데이터과학을 위한 R프로그래밍

8주차. 데이터마이닝과 다중회귀

이혜선 교수
포항공과대학교 산업경영공학과

목차

8주차. 데이터마이닝과 다중회귀

8주차

1차시

다중회귀분석

POSTECH

데이터마이닝, 기계학습

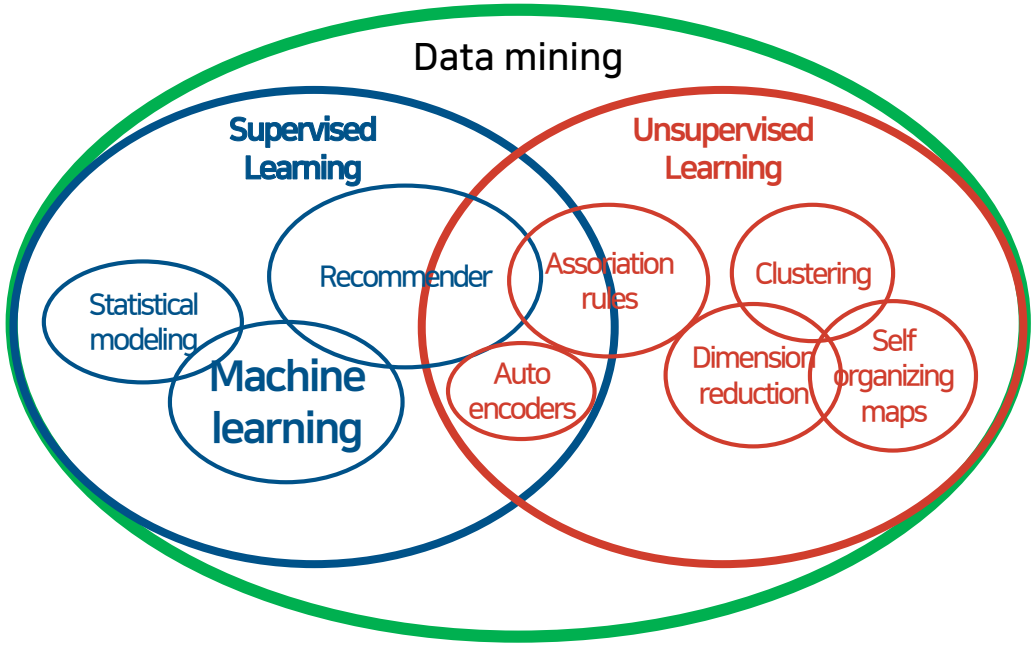
모형화	특징	내용	적용기법
예측	타겟변수 값이 주어지는 경우 (supervised learning)	주어진 데이터를 기반으로 모델을 만든 후, y값을 예측 (y=continuous value)	다중회귀분석 주성분 회귀분석 부분최소자승법 신경망
분류	변수간의 관계	학습표본을 기반으로 분류규칙을 생성. 분류규칙의 성능을 검증하기 위해 실제범주와 추정된 범주를 비교 (y=0/1 혹은 다범주)	로지스틱 회귀모형 의사결정나무 선형판별분석 서포트벡터머신
군집	타겟변수 값이 없는 경우 (unsupervisd learning)	주어진 데이터(X변수들)의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

POSTECH

데이터마이닝 기법

☑ 데이터마이닝, 통계모델, 기계학습, 인공지능..

- Supervised learning,
  Unsupervised learning



참고 SAS

POSTECH

데이터마이닝 기법

데이터마이닝

예측(prediction)

야구선수의 연봉 (차기 년도)
주식변동 (t+1 시점)
일기예보 (비올 확률)
수질오염 (오염 수치)

회귀분석
선형모형
비선형모형

분류(classification)

대출심사 (허가/불가)
신용등급 (A, B, C 등급)
고객분류 (구매빈도, 구매액)
품종분류

의사결정나무
서포트벡터머신
판별분석
로지스틱회귀모형

POSTECH

다중회귀분석

다중회귀모형(multiple regression)

- 종속변수 Y를 설명하는 데 k개의 독립변수 $X_1, \dots, X_k$ 가 있을 때 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_k 의 한 단위 변화에 따른 평균변화량

POSTECH

다중회귀분석

☑ autompg 데이터

```
# lec8_1_MLR.r
# Multiple Regression
# stepwise method

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

# 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
```

종속변수 : mpg (연비) Y

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

POSTECH

다중회귀분석

다중회귀분석

☑ 다중회귀모형 : lm(y변수~x1+x2+x3, data=)

1st model : 전체변수를 모두 포함한 회귀모형

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

```
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

마력(hp)이 높을수록 연비가 좋은가?

데이터 탐색 필요!!

선형회귀식

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

선형회귀식의 결정계수

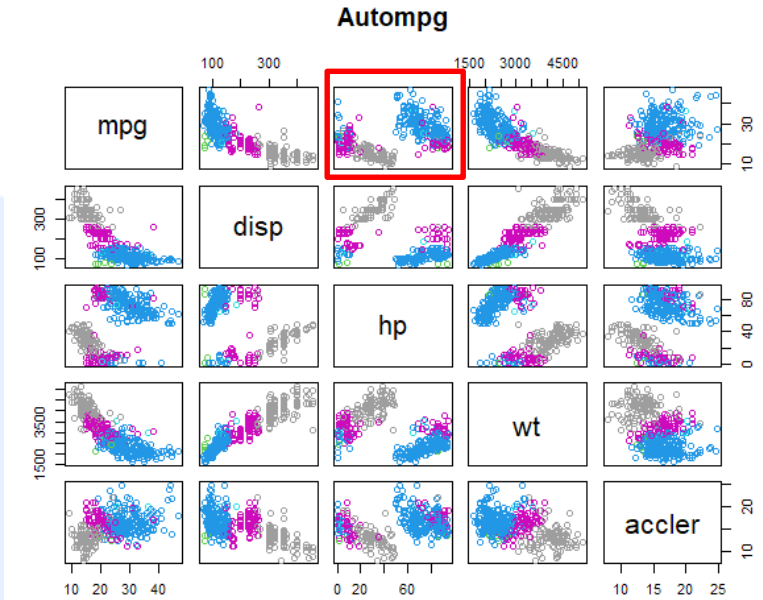
R^2=0.7006

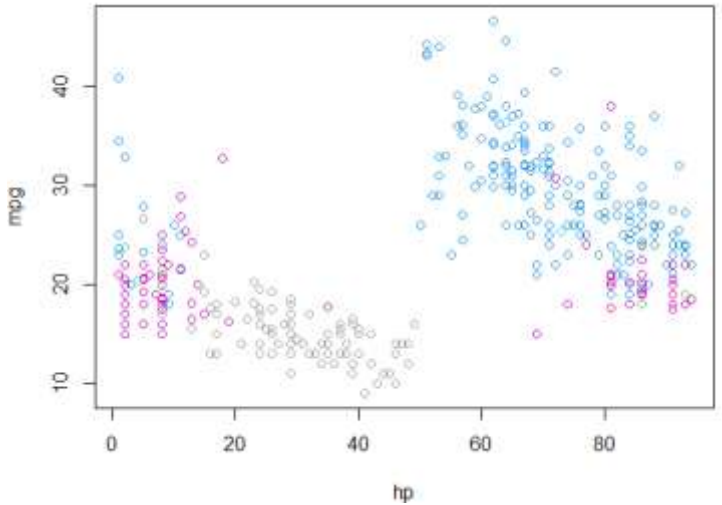
POSTECH

다중회귀분석 – 탐색과 진단

Pairwise scatterplot

- 변수들 관계를 보여주는 산점도





POSTECH

다중회귀분석 – 변수선택방법

단계별 방법(stepwise method)

2nd model : 단계별 선택방법에 의한 회귀모형

- step(모형, direction="both")

```
# 2nd model using variable selection method
# step(r1, direction="forward")
# step(r1, direction="backward")
# stepwise selection
s1<-step(r1, direction="both")
summary(s1)
```

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

변수 제거 : hp

최종 변수 선택 : disp, wt, accler

```
> summary(s1)

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 ' ' 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
```

POSTECH

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 ' ' 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

POSTECH

다중회귀분석 – 탐색과 진단

다중공선성(Multicollinearity)

- 독립변수들 사이에 상관관계가 있는 현상
- 다중공선성이 존재하는 경우 회귀계수 해석 불가능

독립변수들간의 상관계수

```
# check correlation between independent variables
var2<-c("disp","hp","wt", "accler" )
cor(car[var2])

# get correlation for each pair
# cor(disp, wt)
# cor(disp, accler)
# cor(wt, accler)
```

```
> cor(car[var2])
              disp         hp         wt     accler
disp     1.0000000 -0.4785123  0.9328241 -0.5436841
hp      -0.4785123  1.0000000 -0.4807430  0.2566567
wt       0.9328241 -0.4807430  1.0000000 -0.4174573
accler -0.5436841  0.2566567 -0.4174573  1.0000000
```

POSTECH

다중회귀분석 – 탐색과 진단

분산팽창계수(VIF : Variance Inflation Factor) – 다중공선성의 척도

$$VIF_j = \frac{1}{1 - R_j^2},$$

$$j = 1,2, \dots, k$$

- VIF는 다중공산성으로 인한 분산의 증가를 의미
- R_j^2은 X_j를 종속변수로 하고 나머지 변수를 독립변수로 하는 회귀모형에서의 결정계수
- VIF_j > 10 이상이면 다중공선성 고려

- 변수 선택 과정에서 상관계수가 높은 두 변수 중 하나만을 선택
- 더 많은 데이터 수집
- 능형회귀(ridge regression), 주성분회귀(principal components regression)

POSTECH

다중회귀분석 – 탐색과 진단

분산팽창계수 : VIF(다중회귀모형)

- car 패키지 내장 함수

```
# variance inflation factor(VIF)
install.packages("car")
library(car)
vif(lm(mpg ~ disp+hp+wt+accler, data=car))
```

```
> vif(lm(mpg ~ disp+hp+wt+accler, data=car))
    disp       hp       wt   accler
9.948802 1.313565 8.552679 1.557890
```

→ disp와 wt의 VIF가 10에 가까움

크게 문제되지 않다고 볼 수 있음

CHECK POINT 2 multi-collinearlity

CHECK POINT 1 coefficients & R^2

autompg 데이터의 최종모형

- 선형회귀식

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

- 선형회귀식의 결정계수 $R^2=0.7004$

CHECK POINT 3 residual plot

CHECK POINT 4

outlier or other suspicious trend

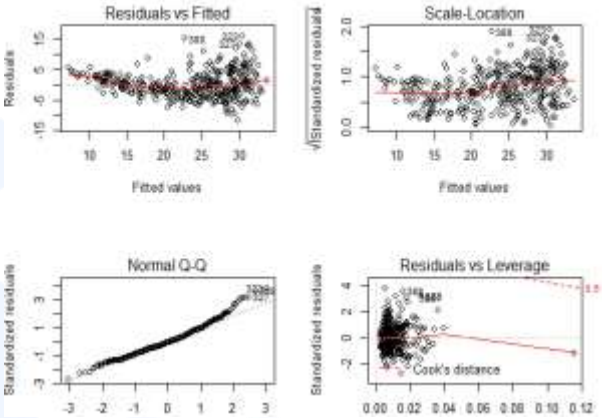
POSTECH

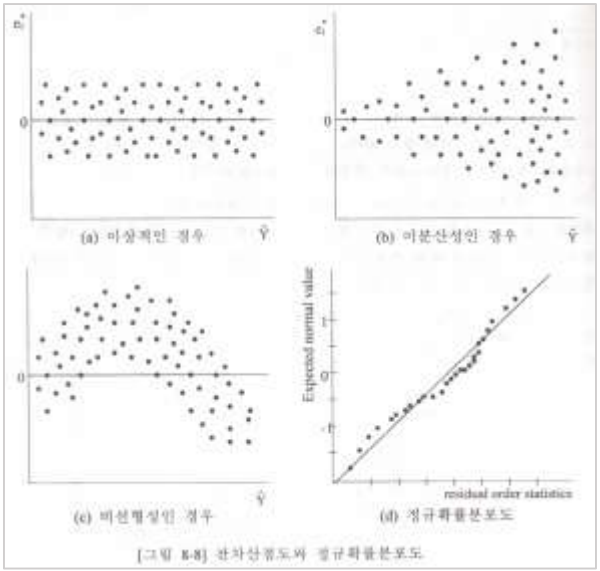
다중회귀분석 – 탐색과 진단

잔차의 산점도

회귀분석의 가정과 진단

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





출처 전치혁, 정민근, 이혜선, 공학응용통계, 홍릉출판사, 2012