데이터과학을 위한 R프로그래밍

7주차. 상관분석과 회귀모형

이혜선 교수
포항공과대학교 산업경영공학과

목차

7주차. 상관분석과 회귀모형

7주차

2차시

선형회귀모형

POSTECH

회귀분석 - 데이터

autompg 데이터

1. mpg: continuous (연비 : 연속형 변수)
2. cylinders: multi-valued discrete (실린더 : 정수 값)
3. displacement: continuous (배기량 : 연속형 변수)
4. horsepower: continuous (마력 : 연속형 변수)
5. weight: continuous (무게 : 연속형 변수)
6. acceleration: continuous (가속 : 연속형 변수)
7. year: multi-valued discrete (모델 연도 : 정수 값)
8. origin: multi-valued discrete (정수 값)
9. car name: string (unique for each instance) (차 종류 이름)

```
> str(car)
'data.frame':   398 obs. of  9 variables:
 $ mpg     : num  18 15 18 16 17 15 14 14 14 15 ...
 $ cyl     : int  8 8 8 8 8 8 8 8 8 8 ...
 $ disp    : num  307 350 318 304 302 429 454 440 455 390 ...
 $ hp      : int  17 35 29 29 24 42 47 46 48 40 ...
 $ wt      : int  3504 3693 3436 3433 3449 4341 4354 4312 4425 3850
 $ accler : num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
 $ year    : int  70 70 70 70 70 70 70 70 70 70 ...
 $ origin : int  1 1 1 1 1 1 1 1 1 1 ...
 $ carname: Factor w/ 305 levels "amc ambassador brougham",..: 50 3
2 2 ...
```

```r
# lec7_2.r : Linear model
# Regression

library(dplyr)

# set working directory
setwd("D:/tempstore/moocr/wk8")

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

# subset with cyl=4,6,8
car1<-filter(car, cyl==4 | cyl==6 | cyl==8)
attach(car1)
table(cyl)
```

```
>   table(cyl)
cyl
  4   6   8
204  84 103
```

POSTECH

회귀분석 - 단순회귀모형

☑ 단순회귀모형 : lm(y변수~x변수, data=)

1. 단순회귀모형
종속변수 : mpg(연비),독립변수:wt(차량무게)

```
# 1. simple Regression(independent variable : wt)
r1<-lm(mpg~wt, data=car1)
summary(r1)
anova(r1)
```

```
> r1<-lm(mpg~wt, data=car1)
> summary(r1)

Call:
lm(formula = mpg ~ wt, data = car1)

Residuals:
    Min      1Q  Median      3Q     Max
-9.6770 -2.7567 -0.3636  2.1120 16.3712

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 46.600189   0.779849   59.76   <2e-16 ***
wt          -0.007759   0.000252  -30.79   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.239 on 389 degrees of freedom
Multiple R-squared:  0.709,     Adjusted R-squared:  0.7083
F-statistic: 947.9 on 1 and 389 DF,  p-value: < 2.2e-16
```

선형회귀식

y(mpg)=46.60-0.0077(wt)

선형회귀식의 결정계수

R^2=0.709

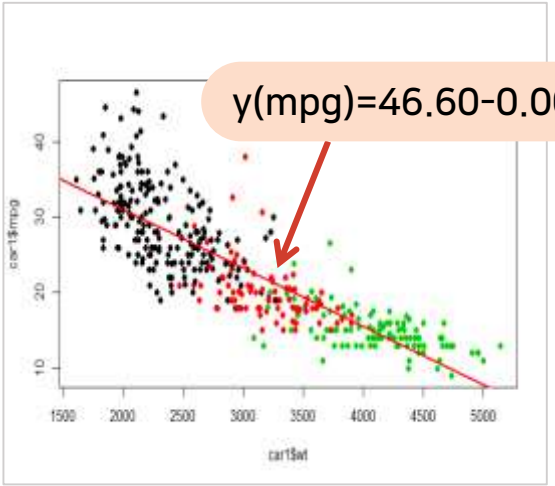
POSTECH

회귀분석 - 단순회귀모형

산점도에 회귀선 그리기

```
# (lec4_2.r) scatterplot with best fit lines
par(mfrow=c(1,1))
plot(wt, mpg,  col=as.integer(car1$cyl), pch=19)
# best fit linear line
abline(lm(mpg~wt), col="red", lwd=2, lty=1)
```

plot(x축변수, y축변수)
abline : add line (선을 추가하는 함수)
lm(y변수~x변수) : lm은 linear model(선형모형)의 약자

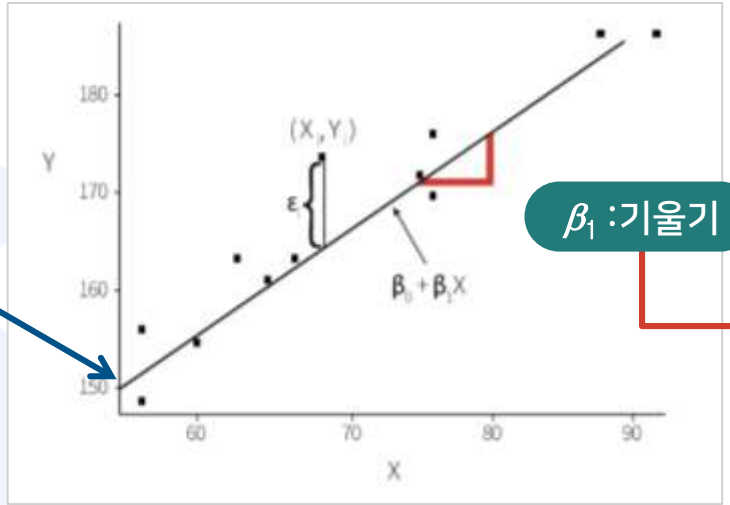


POSTECH

회귀분석의 목적

- 회귀분석의 목적 : 예측(prediction)과 추정(estimation)
  - 선형모형 : 독립변수와 종속변수간의 관계가 선형식으로 적합

모형 : $Y_i=\beta_0 + \beta_1 X_i + \varepsilon_i, \quad i=1, 2, \cdots, n$



β_0 :Y절편

β_1 :기울기

최소자승법 (least squares method)

: 예측값과 관측치간의 오차를 최소화 시키는 회귀계수를 추정

POSTECH

회귀분석 – 모형의 적합도

☑ 회귀식에 의해 설명되는 부분(SSR)과 설명되지 않는 부분(SSE)

```
> anova(r1)
Analysis of Variance Table

Response: mpg
           Df  Sum Sq Mean Sq F value    Pr(>F)
wt          1 17029.1   17029  947.87 < 2.2e-16 ***
Residuals 389  6988.6      18
```

$$R^2 = \frac{SSR}{SST} = \frac{17029}{24017} = 0.709$$

- R^2는 1에 가까울수록 회귀식에 의해 적합되는 부분이 높음
- R^2는 0에 가까우면 주어진 독립변수들에 의해 설명(예측 혹은 적합)되는 부분이 없다고 할 수 있다

SST=total sum of squares
SSR=regression sum of squares
SSE=error(residual) sum of squares

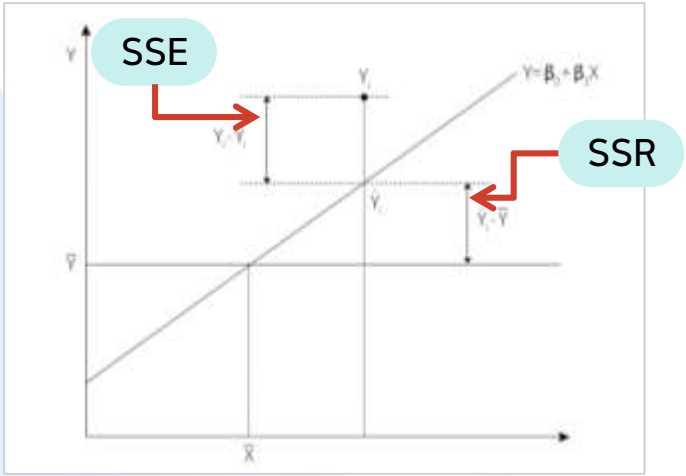
POSTECH

회귀분석 – 모형의 적합도

모형의 적합도와 결정계수 (R^2) : $0 \le R^2 \le 1$

: 전체제곱합(SST)에 대한 회귀제곱합(SSR)의 비율, 즉 모형으로 설명 할 수 있는 부분의 비율

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$



전체제곱의 분할:SST=SSR+SSE

전체제곱합(SST) : $SST = \sum(Y_i - \overline{Y})^2$

회귀제곱합(SSR) : $SSR = \sum(\hat{Y}_i - \overline{Y})^2$

잔차제곱합(SSE) : $SSE = \sum(Y_i - \hat{Y}_i)^2$

POSTECH

회귀분석 : 단순회귀모형

2. 단순회귀모형
종속변수 : mpg(연비), 독립변수: disp(배기량)

```
# 2. simple Regression(independent variable : disp)
r2<-lm(mpg~disp, data=car1)
summary(r2)
anova(r2)
```

```
> summary(r2)

Call:
lm(formula = mpg ~ disp, data = car1)

Residuals:
     Min       1Q   Median       3Q      Max
-10.0627  -3.0037  -0.6113   2.3110  18.5978

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 35.49479    0.48640   72.97   <2e-16 ***
disp        -0.06142    0.00220  -27.93   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

Residual standard error: 4.533 on 389 degrees of freedom
Multiple R-squared:  0.6672,    Adjusted R-squared:  0.6663
F-statistic: 779.8 on 1 and 389 DF,  p-value: < 2.2e-16
```

선형회귀식

y(mpg)=35.49-0.0614(disp)

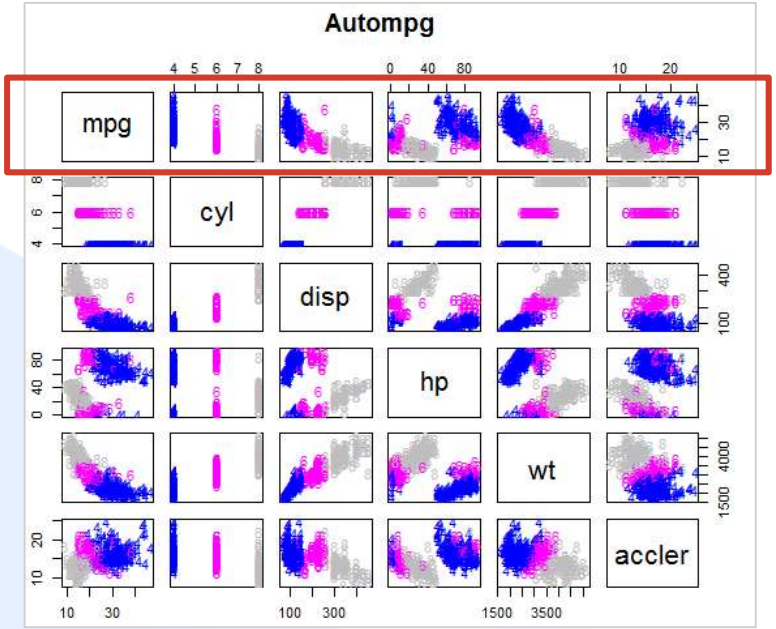
선형회귀식의 결정계수

R^2=0.67

POSTECH

회귀분석 : 단순회귀모형

- pairwise scatterplot




```
# new variable lists
vars1<-c("disp", "wt", "accler", "mpg")
# pariwise plot
pairs(car[vars1], main ="Autompg",cex=1,
```