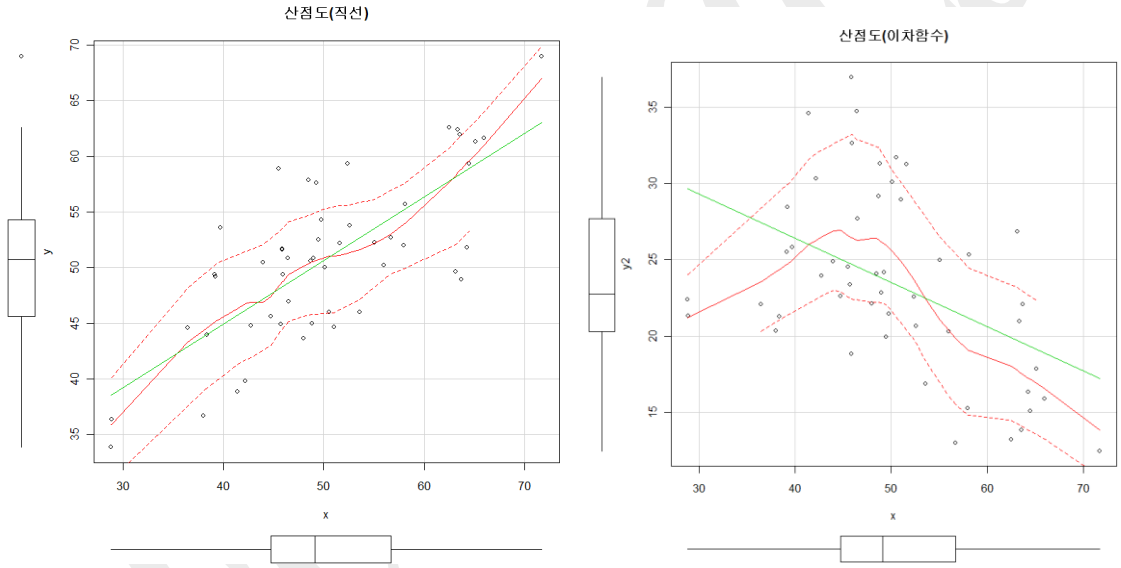
산점도

o 측정형변수 2 개 (x-축과 y-축)를 이용하여 2 차원 공간에 개체 표현
o 변수간의 관계나 개체의 특성 파악에 사용

변수간 함수관계

두 측정형변수의 함수관계를 표현

```
x=rnorm(50,50,10)
y=10+0.6*x+rnorm(50,10,5)
y2=10+0.6*x-0.01*x^2+rnorm(50,10,5)
ds=data.frame(cbind(x,y,y2))
scatterplot(y~x, main="산점도(직선)",ds)
scatterplot(y2~x, main="산점도(이차함수)",ds)
```



o 그린색 직선 : 최소자승추정법(OLS)에 의해 $Y = a + bX + e$의 모수 (a, b)를 추정하고 표현한 것임
o 붉은 색 : LOWESS LOcal WEight Scatterplot Smoothing, 양쪽은 95% 신뢰구간

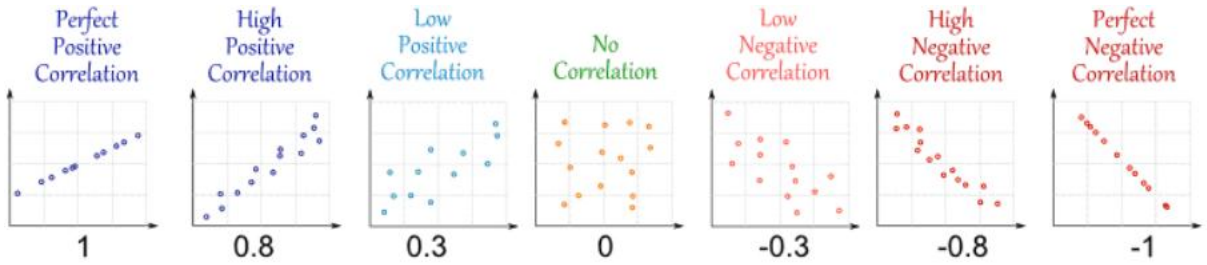
(상관관계) 두 변수간의 직선관계

두 변수의 상관관계 (여기서는 직선관계) 정도를 표현

$$r = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)}\sqrt{\mathrm{var}(Y)}} = \frac{E(X-E(X))E(Y-E(Y))}{\sqrt{\mathrm{var}(X)}\sqrt{\mathrm{var}(Y)}} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}}$$

귀무가설 $H_0 : \rho = 0$ (상관관계 존재 않음)=> $T = \sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}} \sim t(n-2)$

Professor Kwon, Sehyug | Dept. of Statistics, HANNAM Univ. 010.6365.7622 http://wolfpack.hnu.ac.kr



선형 (직선) 관계 $Y_i = a + bX_i + e_i$

최소자승추정법 OLS Ordinary Least Square

$$\min_{a,b} \sum e_i^2 = \sum (Y_i - a - bX_i)^2$$

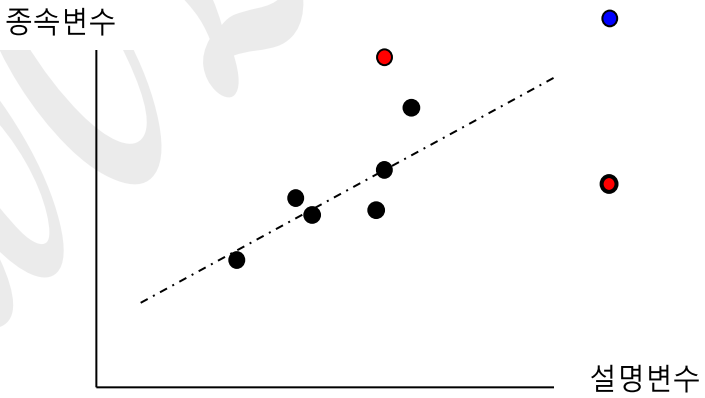
(모형 적합성) $H_0 : b = 0$ > 검정통계량 $TS = \frac{\hat{b}}{s(\hat{b})} \sim t(n-1)$

(결정계수 determinant coefficient) $R^2 = \frac{SSR}{SST} = \frac{\text{모형변동}}{\text{총변동}}$ =(상관계수 제곱) [설명변수 1 개]

결정계수가 70% 이상이면 종속변수 Y 를 유의적으로 설명하는 모형 설정하였음.

(이상치)는 종속변수와 설명변수의 선형 관계식에서 멀리 떨어진 관측치 (Y-축 기준),
삭제함을 원칙으로 함

(영향치)는 선형모형에 영향을 주는 관측치, 다른 관측치와 설명변수 범위 면에서 (X-축 기준)
떨어진 관측치 : 결정계수를 높이는 역할, 실제 사이 설명변수 구간의 관측치를 수집 필요



`summary(lm(y~x,ds))` 선형회귀모형을 OLS 방법에 의해 추정하고 유의성 검정

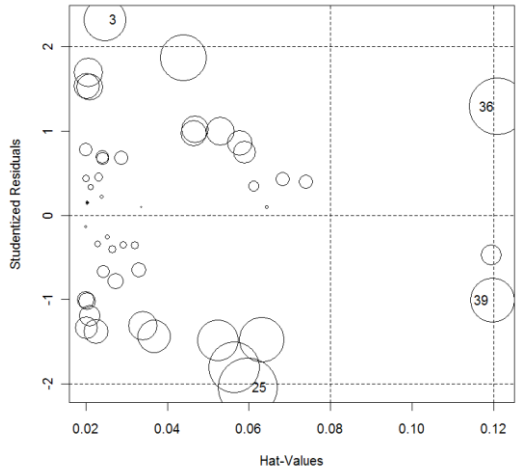
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 22.07979    3.74340   5.898 3.58e-07 ***
x            0.57095    0.07323   7.797 4.48e-10 ***
```

$Y_i = 22.1 + 0.57X_i$ => 통계적으로 유의

```
Multiple R-squared: 0.5588, not enough
```

(이상치, 영향치 진단)

```
influencePlot(lm(y~x,ds),id.n=2)
```




```
ds0=rbind(ds[1:2,],ds[4:50,])
summary(lm(y~x,ds0))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 21.28615    3.60031   5.912 3.65e-07 ***
x            0.58233    0.07028   8.285 9.59e-11 ***
Multiple R-squared: 0.5936,
```

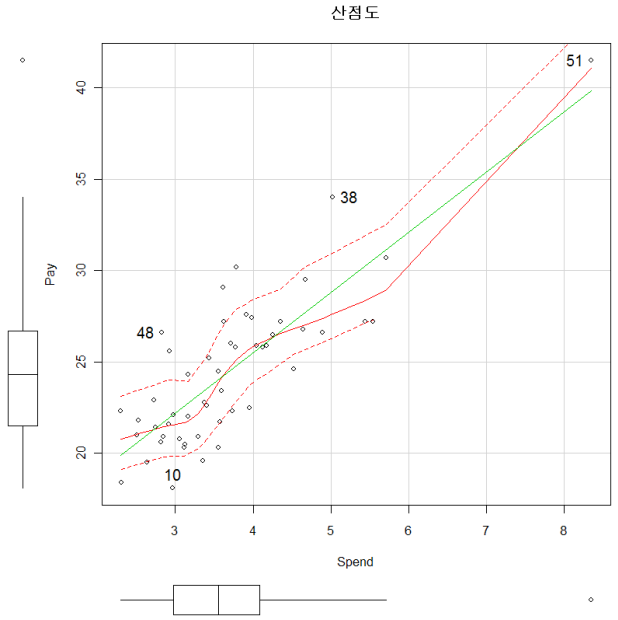
개체 진단

산점도에 개체를 식별할 수 있도록 표현하여 개체나 개체의 그룹에 대한 정보를 얻음

SPENDING.csv

미국 50 개 주 (state, 지역 region)의 교사봉급 (salary)와 학생 일인당 지출예산 (spent)

```
attach(ds.spend)
scatterplot(Pay~Spend, main="산점도",id.method="identify")
```




```
> ds.spend[51,]
    State Region  Pay Spend
51 Alaska     PA 41.5  8.35
```

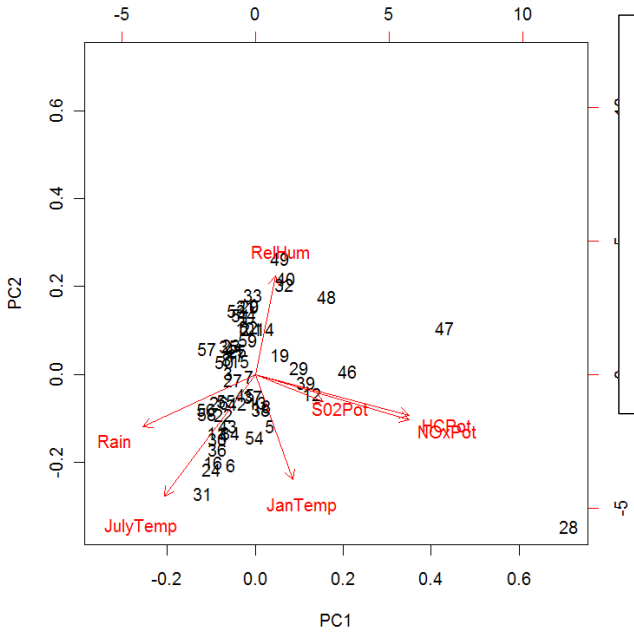
Alaska 주는 지출과 교사 연봉 모두 높음

```
> ds.spend[38,]
   State Region Pay Spend
38  D.C.     SA  34  5.02
```

워싱턴은 지출이 유사한 지역에 비해 교사 봉급 높음

변수가 많을 때 (주성분분석)

```
ds.cl=read.csv("climate.csv")
biplot(prcomp(ds.cl[2:8], scale=TRUE))
```



28 번: 1 월, 7 월 기온 높고 습도 낮으며, 공해 많음, 비는 적게 옴

```
> ds.cl[28,]
                          city
28 Los Angeles, Long Beach, CA
```

습도 높고 온도 낮음

```
> ds.cl[49,]
          city
49 Seattle, WA
```