formula = mpg ~ disp + hp + wt + accler, data = car)

Residuals:
     Min       1Q   Median       3Q      Max
-11.8331  -2.8735  -0.3164   2.4449  16.2079

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 40.8838025  1.9966258  20.476  < 2e-16 ***
disp        -0.0106291  0.0065254  -1.629   0.1041
hp           0.0047774  0.0082597   0.578   0.5633
wt          -0.0061405  0.0007449  -8.243 2.54e-15 ***
accler       0.1722165  0.0976340   1.764   0.0785 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.298 on 393 degrees of freedom
Multiple R-squared:  0.7006,    Adjusted R-squared:  0.6976
F-statistic:   230 on 4 and 393 DF,  p-value: < 2.2e-16
```

Check point 1

마력(hp)가 높을수록 연비가 좋은가??

⇒ 데이터 탐색 필요!!

선형회귀식

mpg = 40.88 - 0.011 disp + 0.0048 hp
- 0.0061 wt + 0.17 accler

선형회귀식의 결정계수

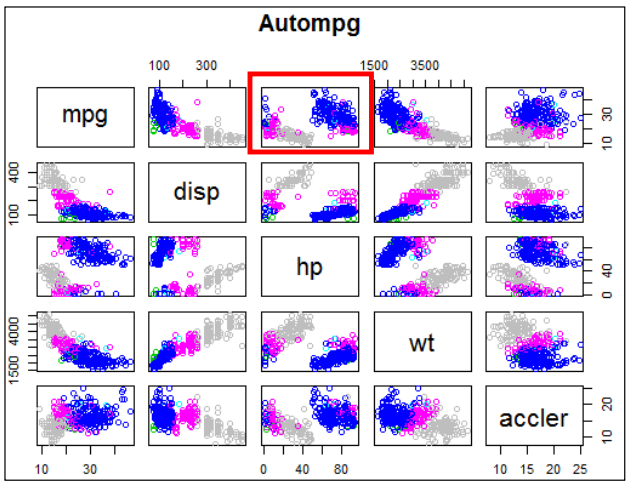
R^2=0.7006

2. 다중회귀분석

• 다중회귀모형 : 데이터탐색(Explanatory Data Analysis)

```
# pariwise plot
var1<-c("mpg","disp","hp","wt", "accler" )
pairs(car[var1], main ="Autompg",cex=1, col=as.integer(car$cyl))
```

Check point1



- 배기량(disp)과 연비(MPG)의 관계는?
- 마력(hp)과 연비(MPG)의 관계는 ?
- 차량무게(wt)와 연비(MPG)의 관계는?

3. 다중회귀분석 – 변수선택방법

• 변수선택방법 – 다수의 독립변수들이 있을때 최종모형은?

(1) 전진선택법(forward selection) - 독립변수중에서 종속변수에 가장 큰 영향을 주는 변수부터 모형에 포함

(2) 후진제거법(backward elimination)

- 독립변수를 모두 포함한 모형에서 가장 영향이 적은(중요하지 않은) 변수부터 제거

(3) 단계별 방법(stepwise method)

- 전진선택법에 의해 변수 추가
- 변수 추가시 기존 변수의 중요도가 정해진 유의수준(threshold)에 포함되지 않으면 앞에서 들어간 변수도 다시 제거됨

단계별방법의 예제

모형에 포함되는 유의수준(0.15)
모형에서 제거되는 유의수준(0.15)

step1 : x4 (p-value=0.01)
step2: x4 (0.01), x10 (0.03)
step3: x4 (0.01), x10 (0.2), x2 (0.12)
step4: x4, x2
step5: x4, x2, x5

3. 다중회귀분석 – 변수선택방법

• **단계별 방법(stepwise method)**

2nd model : 단계별 선택방법에 의한 회귀모형

- **step(모형, direction="both")**

```
# 2rd model using variable selection method
# step(r1, direction="forward")
# step(r1, direction="backward")
# stepwise selection
step(r1, direction="both")
#step(lm(mpg ~ disp+hp+wt+accler, data=car), direction="both")
```

```
> step(r1, direction="both")
Start:  AIC=1165.67
mpg ~ disp + hp + wt + accler

         Df Sum of Sq    RSS    AIC
- hp      1      6.18 7266.2 1164.0
<none>                7260.0 1165.7
- disp    1     49.01 7309.1 1166.3
- accler  1     57.48 7317.5 1166.8
- wt      1   1255.16 8515.2 1227.1

Step:  AIC=1164.01
mpg ~ disp + wt + accler

         Df Sum of Sq    RSS    AIC
<none>                7266.2 1164.0
- disp    1     51.76 7318.0 1164.8
- accler  1     58.62 7324.8 1165.2
+ hp      1      6.18 7260.0 1165.7
- wt      1   1291.30 8557.5 1227.1

Call:
lm(formula = mpg ~ disp + wt + accler, data = car)

Coefficients:
(Intercept)         disp           wt       accler
  41.299076    -0.010895    -0.006189     0.173851
```

R^2가 가장 높은 조합의 변수그룹을 선택
(AIC가 낮은 조합의 변수그룹을 선택)

변수 제거: hp

최종 변수 선택: disp, wt, accler

4. 다중회귀분석 – 최종모형

• **단계별 방법에 따른 최종 다중회귀모형**

2nd model : 단계별 선택방법에 의한 회귀모형

```
# final multiple regression
r2<-lm(mpg ~ disp+wt+accler, data=car)
summary(r2)
```

```
> summary(r2)

Call:
lm(formula = mpg ~ disp + wt + accler, data = car)

Residuals:
     Min       1Q   Median       3Q      Max
-11.7382  -2.8112  -0.3607   2.5231  16.1845

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 41.2990756  1.8614975  22.186  < 2e-16 ***
disp        -0.0108953  0.0065036  -1.675   0.0947 .
wt          -0.0061889  0.0007396  -8.368 1.03e-15 ***
accler       0.1738507  0.0975107   1.783   0.0754 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

Residual standard error: 4.294 on 394 degrees of freedom
Multiple R-squared:  0.7004,    Adjusted R-squared:  0.6981
F-statistic:   307 on 3 and 394 DF,  p-value: < 2.2e-16
```

선형회귀식

mpg = 41.30 - 0.011 disp- 0.0062 wt + 0.17 accler

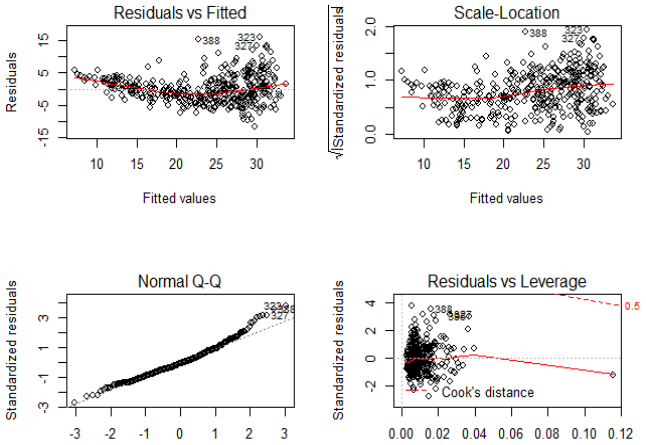
선형회귀식의 결정계수

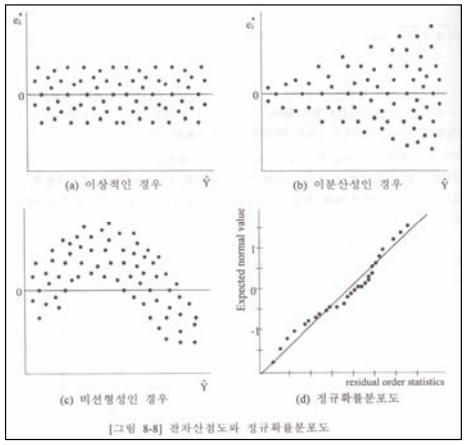
R^2=0.7004

5. 다중회귀분석 – 잔차의 산점도

- 회귀분석의 가정과 진단

```
# residual diagnostic plot
layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
plot(r2)
```





참고 : 전치혁, 정민근, 이혜선, 공학응용통계, 홍릉출판사, 2012