제6강
다중변수 자료의 탐색

Section 01

산점도

1. 산점도

- 다중변수 자료(또는 다변량 자료) : 변수가 2개 이상인 자료
- 다중변수 자료는 2차원 형태를 나타내며, 이는 매트릭스나 데이터 프레임에 저장하여 분석
- 산점도(scatter plot)란 2개의 변수로 구성된 자료의 분포를 알아보는 그래프

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa

그림 6-1 다중변수 자료인 iris 데이터셋

1. 산점도

1. 두 변수 사이의 산점도

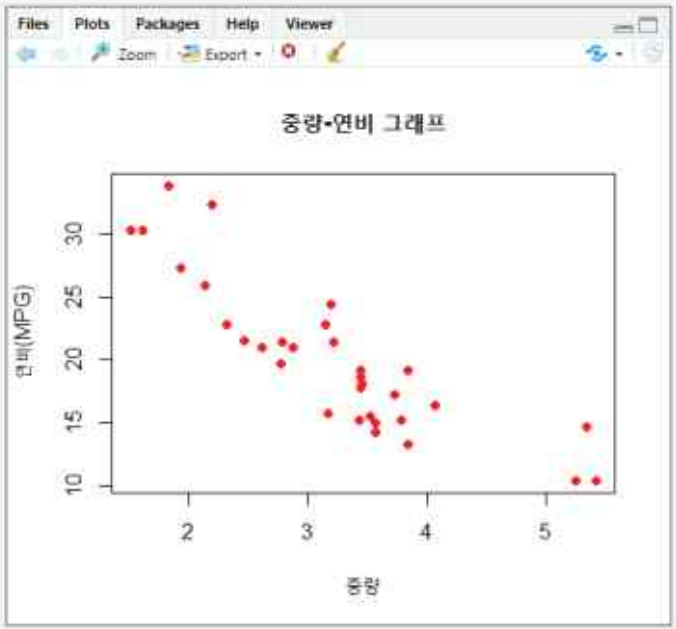
- mtcars 데이터셋에서 자동차의 중량(wt)과 연비(mpg) 사이의 관계

코드 6-1

```
wt <-mtcars$wt                          # 중량 자료
mpg <- mtcars$mpg                       # 연비 자료
plot(wt, mpg,                           # 2개 변수(x축, y축)
          main="중량-연비 그래프",       # 제목
          xlab="중량",                  # x축 레이블
          ylab="연비(MPG)",             # y축 레이블
          col="red",                    # point의 color
          pch=19)                       # point의 종류
```

산점도는 2개의 변수로 구성된 자료의 분포를 알아보는 기법임을 기억하자.
아울러 두 변수의 데이터 분포를 나타내는 것이기에 두 개의 변수에 대한 자료가 필요하다.

1. 산점도



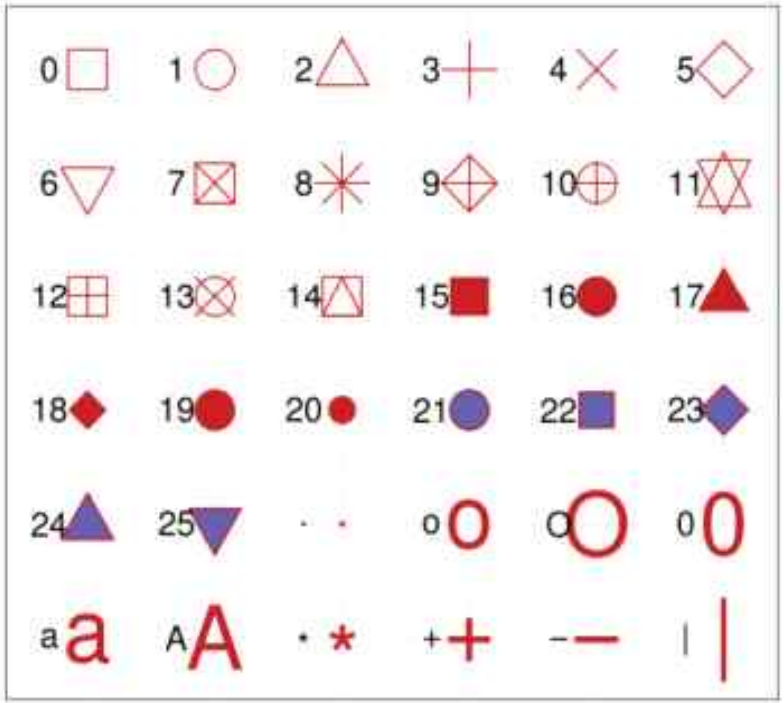


그림 6-2 pch 값에 따른 점의 모양

- 중량이 증가할수록 연비는 감소하는 경향을 확인

1. 산점도

2. 여러 변수들 간의 산점도

산점도는 기본적으로 2개의 변수에 대해서 작성하는 것이기 때문에 변수가 여러 개인 자료의경우는 다소 불편하다.이에 대체하는 방법으로 pairs()함수를 이용하면 편리하다.

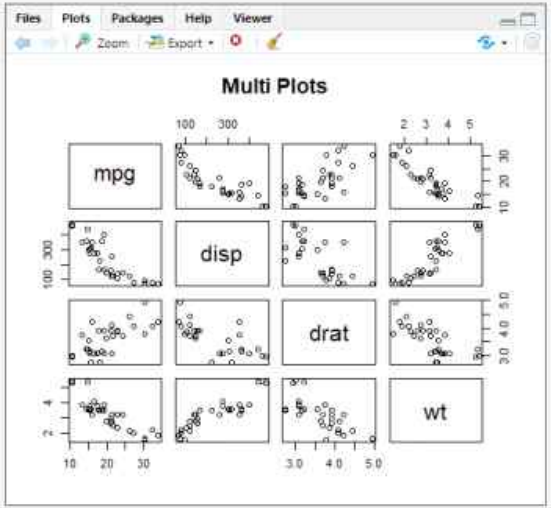
코드 6-2

```
vars <- c("mpg","disp","drat","wt")          # 대상 변수
target <- mtcars[,vars]
head(target)
pairs(target,                                # 대상 데이터
      main="Multi Plots")
```

```
> vars <- c("mpg","disp","drat","wt")      # 대상 변수
> target <- mtcars[,vars]
> head(target)
                   mpg  disp  drat    wt
Mazda RX4         21.0   160  3.90 2.620
Mazda RX4 Wag     21.0   160  3.90 2.875
Datsun 710        22.8   108  3.85 2.320
Hornet 4 Drive    21.4   258  3.08 3.215
Hornet Sportabout 18.7   360  3.15 3.440
Valiant           18.1   225  2.76 3.460
```

1. 산점도

```
> pairs(target,                    # 대상 데이터
+       main="Multi Plots")
```



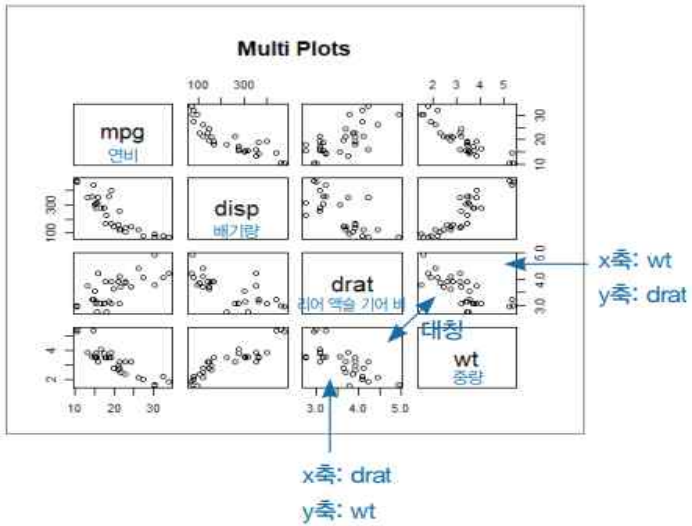


그림 6-3 다중 산점도의 예

위 그래프는 4개의 변수에 대해서 다중 산점도를 작성하였다. 다중 산점도에는 4개의 변수가 대각선으로 표기되며, 두 변수가 만나는 지점에 두 변수의 산점도가 나타난다. 다중 산점도는 대각선을 기준으로 하여 오른쪽 위의 산점도와 왼쪽 아래의 산점도들이 대칭을 이루고 있다. 즉, x축과 y축이 바뀌었다는 것이다. 다중 산점도의 예를 보면 disp, wt산점도는 한쪽이 증가하면 다른 쪽도 증가하고, drat, wt산점도의 경우는 한 쪽이 증가하면 한 쪽이 감소하는 것을 알수가 있다. 이와 같이 다중 산점도는 여러 변수들 간의 추세를 한눈에 파악할 수 있는 장점이 있다.

1. 산점도

3. 그룹 정보가 있는 두 변수의 산점도

- 그룹 정보를 알고 있다면 산점도를 작성 시 각 그룹별 관측값들을 다른 색깔과 점의 모양으로 표시할 수 있음
- 이렇게 작성된 산점도는 두 변수 간의 관계 뿐만 아니라 그룹 간의 관계도 파악할 수 있어서 편리

코드 6-3

```
iris.2 <- iris[,3:4]                    # 데이터 준비
point <- as.numeric(iris$Species)       # 점의 모양
point                                   # point 내용 출력
color <- c("red","green","blue")        # 점의 컬러
plot(iris.2,
        main="Iris plot",
        pch=c(point),
        col=color[point])
```

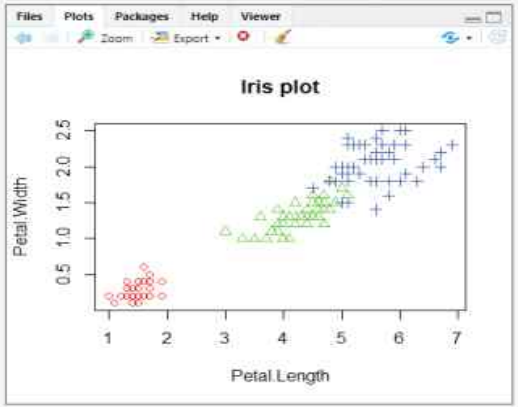
```
> iris.2 <- iris[,3:4]                  # 데이터 준비
```

1. 산점도

```
> point <- as.numeric(iris$Species)        # 점의 모양
> point                                    # point 내용 출력
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [32] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [63] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [94] 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[125] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
> color <- c("red","green","blue")         # 점의 컬러
```

좌측 코드는 품종별로 점의 모양을 달리하기 위하여 숫자 타입으로 변환시켜 color속성에 대입한다.

```
> plot(iris.2,
+      main="Iris plot",
+      pch=c(point),
+      col=color[point])
```



- Petal.Length(꽃잎의 길이)의 길이가 길수록 Petal.Width(꽃잎의 폭)도 커짐
- setosa 품종은 다른 두 품종에 비해 꽃잎의 길이와 폭이 확연히 작음
- virginica 품종은 다른 두 품종에 비해 꽃잎의 길이와 폭이 제일 큼
- 이와 같이 산점도를 그릴 때 그룹 정보를 표시하면 변수 간의 관계와 그룹 간의 관계를 함께 관찰할 수가 있다.

감사합니다.