Wk9-2 : 데이터마이닝과 예측
- 다중회귀분석 II -

3. 다중회귀분석 – 변수선택방법

• 변수선택방법 – 다수의 독립변수들이 있을 때 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

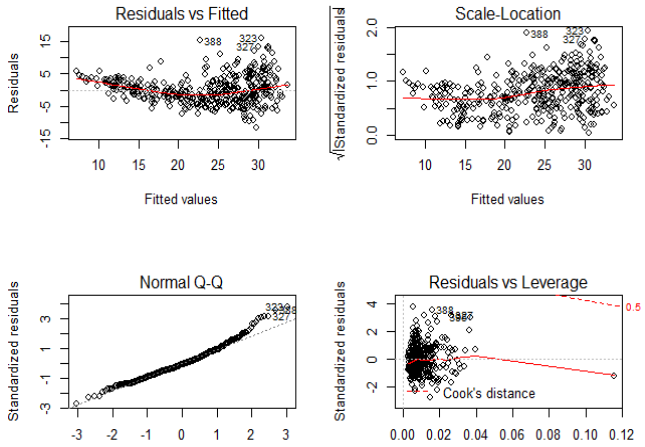
선형회귀식의 결정계수

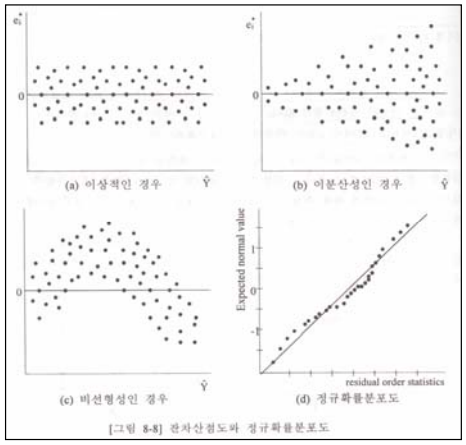
R^2=0.7004

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)
```



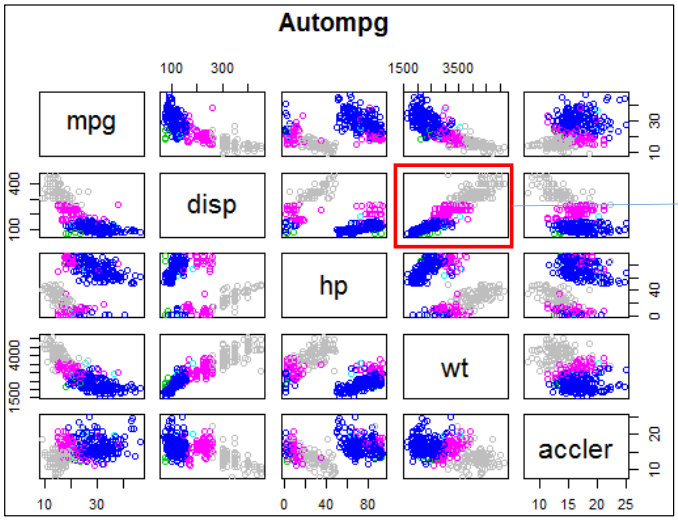


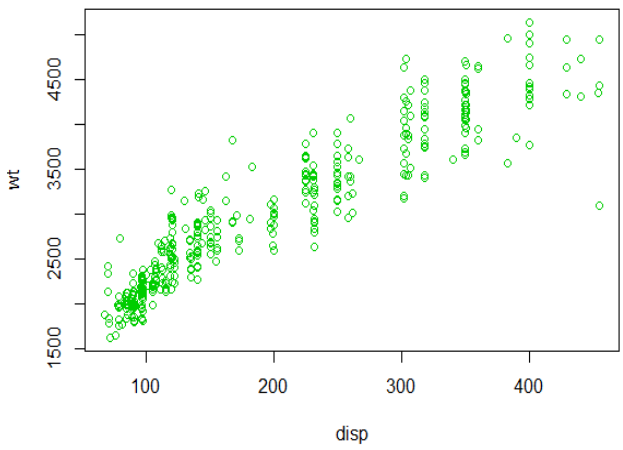
참고 : 전치혁, 정민근, 이혜선, 공학응용통계, 홍릉출판사, 2012

6. 다중회귀분석 – 탐색과 진단

• 다중공선성(Multicollinearity)

- 독립변수들 사이에 상관관계가 매우 높은 경우 발생하는 현상





6. 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
```


6. 다중회귀분석 – 탐색과 진단
9.2 데이터마이닝-다중회귀분석 II

• 분산팽창계수(VIF; Variance Inflation Factor) – 다중공선성의 척도

$$VIF_j = \frac{1}{1 - R_j^2}, \qquad j = 1,2,\dots,k$$

- VIF는 다중공산성으로 인한 분산의 증가를 의미

- R_j^2은 X_j를 종속변수로 하고 나머지 변수를 독립변수로 하는 회귀모형에서의 결정계수

- VIF_j > 10 이상이면 다중공산성 고려

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

6. 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
```

Check point 2 : multi-collinearlity

disp와 wt의 VIF가 10에 가까움

⇒ 크게 문제되지 않다고 볼수 있음

Check point 1 : coefficients & R^2

week9_1의 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

6. 다중회귀분석 – 탐색과 진단

• 변수선택에 대한 R^2 확인 (변수선택방법)

```
# compare R-sqaured in regression
# which one is the most important variable?
summary(lm(mpg ~ disp))
summary(lm(mpg ~ hp))
summary(lm(mpg ~ wt))
summary(lm(mpg ~ accler))
```

연비(MPG)를 예측하기 위해 한 개의 변수만 선택한다면?

R^2?

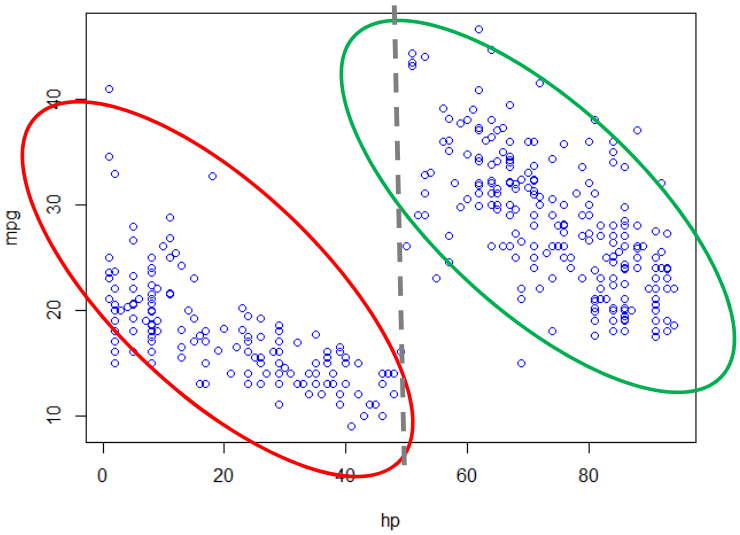
7. 다중회귀모형에 대한 탐색적 분석

• 다중회귀모형 : 데이터탐색(Explanatory Data Analysis) 3rd model : a possible fitting method

```
# more checking point
plot(hp, mpg, col="blue")
```

- 마력(hp)과 연비(MPG)의 관계는 ?
  - 마력이 높을수록 연비는 낮다 (사전지식)
- 마력과 연비간의 산점도에서 발견한 문제는?
  - 두개의 클러스터를 발견

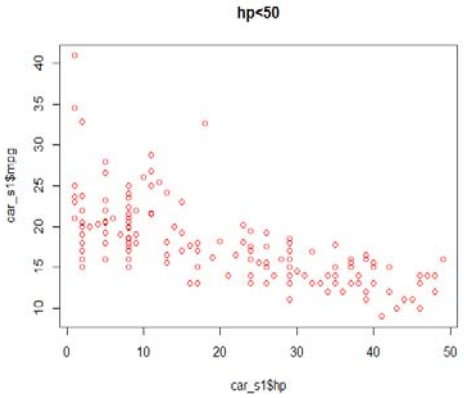
Check point 1 : sign of coefficients



6. 다중회귀분석 – 탐색과 진단

• 다중회귀모형 : 데이터탐색(Explanatory Data Analysis) - subset 생성 (hp<50)

```
# subset data
par(mfrow=c(1,1))
car_s1<-subset(car, hp<50)
plot(car_s1$hp, car_s1$mpg,col=10,  main="hp<50")
```




```
> summary(lm(car_s1$mpg ~ car_s1$hp))

Call:
lm(formula = car_s1$mpg ~ car_s1$hp)

Residuals:
    Min      1Q  Median      3Q     Max
-6.6155 -1.9918 -0.6165  1.4581 19.0596

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 22.06537    0.47732   46.23   <2e-16 ***
car_s1$hp   -0.22495    0.01906  -11.80   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

Residual standard error: 3.562 on 171 degrees of freedom
Multiple R-squared:  0.449,     Adjusted R-squared:  0.4458
F-statistic: 139.3 on 1 and 171 DF,  p-value: < 2.2e-16
```

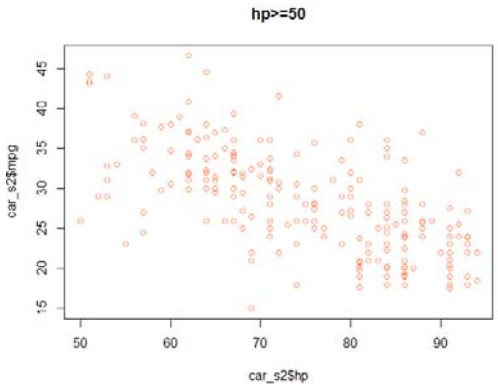
선형회귀식
mpg = 22.07 - 0.22 hp

선형회귀식의 결정계수
R^2=0.45

6. 다중회귀분석 – 탐색과 진단

• 다중회귀모형 : 데이터탐색(Explanatory Data Analysis) - subset 생성 (hp>=50)

```
# subset data hp>=50
car_s2<-subset(car, hp>=50)
plot(car_s2$hp, car_s2$mpg, col="coral", main="hp>=50")
```




```
> summary(lm(car_s2$mpg ~ car_s2$hp))

Call:
lm(formula = car_s2$mpg ~ car_s2$hp)

Residuals:
     Min       1Q   Median       3Q      Max
-15.1987  -3.4734   0.0966   2.7712  14.0820

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  53.0605     2.1498   24.68   <2e-16 ***
car_s2$hp    -0.3313     0.0283  -11.71   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 4.991 on 223 degrees of freedom
Multiple R-squared:  0.3807,    Adjusted R-squared:  0.3
F-statistic: 137.1 on 1 and 223 DF,  p-value: < 2.2e-16
```

선형회귀식
mpg = 53.06 - 0.33 hp

선형회귀식의 결정계수
R^2=0.38