4강. 이변량 자료의 시각화 1

◈ 담당교수 : 허명회

들어가기

■ 주요용어

용어	해설
산점도	연속형 자료에 대한 시각화 기법으로 두 변수 간 상관회귀 관계를 살펴보는 데 매우 효과적이다
이변량 밀도 추정	이변량 자료로부터 모분포의 밀도를 추정하기 위해서 커널 함수를 사용한다. 등고선을 덧붙여 시각화 효과를 높인다.
육각형 칸에 넣기	관측 개체를 육각형 칸에 넣어 얻은 돗수(count)를 칸 별로 색의 농담(濃淡)으로 나타낸 일종의 산점도이다.
회귀적 관계	이변량 자료로부터 추정된 회귀함수 y = f(x)를 산점도에 넣어 두 변수 간 관계를 시각화한다. 회귀 함수 형태는 직선과 곡선이 있다.

연습문제

1. R에서 이변량 밀도 추정을 위한 함수는 무엇인가?

정답 : bkde2D( )

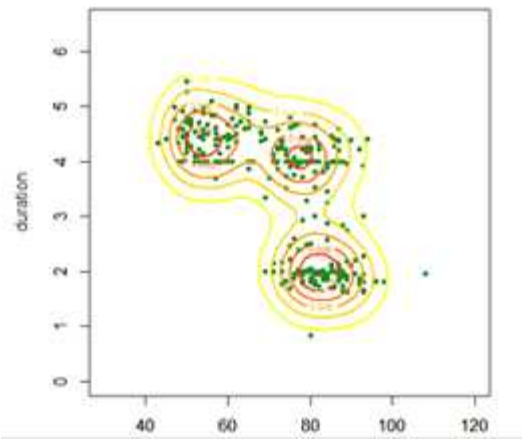
2. R에서 이변량 회귀 관계를 비모수적으로 추정하는 함수는 무엇인가?

정답 : lowess( )

3. R MASS 팩키지의 geyser 자료에서 Old Faithful 간헐천의 대기시간(waiting)과 분출시간(duration) 간 산점도를 작성하고 밀도 등고선을 넣어라.

정답 : library(MASS)
data(geyser)

```
windows(height=7, width=6.4)
plot(geyser$waiting, geyser$duration, xlim=c(30,120),
    ylim=c(0,6.5), col="forestgreen", pch=20, xlab="waiting",
    ylab="duration", main="geyser")
library(KernSmooth)
density <- bkde2D(geyser, bandwidth=c(5,0.5))
par(new=T)
contour(density$x1, density$x2, density$fhat, xlim=c(30,120),
 ylim=c(0,6.5), col=heat.colors(7)[7:1], nlevels=7, lwd=2)
```

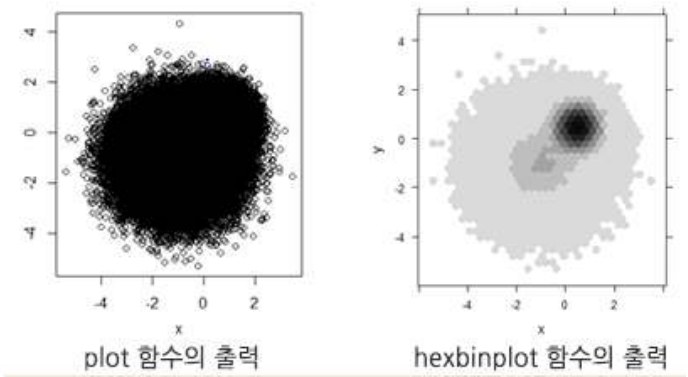


4. 다음 모의생성 자료 (x, y)에 대해 적절한 산점도를 그려라.

```
n ← 100000
x ← c(rnorm(n, -1, 1), rnorm(2*n, 0.5, 0.5))
y ← c(rnorm(n, -1, 1), rnorm(2*n, 0.5, 0.5))
```

정답 : plot(y ~ x)

```
library(hexbin)
hexbinplot(y ~ x, colorkey=F)
```



plot 함수의 출력 hexbinplot 함수의 출력

정리하기

1. 산점도(scatterplot)는 연속형 이변량 자료를 시각화하는 기본 그래프로서 관측개체의 이변량 좌표에 점을 찍어 얻는다. 일반적 상황에서는 'aspect', 즉 가로 대 세로의 비를 1로 한다. 수평 축과 수직 축에 이름이 붙어야 하고 좌우상하에 10% 정도의 여백이 있어야 보기가 좋다.

2. 이변량 커널 밀도추정(kernel density estimation)은 커널 함수를 써서 모분포의 확률밀도함수를 추정하는 방법이다. 이 때, 띠 너비를 크게 하면 밀도추정 함수가 밋밋해지고 띠 너비를 작게 하면 밀도추정 함수가 울퉁불퉁해진다.

3. 개체 수가 10,000 이상인 큰 자료에 대하여는 단순한 산점도는 적절하지 않다. 육각형 칸에 넣기(hexagonal binning)와 같은 특별한 기법이 적용되어야 한다.

4. X와 Y의 산점도에서 $X=x$에서 Y의 평균 반응이 회귀함수 $y=f(x)$이다. 직선 형태의 회귀함수 또는 곡선 형태의 회귀함수를 산점도에 넣어 시각화 효과를 높일 수 있다.

참고자료

1. 허명회 (2010). R을 활용한 탐색적 자료분석 (1판 3쇄), 자유아카데미.
2. 허명회 (2014). 데이터 시각화, 자유아카데미.