1

상관분석

1. 개념

두 양적(순서형 포함) 변수(X, Y)간의 직선 관계 정도를 계수로 측정함

직선관계가 유의하다는(한 변수가 증가하면 다른 변수도 직선적으로 증가하거나 감소함) 것은 두 변수가 유사하다는 의미 - 변수의 유사성 척도

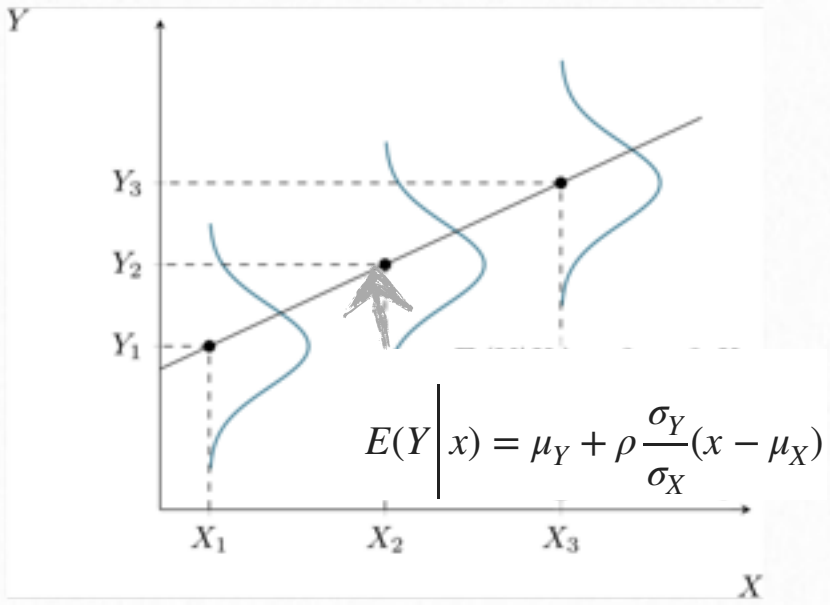
1) 수리적 접근 ($Y = a + bX + e \sim N(0, \sigma^2)$)

$X = x$에 대하여 Y는 정규분포를 따른다고 가정하자

(1) $Y \mid x \sim Normal$

(2) $E(Y \mid x) = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X}(x - \mu_X)$

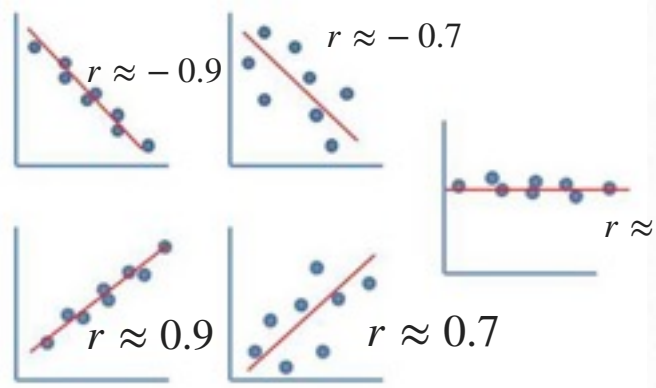
(3) $V(Y \mid x) = \sigma_Y^2$

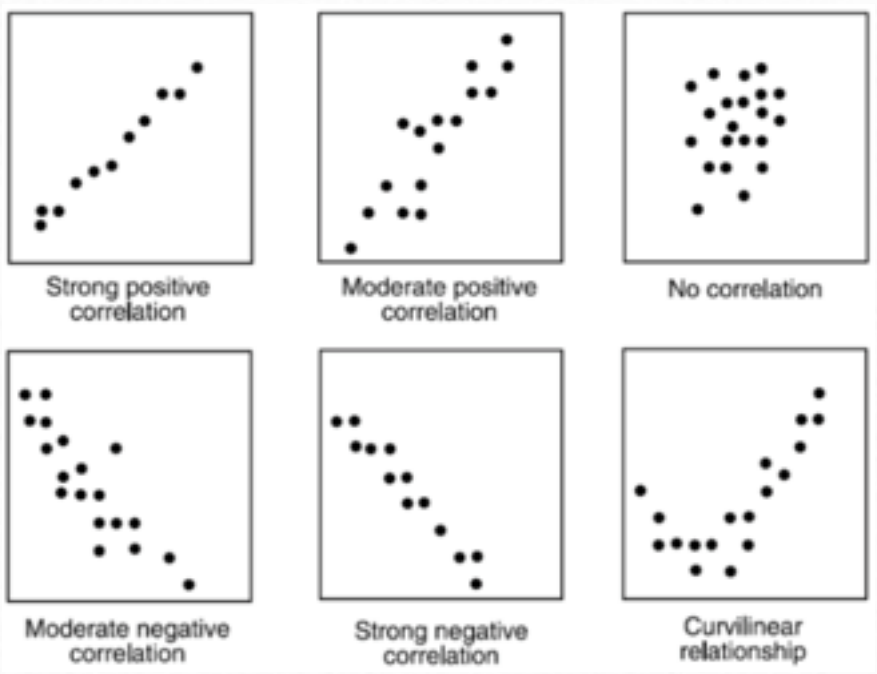


2) 산점도 scatter plot

두 양적변수의 함수관계를 이차원 공간의 점들로 표현하는 그래프

X-축은 독립변수(설명변수)에 해당하는 확률변수, Y-축은 종속변수에 해당하는 변수의 관측값을 좌표로 활용한다.





*) 출처 -위키피디아

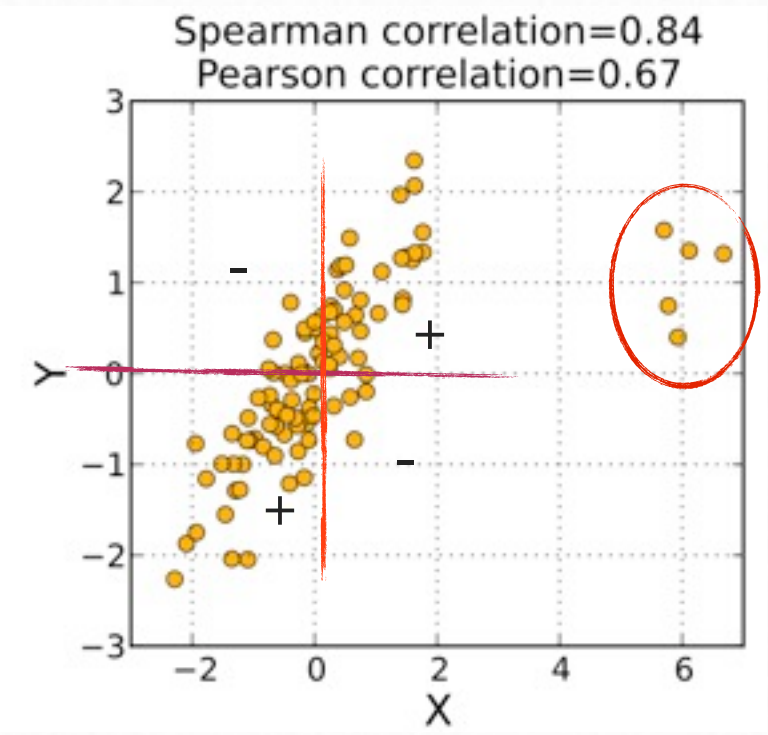
2. 상관계수 correlation coefficient

1) Karl Peason 공식

$\rho = \frac{COV(X,Y)}{\sqrt{V(X)}\sqrt{V(Y)}}$: 모집단

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

분모는 확률변수의 표준편차이므로 상관계수의 부호를 결정하는 분자항이다. $(x_i - \bar{x})(y_i - \bar{y})$의 부호는 아래 그림(수평선은 Y의 평균, 수직선은 X의 평균, 오른쪽의 관측치 5개를 제외한 경우)에서 시각적으로 확인할 수 있음.



*) 출처 : 위키피디아

대부분의 데이터 범위 밖에 있는 관측치(타원형 내 관측치)는 상관계수 값을 높이는 역할을 한다. 그러므로 상관계수를 계산하기 전에 반드시 산점도를 그려 데이터의 범위를 많이 벗어난 관측치가 있는지 확인하여 상관분석의 활용도를 높일 필요가 있음.

2) Spearman 순위 상관계수

(방법 1) $r_s = Corr(R_{X_i}, R_{Y_i})$ where R_{X_i}는 X_i의 순위이며, R_{Y_i}는 Y_i의 순위이다.

(방법 2) $r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$, $d_i = R_{X_i} - R_{Y_i}$

3) Kendall τ 상관계수

$\tau = \frac{\#of\_concordant\_pairs - \#of\_disconcordant\_pairs}{n(n-1)/2}$

concordant = 만약 $(x_i > x_j)$, $(y_i > y_j)$이거나 $(x_i < x_j)$이면, $(y_i < y_j)$이면 두 관측치는 concordant 쌍이라 함

τ 값이 클수록 데이터 순위의 일치도는 높아지므로 상관관계가 높다.

3. 상관계수 유의성 검정

1) 가설

귀무가설 : 두 변수의 **직선** 상관관계는 유의하지 않다. <=> 서로 독립이다. $\rho = 0$

대립가설 : 두 변수의 **직선** 상관관계는 유의하다. $\rho \neq 0$

2) 데이터 검증

1) 데이터는 이변량 정규분포에 근사해야 한다. 단 n>20 인 대표본에서는 문제 없음

2) 산점도를 그려 데이터 범위(X-) 밖의 관측치 존재 여부를 체크한다. - 존재한다면 제외하거나 활용 시 주의해야 한다.

3) 검정통계량

$TS = \frac{r}{\sqrt{(1-r^2)(n-2)}} \sim t(n-2)$, n=표본크기, r=상관계수

(만약) 귀무가설이 $\rho = \rho_0 \neq 0$ (임의의 상관계수와 동일한 경우)이라면

$$TS = \frac{1}{2}\ln(\frac{1+r}{1-r}) \sim Normal(\frac{1}{2}\ln(\frac{1+\rho}{1-\rho}), \frac{1}{n-3})$$

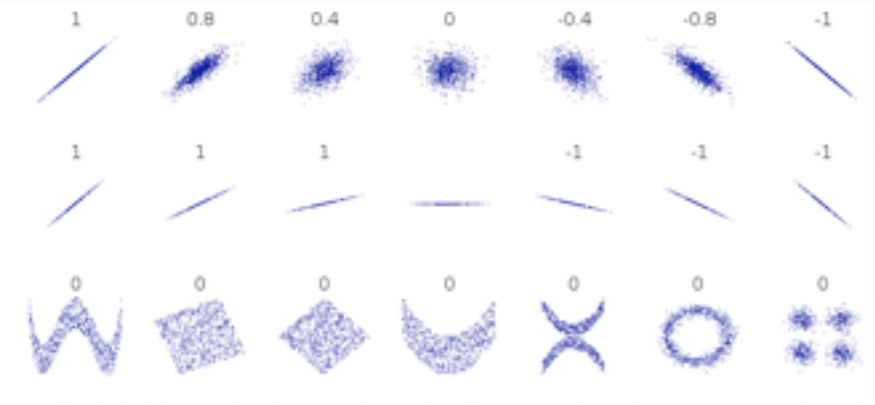
4) 결론

유의확률 $P(t(n-2) > \left|TS\right|)$이 유의수준보다 작다면 귀무가설을 기각하여 상관관계의 유의하다고 결론 내리고 표본상관계수의 부호를 이용하여 해석

- 귀무가설이 기각, 표본상관계수 부호 + => 두 변수는 양의 상관관계가 있고 한 변수의 값이 증가(감소)하면 다른 변수의 값도 증가(감소)한다.
- 귀무가설이 기각, 표본상관계수 부호 - => 두 변수는 음의 상관관계가 있고 한 변수의 값이 증가(감소)하면 다른 변수의 값도 감소(증가)한다.

4. 상관계수 해석

상관 계수는 두 변수간의 선형 관계 (linear association)에 대한 척도



*) 출처 : 위키피디아

(1) -1과 1사이의 값이다.

(2) 1에 가까우면 양의 선형 상관 관계가 존재한다. 한 변수의 값이 증가(감소)하면 다른 변수 값도 증가(감소)한다.

(3) -1에 가까우면 음의 선형 상관 관계가 존재한다. 한 변수의 값이 증가(감소)하면 다른 변수 값은 감소(증가)한다.

(4) 두 변수의 상관 관계가 높다는 것은 두 변수가 동일한(comparable) 개념을 측정한다는 의미도 담고 있다(두 변수가 유사함). 그러므로 변수를 축약하거나 개체를 분류하는데 사용되는 다변량 분석에서는 공분산, 혹은 상관계수 개념을 사용

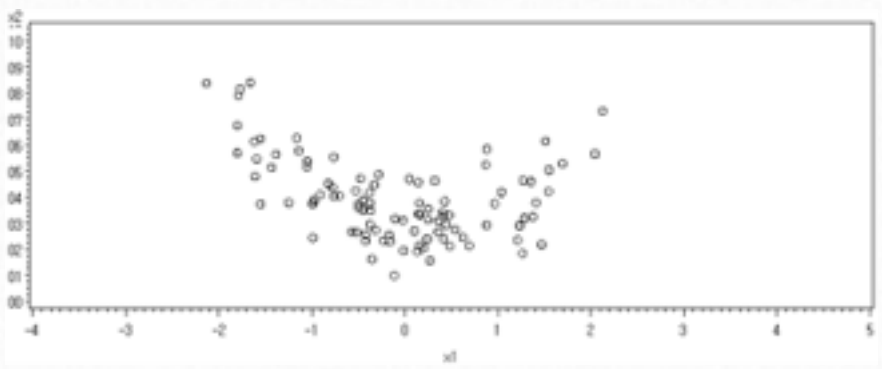
(5) 회귀분석과 관계 : $Y = a + bX + e$,

(a) 독립변수 X가 Y에 선형적 영향을 미치는지 검정 <=> 기울기 b=0(영향을 미치는 않음) 유의성 검정 <=> 상관계수의 유의성 검정과 동일

(b) 그러므로 회귀계수 b의 부호와 상관계수 r의 부호는 동일하고 $\hat{b} = \sqrt{\frac{S_{XY}}{S_{XX}}} r$ ($S_{XX} = \sum (x_i - \bar{x})^2$, $S_{XY} = \sum (x_i - \bar{x})(y_i - \bar{y})$)

(c) 모형의 적합성을 나타내는 결정계수는 상관계수의 제곱과 같다. $R^2 = \frac{SSR}{SST} = r^2$

(4) 상관 계수가 0에 가깝다는 것은 선형 상관 관계가 없다는 것이지 함수 관계가 없다는 것은 아니다. 두 변수는 이차식에 의한 ($Y_i = 100 + X_i^2 - 0.4X_i$) 생성된 데이터이나 상관계수는 0에 가깝다.



(5) 상관 계수의 크기는 자료의 크기가 커지면 증가하므로, 값 자체의 크기가 의미가 있는 곳이 아니라

자료로부터 검정 시 계산되는 유의확률(p-값)에 의해 상관관계를 분석하면 된다.

(6) 다음은 유의성을 검정하지 않고도 유의한 정도를 알아볼 수 있는 기준이다.

(a) 실험실 자료와 같이 연구자가 자료 수집을 control 할 수 있는 경우는 0.9 이다.
(b)연구자가 control하기 어려운 경우는 0.7 정도이다.
(c) 일반적으로 자료의 수가 20-30정도인 경우 0.6 정도를 생각한다.
(d) 설문 조사의 리커드 척도와 같이 변수가 가질 수 있는 값이 한정된 경우 (1-5점, 물론 여러 문항을 합쳐 평균을 이용하는 경우에는 다소 문제가 해결되지만) 상관 계수는 매우 낮다.

5. 사례분석

```
DS<-read.csv("SMSA_USA.csv")
names(DS)
library(car)
scatterplot.matrix(~X1월기온+X7월기온+상대습도+강수량, data=DS,
  diagonal=c("density"), main="Scatter Plot", spread=F)

cor(DS[,2:5], method="pearson") #spearman, kendall
cor.test(DS[,2:2],DS[,3:3],method="pearson")

library(Hmisc)
rcorr(as.matrix(DS[,2:5])) #Pearson only 상관계수 행렬
```

1) 산점도 행렬 해석

히스토그램

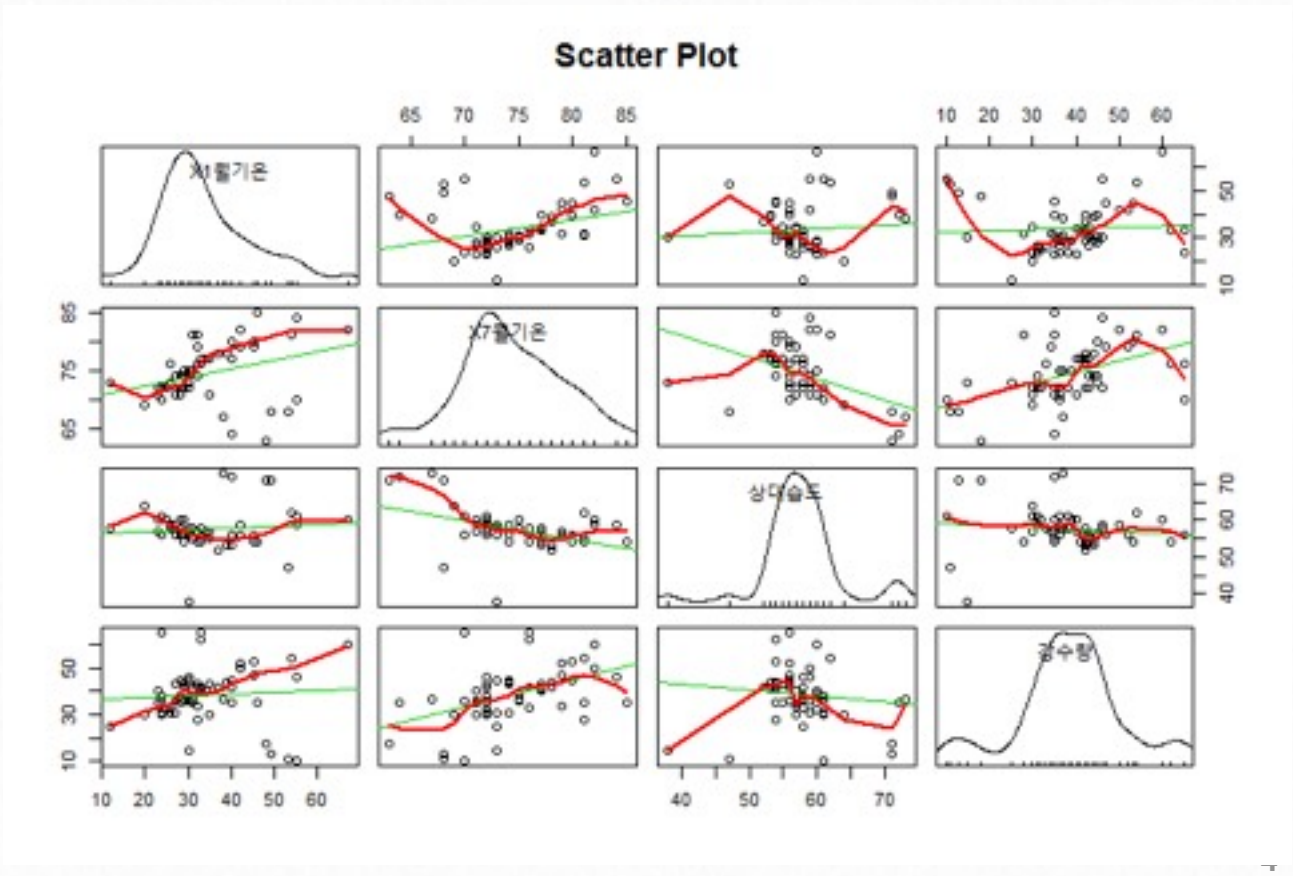
- 1월 기온 우로 치우친 경향
- 7월 기온 : 좌우대칭
- 상대습도 : 좌우 대칭, 좌측 이상치 보임
- 강수량 : 좌우 대칭, 양측 낮은 봉우리

산점도

- 우측 삼각형과 좌측 삼각형 동일 산점도
- 초록색 : 적합 선형 회귀선
- 붉은색 : 함수 형태
- 상대습도와 강수량 선형 관계 가장 높음

```
> names(DS)
 [1] "도시이름"          "X1월기온"          "X7월기온"
 [4] "상대습도"          "강수량"            "사망률지수"
 [7] "교육기간"          "인구밀도.천명."    "비백인비율..."
[10] "사무직종사자비율" "총인구"            "가구당인구"
[13] "가계소득"          "HCPot"             "NOxPot"
[16] "SO2Pot"            "NOx"               "남부여부"
```

도시이름=DS[, 1:1]
상대습도~사망률지수=DS[,4:6]



2) 상관계수 행렬

cor() 함수 - 상관계수 값만 출력

```
> cor(DS[,2:5], method="pearson") #spearman, kendall
           X1월기온   X7월기온     상대습도      강수량
X1월기온 1.00000000  0.3221455  0.08552171  0.05856608
X7월기온 0.32214555  1.0000000 -0.44139661  0.47225673
상대습도 0.08552171 -0.4413966  1.00000000 -0.11777277
강수량   0.05856608  0.4722567 -0.11777277  1.00000000
```

cor.test() 함수 - 두 변수 값만 가능, "1월기온"과 "7월기온" 상관계수는 0.32이고 유의하다(유의확률=0.013, 유의수준 5%). 그러므로 1월 기온이 높은 도시는 7월 기온도 높아지는 경향을 보인다.

```
> cor.test(DS[,2:2],DS[,3:3],method="pearson")

        Pearson's product-moment correlation

data:  DS[, 2:2] and DS[, 3:3]
t = 2.5691, df = 57, p-value = 0.01284
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.07200323 0.53416174
sample estimates:
      cor
0.3221455
```

rcorr() 함수 : 데이터는 행렬 형태이어야 함, 유의수준과 상관계수 값 동시 출력, (1월 기온, 7월 기온-유의확률 0.013) 양의 상관관계, (7월기온, 상대습도, 유의확률<0.001)(음의 상관관계-7월기온이 높은 도시는 상대습도가 낮은 경향), (7월기온, 강수량, 유의확률<0.001) 양의 상관관계

```
> rcorr(as.matrix(DS[,2:5])) #Pearson only 상관계수 행렬
         X1월기온 X7월기온 상대습도 강수량
X1월기온     1.00     0.32     0.09   0.06
X7월기온     0.32     1.00    -0.44   0.47
상대습도     0.09    -0.44     1.00  -0.12
강수량       0.06     0.47    -0.12   1.00

n= 59


P
         X1월기온 X7월기온 상대습도 강수량
X1월기온           0.0128   0.5196   0.6595
X7월기온 0.0128            0.0005   0.0002
상대습도 0.5196   0.0005            0.3743
강수량   0.6595   0.0002   0.3743
```

|Bubble plot| (optional)

1) 개념

- 확률변수가 3개, 이차원 산점도에 그리는 방법
- 버블 크기가 3번째 변수 관측값의 크기에 비례함

2) 그리기

```
DS1<-DS[which(DS$HCPot<200),] #HCPot <200 Obs. only
DS2<-DS1[1:20,]   #1~20 observation,
rcorr(as.matrix(DS2[,14:16])) #Pearson only

symbols(DS2$HCPot, DS2$NOxPot, circles=DS2$SO2Pot)
radius <- sqrt( DS2$SO2Pot/ pi )
symbols(DS2$HCPot, DS2$NOxPot, circles=radius)
symbols(DS2$HCPot, DS2$NOxPot, circles=radius, inches=0.35,
        fg="white", bg="green", xlab="HCPot", ylab="NOxPot")
text(DS2$HCPot, DS2$NOxPot, DS2$도시이름, cex=0.5)
```

- HCPot 관측치가 200 미만인 도시만 선택하고 1~20번째 관측치만 선택 - 그래프 해석의 용이하게 하기 위하여
- 피어슨 상관계수 계산 결과 3변수 매우 유의

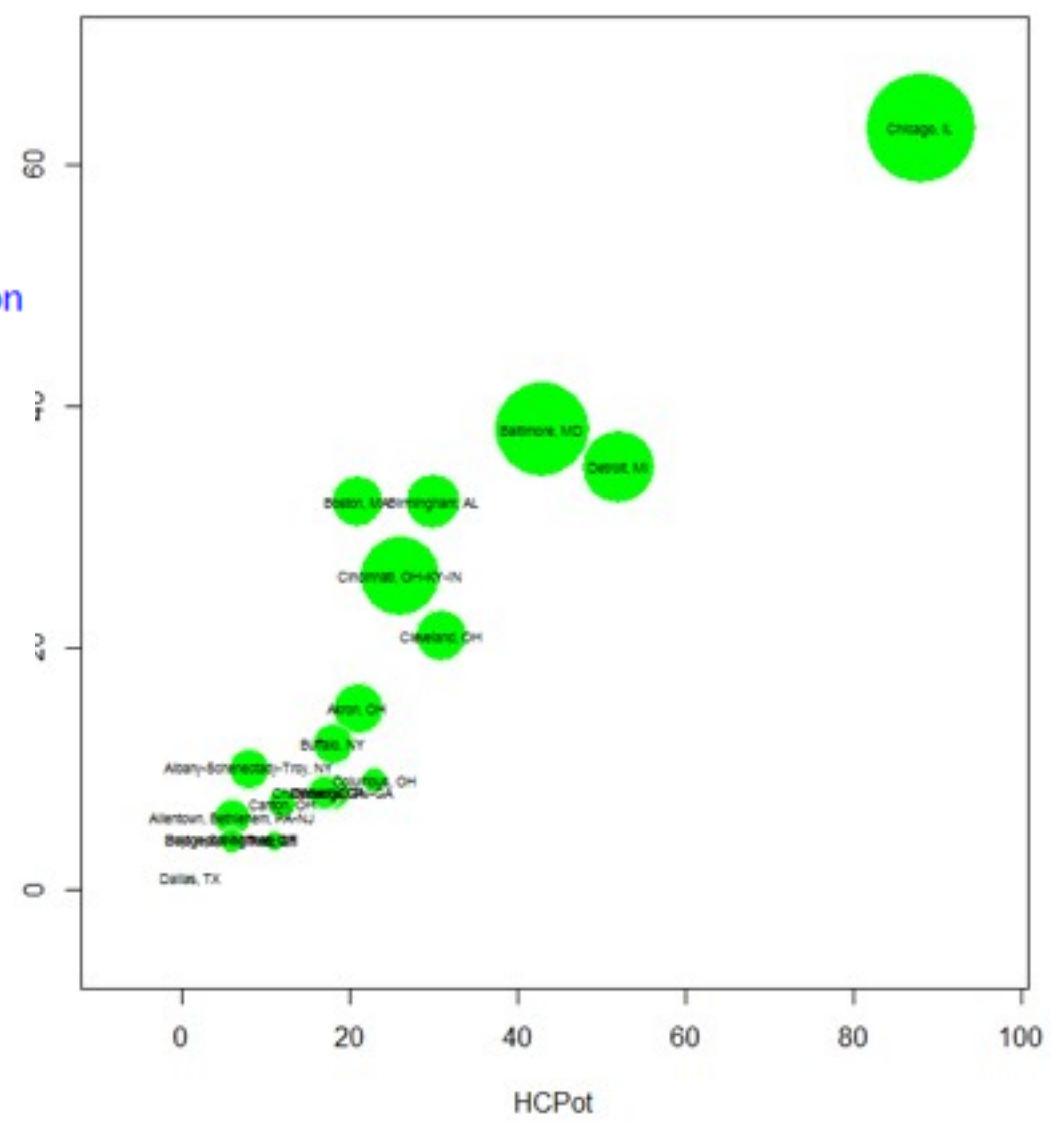
```
> rcorr(as.matrix(DS2[,14:16])) #Pearson
       HCPot NOxPot SO2Pot
HCPot   1.00   0.93   0.90
NOxPot  0.93   1.00   0.93
SO2Pot  0.90   0.93   1.00

n= 20


P
       HCPot NOxPot SO2Pot
HCPot         0      0
NOxPot 0             0
SO2Pot 0      0
```

[버블 산점도 해석]

- X-축=HCPot, Y-축=NOxPot, 버블 크기=SO2Pot
- 버블의 크기가 수평적으로 커지는 경향이 있음 - X축 HCPot 변수와 SO2Pot 변수는 양의 상관 관계
- 버블의 크기가 수직적으로 커지는 경향이 있음 - Y축 NOxPot변수와 SO2Pot 변수는 양의 상관 관계
- 만약 버블이 수직(수평)적으로 작아지는 경향이 있으면 음의 상관관계
- 일리노이 주 시카고의 3 변수의 관측값이 가장 크고 텍사스 주 달라스가 3변수 관측값이 가장 작다.



[Bubble plot] plot_ly 이용

library(plotly)

plot_ly(x=DS2$HCPot,y=DS2$NOxPot,size=DS$S02Pot,mode="markers")