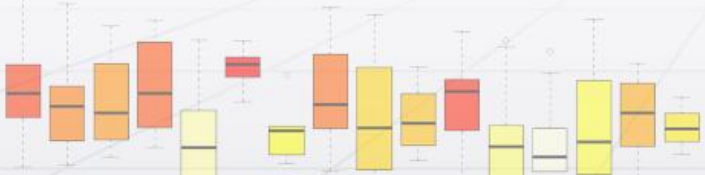
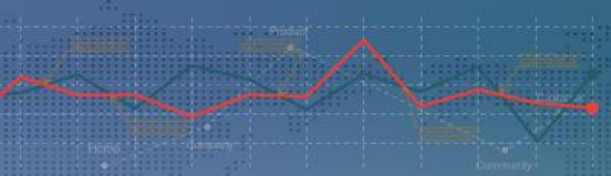
4강 이변량 데이터의 시각화 (1)

고려대학교 통계학과 허명회 교수

1. 산점도
2. 이변량 밀도
3. '큰' 자료의 산점도
4. 회귀적 관계





1. 산점도 (scatterplot)

- 산점도
- 보기 1. 중간·기말 성적 자료

1 산점도

산점도 (scatterplot)는….

- 이변량 연속형 자료점들을 2차원 평면에 넣은 그래프

$$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$$

- 회귀적 관계 ($X \rightarrow Y$) : X를 수평 축에, Y를 수직 축에 넣음
- 기본원칙 : 가로와 세로의 비는 1 : 1로 (aspect = 1)

1 산점도

보기 1. 중간·기말 성적 자료

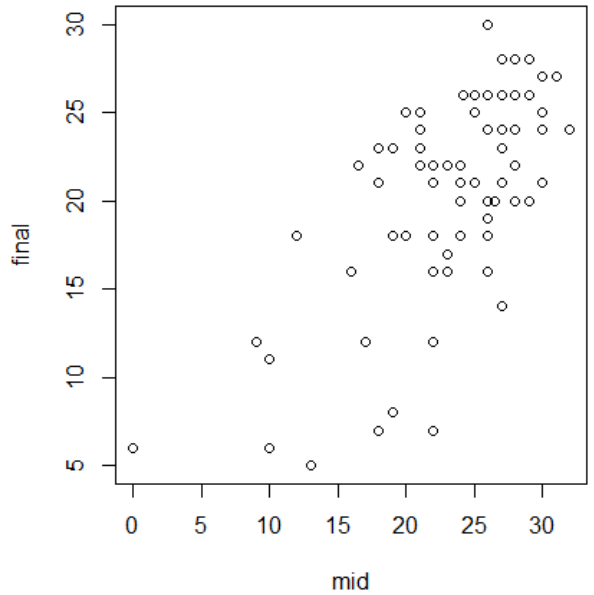
R 스크립트

```
exam <- read.table("exam scores_2012.txt", header=T)
str(exam)
  'data.frame':    85 obs. of  2 variables:
   $ mid  : num  29 30 29 13...
   $ final: int  NA 21 20 5 ...
attach(exam)
windows(height=5.5,width=5)
plot(mid,final)
```

1 산점도

보기 1. 중간·기말 성적 자료

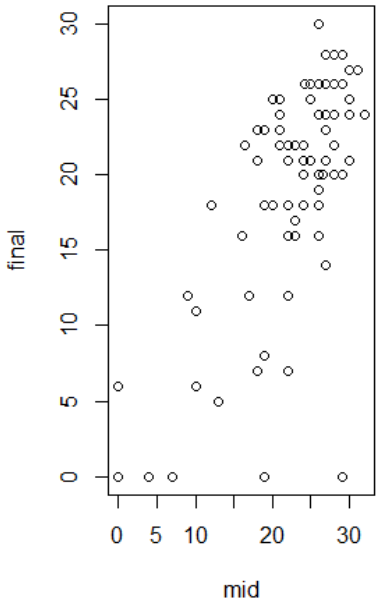
- 출력 그래프 (aspect = 1)

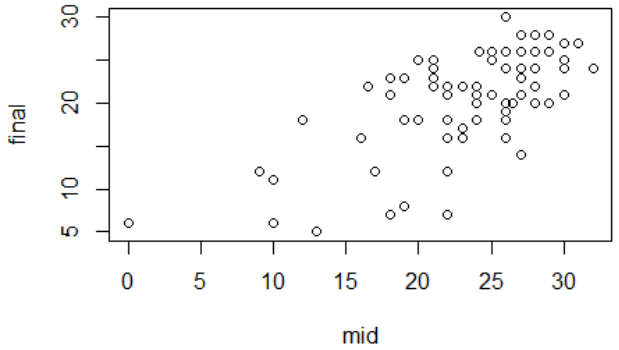


1 산점도

보기 1. 중간·기말 성적 자료

- 출력 그래프 (aspect ≠ 1)





1 산점도

보기 1. 중간·기말 성적 자료

- 결측값 (missing value)

```
> summary(exam)
      mid              final
 Min.   :  0.00   Min.   :  5.00
 1st Qu.: 19.50   1st Qu.: 18.00
 Median : 24.00   Median : 22.00
 Mean   : 22.76   Mean   : 20.38
 3rd Qu.: 27.00   3rd Qu.: 24.00
 Max.   : 32.00   Max.   : 30.00
 NA's   :  2      NA's   :  6
```

* NA = not available (결측)

1 산점도

보기 1. 중간·기말 성적 자료

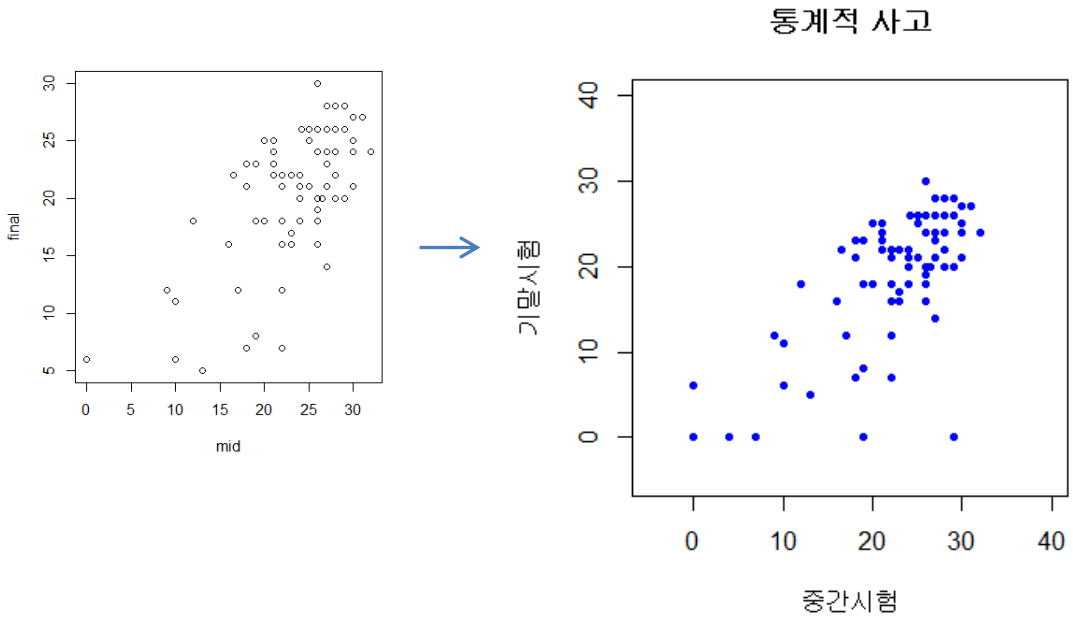
- R 스크립트 (추가)

```
mid[is.na(mid)] <- 0
final[is.na(final)] <- 0
windows(height=5.5,width=5)
plot(mid,final,pch=20,
    xlim=c(-5,40),ylim=c(-5,40),
    col="blue",
    xlab="중간시험",ylab="기말시험",
    main="통계적 사고")
```

1 산점도

보기 1. 중간·기말 성적 자료

출력 그래프



1 산점도

보기 1. 중간·기말 성적 자료

- 격자형 패턴의 수정 : 정수값 자료의 경우

$$x \rightarrow x + \varepsilon_x \,, \quad y \rightarrow y + \varepsilon_y$$

여기서 ε_x, ε_y는 $(-0.5,\ 0.5)$ 에서의 균일분포 임의수
(uniform random number)

1 산점도

보기 1. 중간·기말 성적 자료

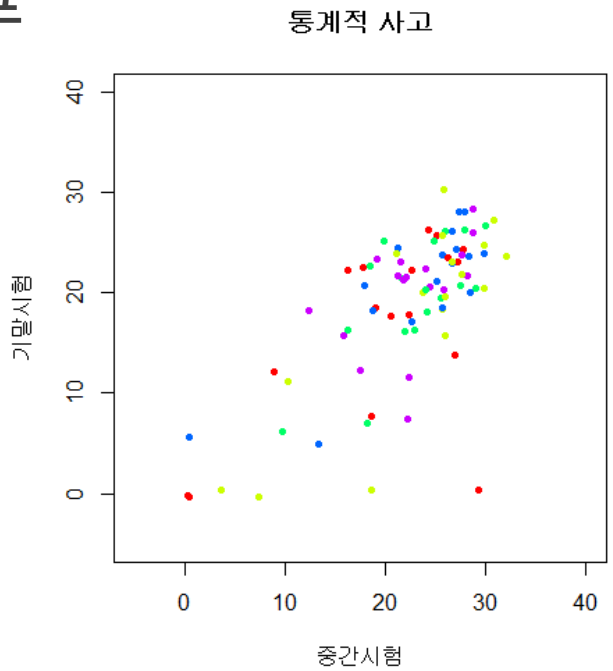
- R 스크립트

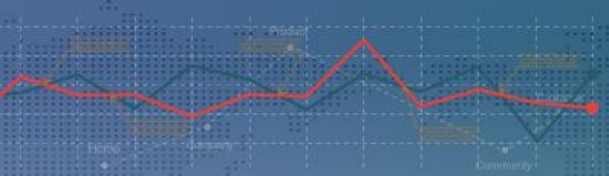
```
n <- length(mid)
plot(mid+runif(n,-0.5,0.5), final+runif(n,-0.5,0.5),
      pch=20, col=rainbow(5),
      xlim=c(-5,40) ,ylim=c(-5,40),
      xlab="중간시험", ylab="기말시험",
      main="통계적 사고")
```

1 산점도

보기 1. 중간·기말 성적 자료

출력 그래프





2. 이변량 밀도

- 밀도추정
- 보기 1. 중간·기말 성적 자료
- R 스크립트

2 이변량 밀도

밀도추정(density estimation)

- $(x_1, y_1), \cdots, (x_n, y_n)$ 로부터 모분포 확률밀도 $f(x, y)$를 추정:

$$\hat{f}(x, y;\ b_x, b_y) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{b_x}\frac{1}{b_y}K\left(\frac{x - x_i}{b_x}\right)K\left(\frac{y - y_i}{b_y}\right)$$

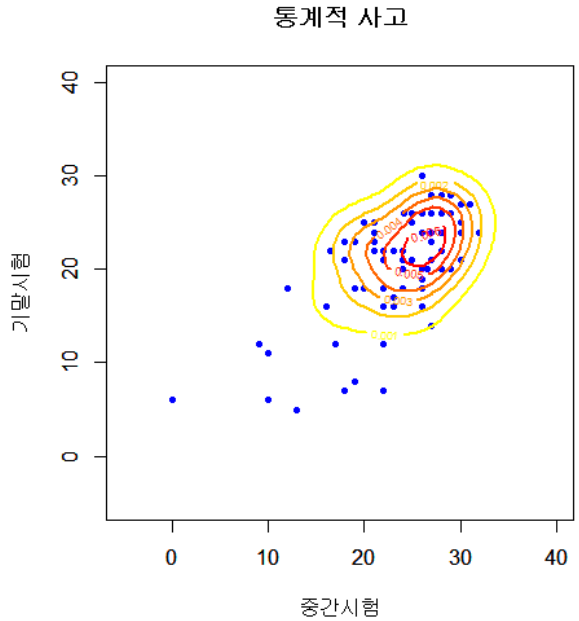
  - $K(z)$는 커널함수, b_x와 b_y는 띠 너비(bandwidth).

$$K(z) = \frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}$$

2 이변량 밀도

밀도추정(density estimation)

보기 1. 중간·기말 성적 자료

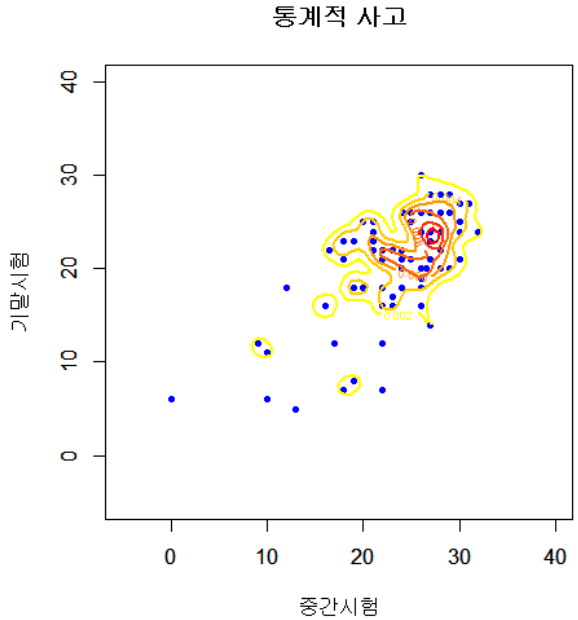


← $b_x, b_y = 2.5$ 로 지정

2 이변량 밀도

밀도추정(density estimation)

보기 1. 중간·기말 성적 자료



← $b_x, b_y = 1$로 지정

밀도추정(density estimation)

보기 1의 R 스크립트

```
exam <- exam[!is.na(exam$mid) & !is.na(exam$final),]
windows(height=5.5,width=5)
plot(exam$mid,exam$final,pch=20,... )
library(KernSmooth)
density <- bkde2D(exam,bandwidth=c(2.5,2.5))
par(new=T)
contour(density$x1,density$x2,density$fhat,
     xlim=c(-5,40),ylim=c(-5,40),
     col=heat.colors(7)[7:1],
     nlevels=7,lwd=2)
```

2 이변량 밀도

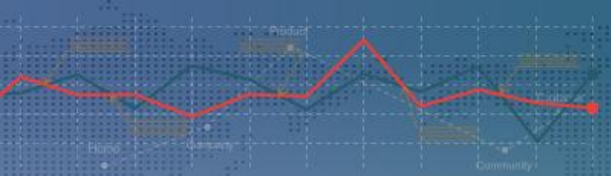
R 스크립트 상세 설명

```
density <- bkde2D(exam, bandwidth=c(2.5,2.5))
```

bivariate kernel density estimate

```
contour(density$x1, density$x2, density$fhat,
      xlim=c(-5,40),ylim=c(-5,40),
      col=heat.colors(7)[7:1],
      nlevels=7, lwd=2)
```

line width



3. ‘큰’ 자료의 산점도

- 자료 크기 n ≥ 10,000
- 육각형 칸에 넣기 (hexagonal binning)

3 '큰' 자료의 산점도

자료 크기 $n \geq 10,000$

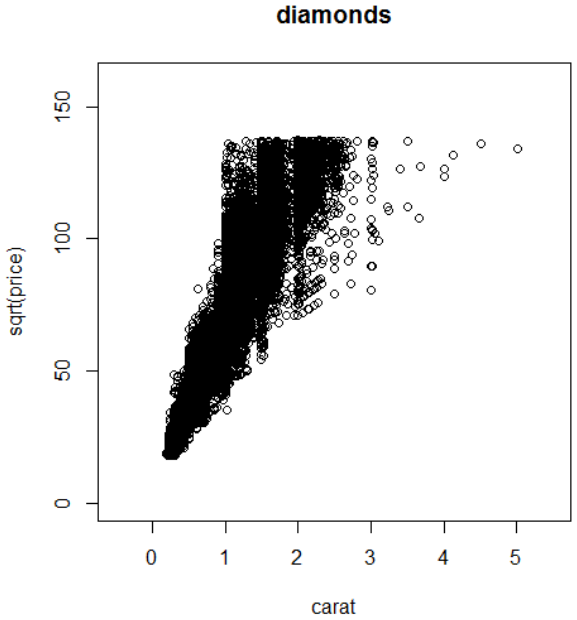
- 그래프 영역이 자료 점들로 포화 → 먹칠
- 보기 1. diamond 자료

```
library(ggplot2)
data(diamonds)
str(diamonds); attach(diamonds)
windows(height=6, width=6)
plot(carat,sqrt(price),main="diamonds",
     xlim=c(-0.5,5.5),ylim=c(0,160))
```

3 ‘큰’ 자료의 산점도

자료 크기 $n \geq 10,000$

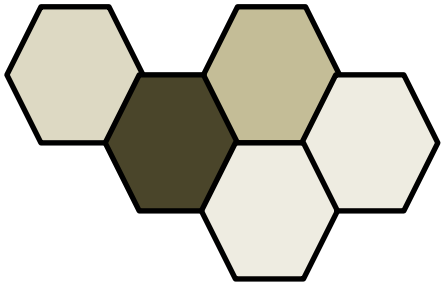
- 보기 1. `diamond` 자료 ($n = 53{,}940$): 출력 그래프



3 ‘큰’ 자료의 산점도

육각형 칸에 넣기 (hexagonal binning)

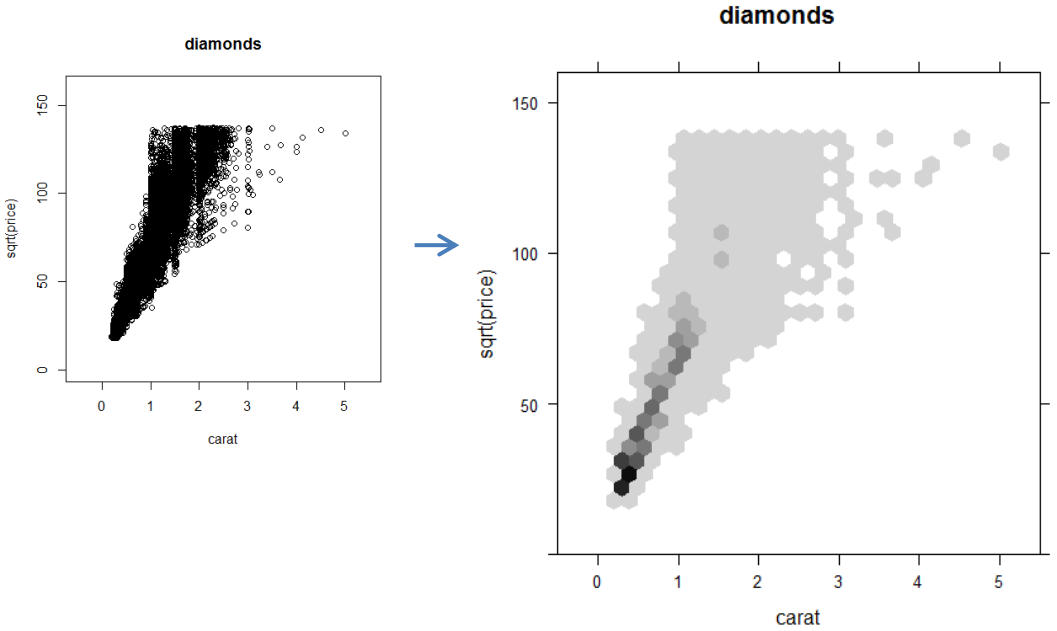
- 그래프 영역을 육각형의 벌집으로 나누어 개체들을 해당하는 칸에 넣고 칸의 돗수를 색의 불투명도로 표현



3 ‘큰’ 자료의 산점도

육각형 칸에 넣기 (hexagonal binning)

- 보기 1. `diamond` 자료에 적용된 “육각형 칸에 넣기”

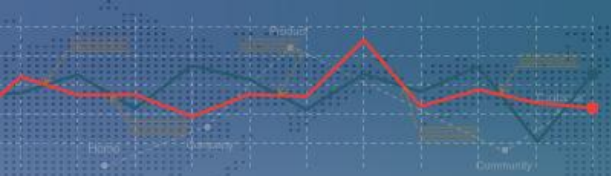


3 '큰' 자료의 산점도

육각형 칸에 넣기 (hexagonal binning)

- R 스크립트 (보기 1)

```
library(hexbin)
windows(height=7, width=6.4)
hexbinplot(sqrt(price)~carat, data=diamonds,
     main="diamonds",
     xlim=c(-0.5,5.5), ylim=c(0,160),
     xbins=25, aspect=1, colorkey=F)
```



4. 회귀적 관계

- 회귀함수의 두 종류
- 선형회귀
- 비모수적 회귀
- R 스크립트

4 회귀적 관계

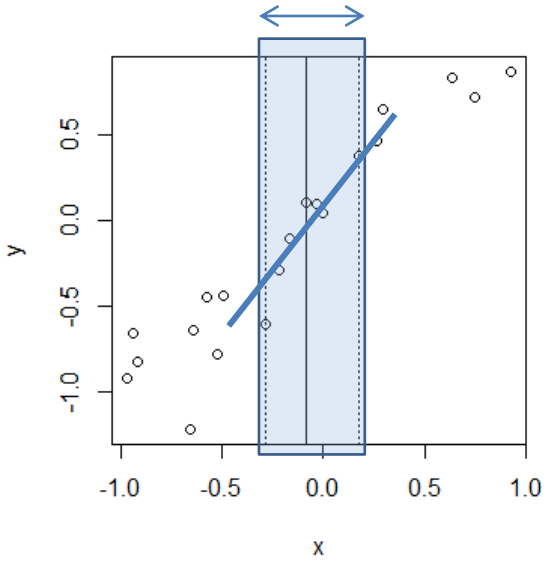
회귀함수의 두 종류

- 회귀함수 : $f(x) = E[Y \mid X = x]$
- 회귀모형 : $Y = f(x) + \varepsilon, \quad \varepsilon \sim (0, \sigma)$
  - 선형회귀 : $f(x) = \beta_0 + \beta_1 x$, 최소제곱법
  - 비모수적 회귀 : lowess
    (locally weighted scatterplot smoothing)

* 허명회 (2010) 『R을 활용한 탐색적 자료분석』, 188–190쪽

회귀함수의 두 종류

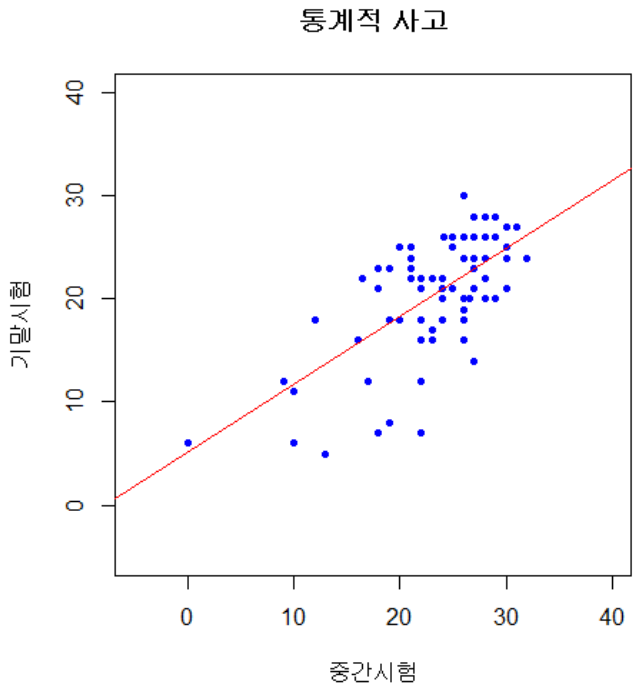
- lowess (locally weighted scatterplot smoothing)



- 띠 너비 (bandwidth) b_x
- 띠에 포함되는 데이터의 상대적 비율 = f

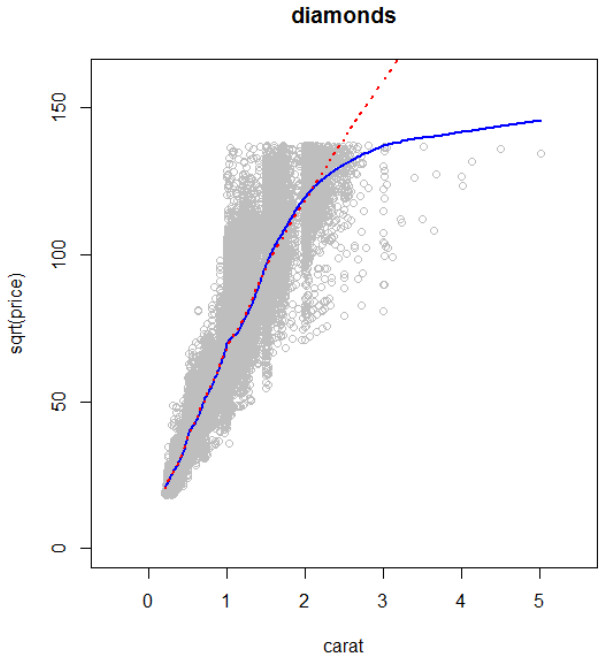
4 회귀적 관계

선형회귀 : $f(x) = \beta_0 + \beta_1 x$



4 회귀적 관계

비모수적 회귀 : 띠너비 f 에 따라 유연하게 데이터 패턴을 반영



실선: $f = 0.1$

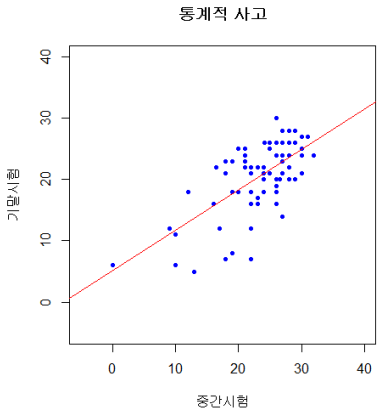
점선: $f = 0.25$

4 회귀적 관계

R 스크립트

선형회귀

```
windows(height=5.5,width=5)
plot(exam$mid,exam$final,pch=20,...)
abline(lm(exam$final~exam$mid),col="red")
```

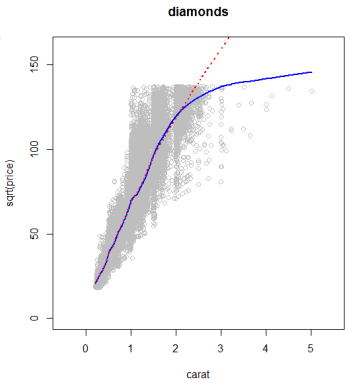


4 회귀적 관계

R 스크립트

비모수적 회귀

```
windows(height=7,width=6.4)
diamonds$sqrt.price <- sqrt(price)
plot(sqrt(price) ~ carat,col="gray", …)
lines(
  lowess(diamonds$sqrt.price~diamonds$carat,f=0.1),
    lwd=2,col="blue")
```



정리

- 산점도 : 두 연속형 변수 간 관계를 보여준다.
- 이변량 밀도 : 밀도 등고선(contour line) 넣기
- '큰' 자료의 산점도 : 육각형 칸에 넣기(hexagonal binning)
- 회귀적 관계 : 선형회귀 vs 비모수적 회귀 lowess

- 연속형 변수가 아니라 범주형 변수라면? → 다음 강의