데이터과학을 위한 R프로그래밍

4주차. 데이터시각화

이혜선 교수
포항공과대학교 산업경영공학과

목차

4주차. 데이터시각화

4주차

2차시

R 그래픽 II
(상자그림, 산점도)

POSTECH

상자그림 (Boxplot, 1차원)

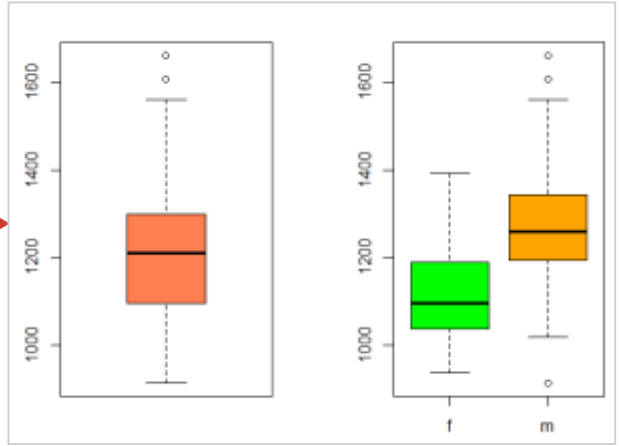
☑ 상자그림 : boxplot(변수이름, col=c("colname"))

```
# 2. boxplot
par(mfrow=c(1,2))
# 2-1 boxplot for all data
boxplot(wt, col=c("coral"))
```

☑ 그룹별 상자그림 : boxplot(변수이름~그룹이름, col=c("col1", "col2")

```
# 2-2 boxplot by group variable (female, male)
boxplot(wt~sex, col = c("green", "orange"))
```

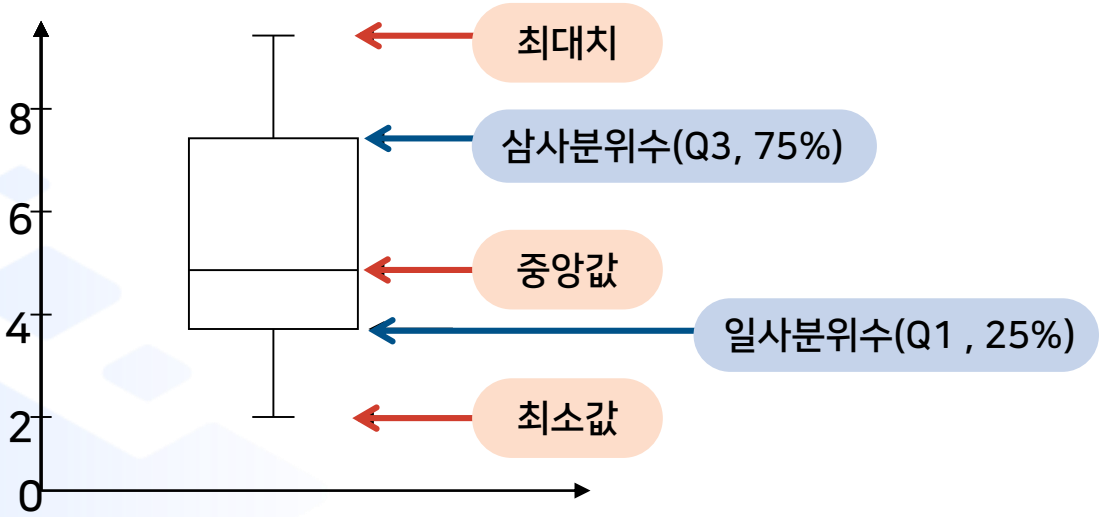
par(mfrow=c(1,2)) : 그래프화면의 분할을 행(row)는 1행으로, 열은 2열로 하라는 의미



상자그림 (Boxplot, 1차원)

상자그림 설명

- 데이터의 분포를 사분위수를 중심으로 설명해주는 그림



✠ Q1, Q3로부터 ±1.5 IQR 넘는 값 (이상치로 볼 수 있음)

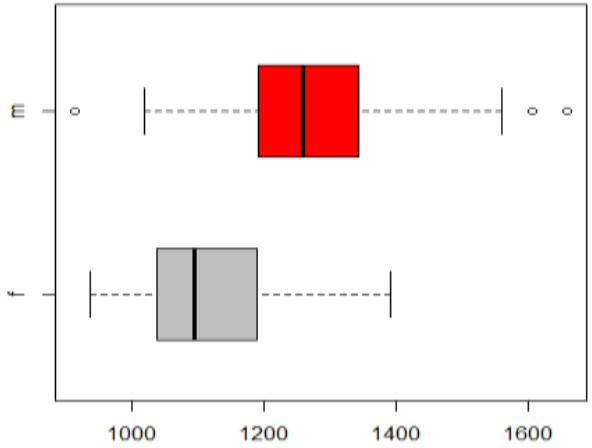
POSTECH

상자그림 (Boxplot, 1차원)

☑ 수평 상자그림 : boxplot(변수이름, col=c("colname"), horizontal=TRUE)

```
# 2-3 horizontal boxplot
par(mfrow=c(1,1))
boxplot(wt~sex, boxwex=0.5, horizontal=TRUE, col = c("grey", "red"))
```

수평으로 상자그림을 그릴 수 있음



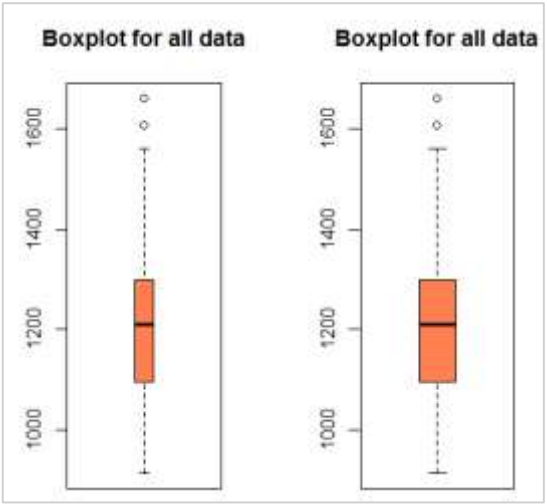
POSTECH

상자그림 (Boxplot, 1차원)

☑ 상자그림 : boxplot(변수이름, col=c("colname"), boxwex=)

```
# 2-4 box width boxwex (width of box)
par(mfrow=c(1,2))
boxplot(wt, boxwex = 0.25, col=c("coral"),  main="Boxplot for all data")
boxplot(wt, boxwex = 0.5, col=c("coral"), main="Boxplot for all data")
```

boxwex : 그림상자의 폭을 조정



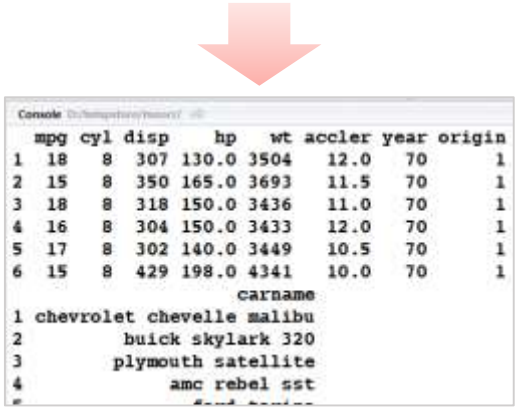
POSTECH

autompg(차의 연비) 데이터

- autompg 데이터 (lec3_3.R에서 사용)
  - mpg: continuous (연비 : 연속형변수)
  - cylinders: multi-valued discrete (실린더 : 정수값)
  - displacement: continuous (배기량 : 연속형변수)
  - horsepower: continuous (마력 : 연속형변수)
  - weight: continuous (무게 : 연속형변수)
  - acceleration: continuous (가속 : 연속형변수)
  - year: multi-valued discrete (모델연도 : 정수값)
  - origin: multi-valued discrete (정수값)
  - car name: string (unique for each instance) (차종류 이름)

```
# use autompg data (lec3_3.R)
car<-read.csv("autompg.csv")
head(car)

attach(car)
```



```
  mpg cyl disp    hp   wt accler year origin
1  18   8  307 130.0 3504    12.0   70      1
2  15   8  350 165.0 3693    11.5   70      1
3  18   8  318 150.0 3436    11.0   70      1
4  16   8  304 150.0 3433    12.0   70      1
5  17   8  302 140.0 3449    10.5   70      1
6  15   8  429 198.0 4341    10.0   70      1
                  carname
1 chevrolet chevelle malibu
2         buick skylark 320
3        plymouth satellite
4             amc rebel sst
```

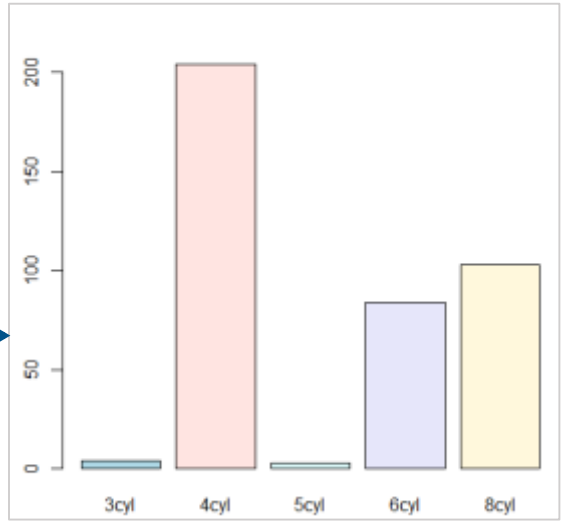
POSTECH

막대 그림 (barplot, 2차원)

- barplot(변수빈도, col=c("col1", "col2", ..))

막대그림을 그리기 위해서는 우선 table(변수이름)을 이용하여 빈도를 계산함

```
# 3. bar plot with cylinder count
par(mfrow=c(1,1))
table(cyl)
freq_cyl<-table(cyl)
names(freq_cyl) <- c ("3cyl", "4cyl", "5cyl", "6cyl",
                      "8cyl")
barplot(freq_cyl, col = c("lightblue", "mistyrose", "lightcyan",
                          "lavender", "cornsilk"))
```



autompg 데이터의 cylinder 빈도

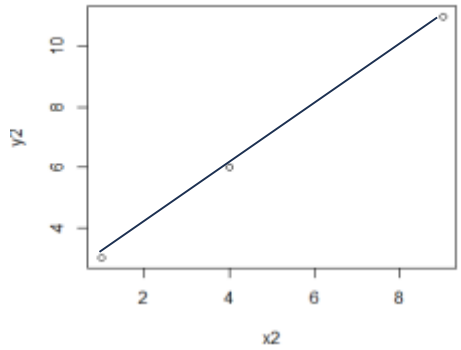
```
> table(cyl)
cyl
  3   4   5   6   8
  4 204   3  84 103
```

POSTECH

산점도 (scatterplot) 2차원

산점도 : plot(x, y)

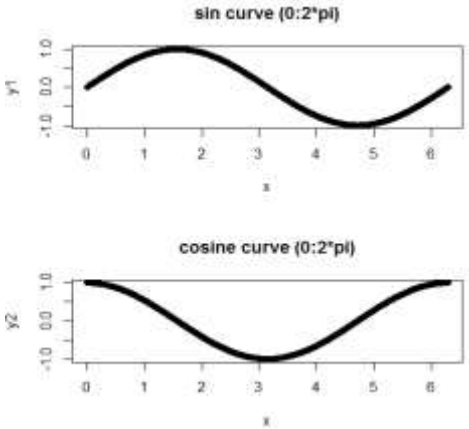
```
# 4-1 simple plot
par(mfrow=c(1,1))
x2<-c(1,4,9)
y2<-2+x2
plot(x2, y2)
```



x와 y간의 관계를 보여주는 그래프

```
par(mfrow=c(2,1))
x<-seq(0, 2*pi, by=0.001)
y1<-sin(x)
plot(x,y1, main="sin curve (0:2*pi)")

y2<-cos(x)
plot(x,y2,main="cosine curve (0:2*pi)" )
```

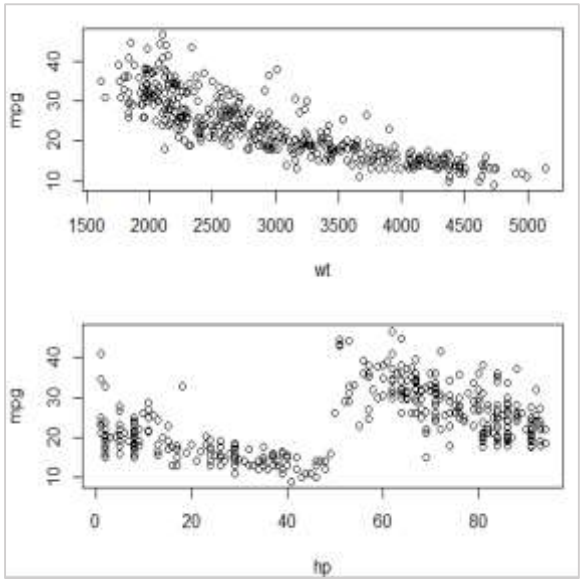


POSTECH

산점도 (scatterplot) 2차원

- wt(차의 무게)과 mpg(연비)간의 산점도 : plot(wt, mpg)
- hp(마력)과 mpg(연비)간의 산점도 : plot(hp, mpg)

```
par(mfrow=c(2,1))
plot(wt, mpg)
plot(hp, mpg)
```



산점도에 대한 해석과 설명

- 차의 무게가 무거울수록 연비는 낮다.
- 마력과 연비간의 산점도에서는 두 개의 클러스터가 보임(클러스터내에서는 마력이 높을수록 연비가 낮음)

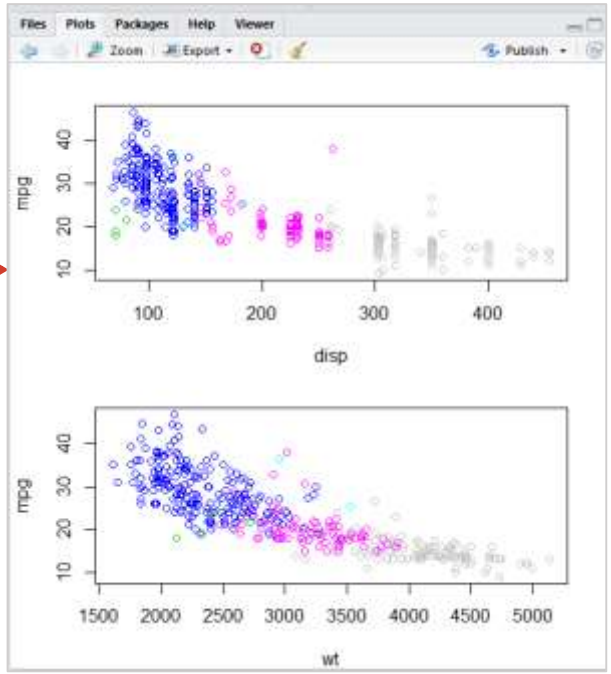
POSTECH

산점도 (scatterplot) 2차원

- plot(x, y, col=as.integer(그룹변수))

색으로 표시

```
par(mfrow=c(2,1), mar=c(4,4,2,2))
plot(disp, mpg, col=as.integer(car$cyl))
plot(wt, mpg,  col=as.integer(car$cyl))
```



autompg 데이터의 cylinder 빈도

```
> table(car$cyl)

  3   4   5   6   8
  4 204   3  84 103
```

POSTECH

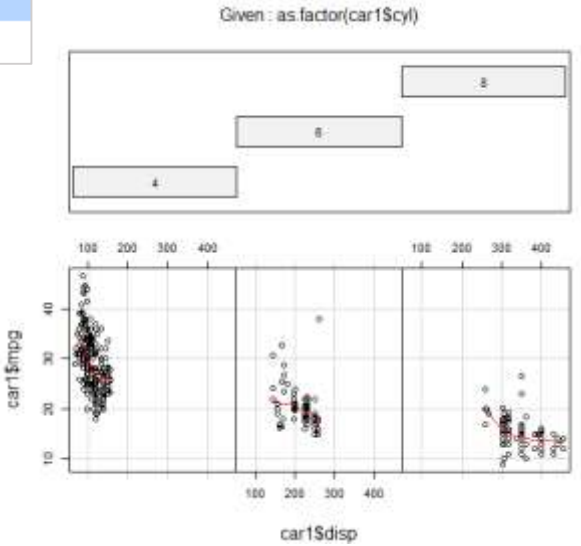
산점도 (scatterplot) 2차원

- Conditioning plot : coplot(y~x | z) z는 factor(그룹)

그룹에 따른 (x와 y)간 산점도

Subset 데이터 활용 : 4,6,8cylinder

```
car1<-subset(car, cyl==4 | cyl==6 | cyl==8)
coplot(car1$mpg ~ car1$disp | as.factor(car1$cyl), data = car1,
       panel = panel.smooth, rows = 1)
```



그룹별 산점도

- cylinder에 따른 차이를 보여줌
- 4cyl, 6cyl, 8cyl별로 (배기량과 연비)간 관계를 구체적으로 해석할 수 있음

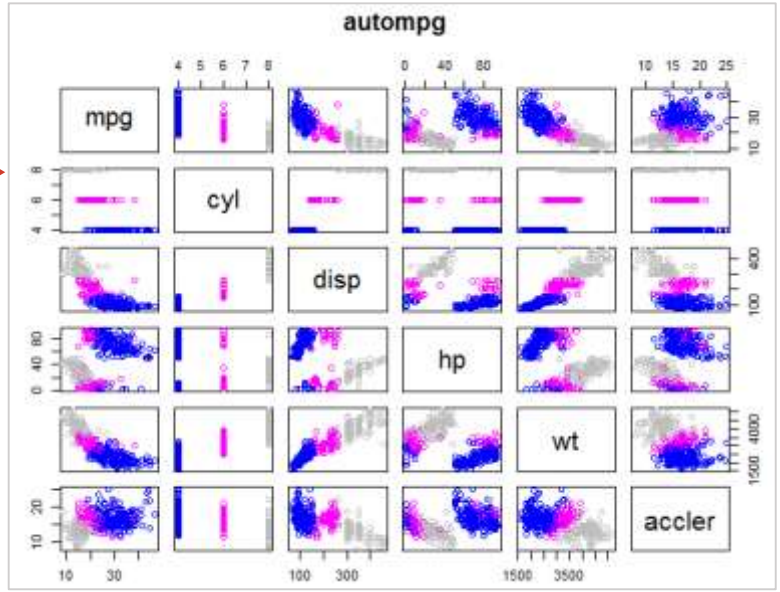
POSTECH

산점도 (scatterplot) 2차원

- pairwise scatterplot : pairs(변수리스트)

Subset 데이터 활용 : 4,6,8cylinder

```
pairs(car1[,1:6], col=as.integer(car1$cyl), main = "autompg")
```

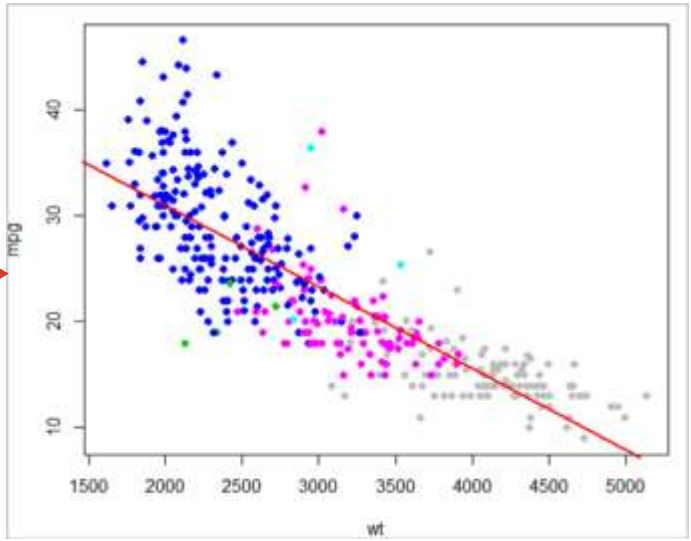


POSTECH

산점도 (scatterplot) 2차원

- 최적 적합 함수 추정 (선형회귀모형, 비선형회귀모형)
  - lm(y변수~x변수) : 여기서 lm은 linear model(선형모형)의 약자
  - abline : add line (선을 추가하는 함수)

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



POSTECH

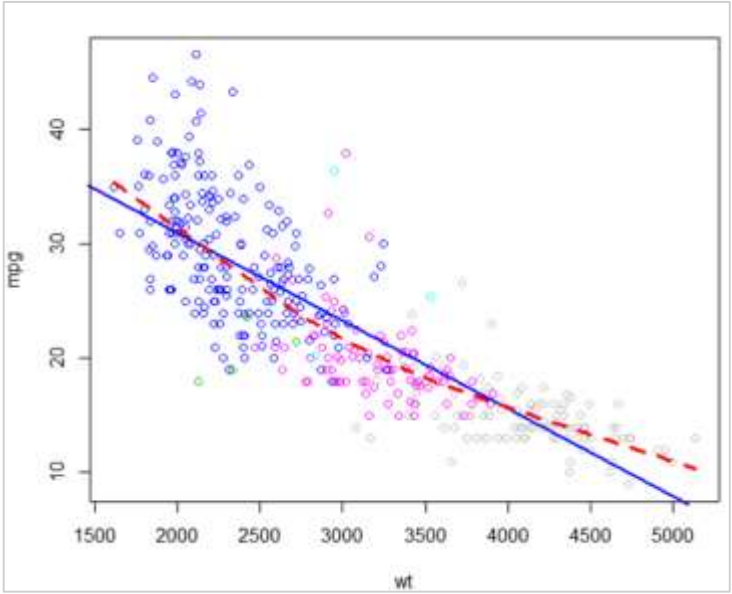
산점도 (scatterplot) 2차원

최적 적합 함수 추정 (비선형회귀모형, lowess 이용)

- lowess : locally-weighted polynomial regression (see the references).

```
plot(wt, mpg,  col=as.integer(car$cyl))
# best fit linear line
abline(lm(mpg~wt), col="blue", lwd=2, lty=1)

# lowess : smoothed line, nonparmetric fit line
lines(lowess(wt, mpg), col="red", lwd=3, lty=2)
help(lowess)
```



참고문헌 : lowess

Cleveland, W. S. (1979) Robust locally weighted regression and smoothing scatterplots. *J. American Statistical Association* **74**, 829–836.
Cleveland, W. S. (1981) LOWESS: A program for smoothing scatterplots by robust locally weighted regression. *The American Statistician* **35**, 54