Part 2. R 통계분석 (데이터 분석 전문가 양성과정)

12

상관관계와 상관분석

경북대학교 배준현 교수
(joonion@knu.ac.kr)

12. 상관관계와 상관분석

- **상관분석:** *correlation analysis*
  - 두 사건 간의 연관성을 분석하고자 할 때
    - 기업의 연구개발 투자와 신제품 출시 비율 간의 관계
    - 한 나라의 일인당 GDP와 국민의 기대수명 간의 관계
    - 어떤 제품의 광고비와 그 제품의 매출액 간의 관계
  - 상관: *correlation*
    - 두 사건 , 즉 두 변수 간의 선형적 관계
    - 이때 두 변수는 일반적으로 연속형 변수

12. 상관관계와 상관분석

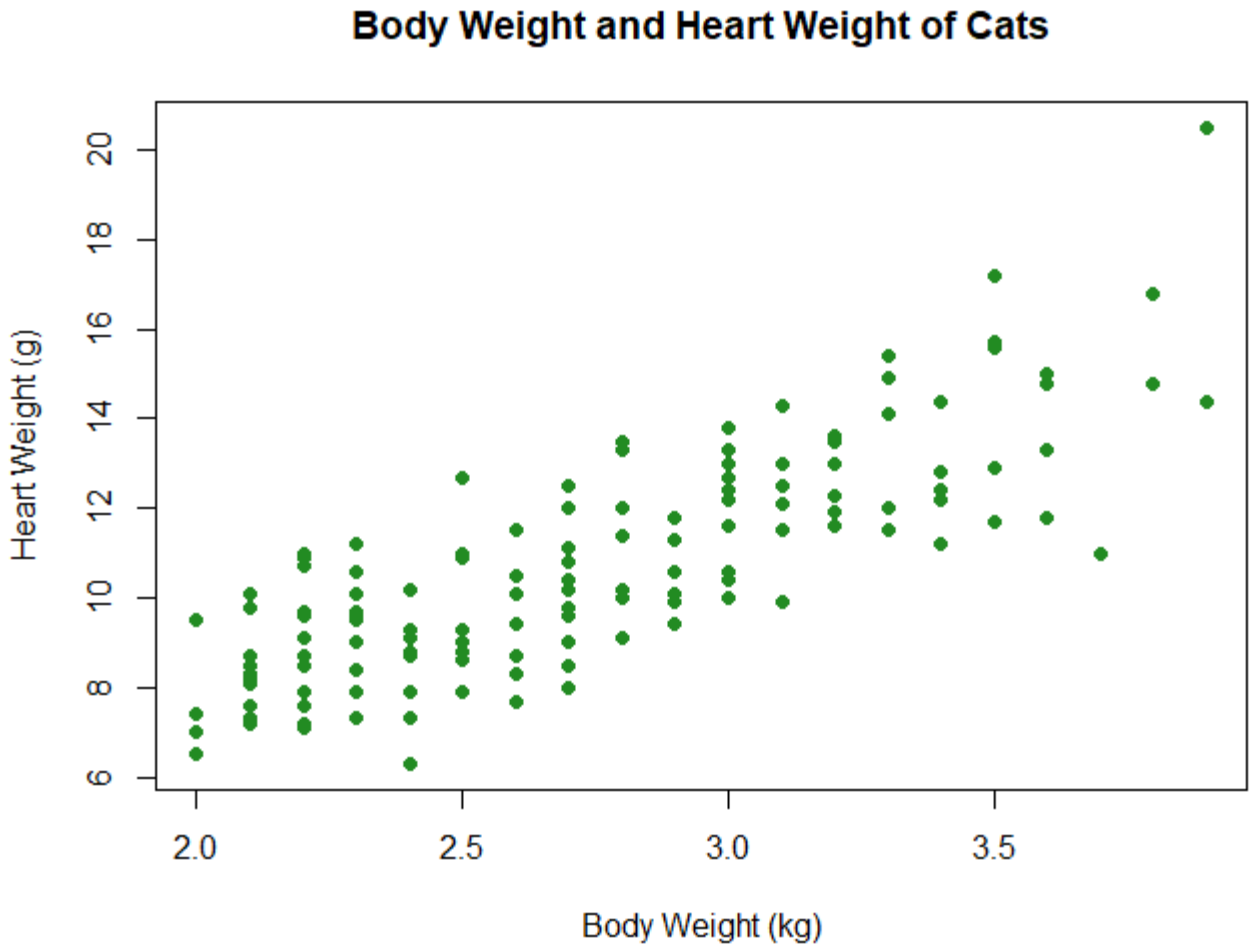
- MASS 패키지의 cats 데이터셋으로 상관분석 수행

```
> library(MASS)
> str(cats)
'data.frame':  144 obs. of  3 variables:
 $ Sex: Factor w/ 2 levels "F","M": 1 1 1 1 1 1 1 1 1 1 ...
 $ Bwt: num  2 2 2 2.1 2.1 2.1 2.1 2.1 2.1 2.1 ...
 $ Hwt: num  7 7.4 9.5 7.2 7.3 7.6 8.1 8.2 8.3 8.5 ...
```

12. 상관관계와 상관분석

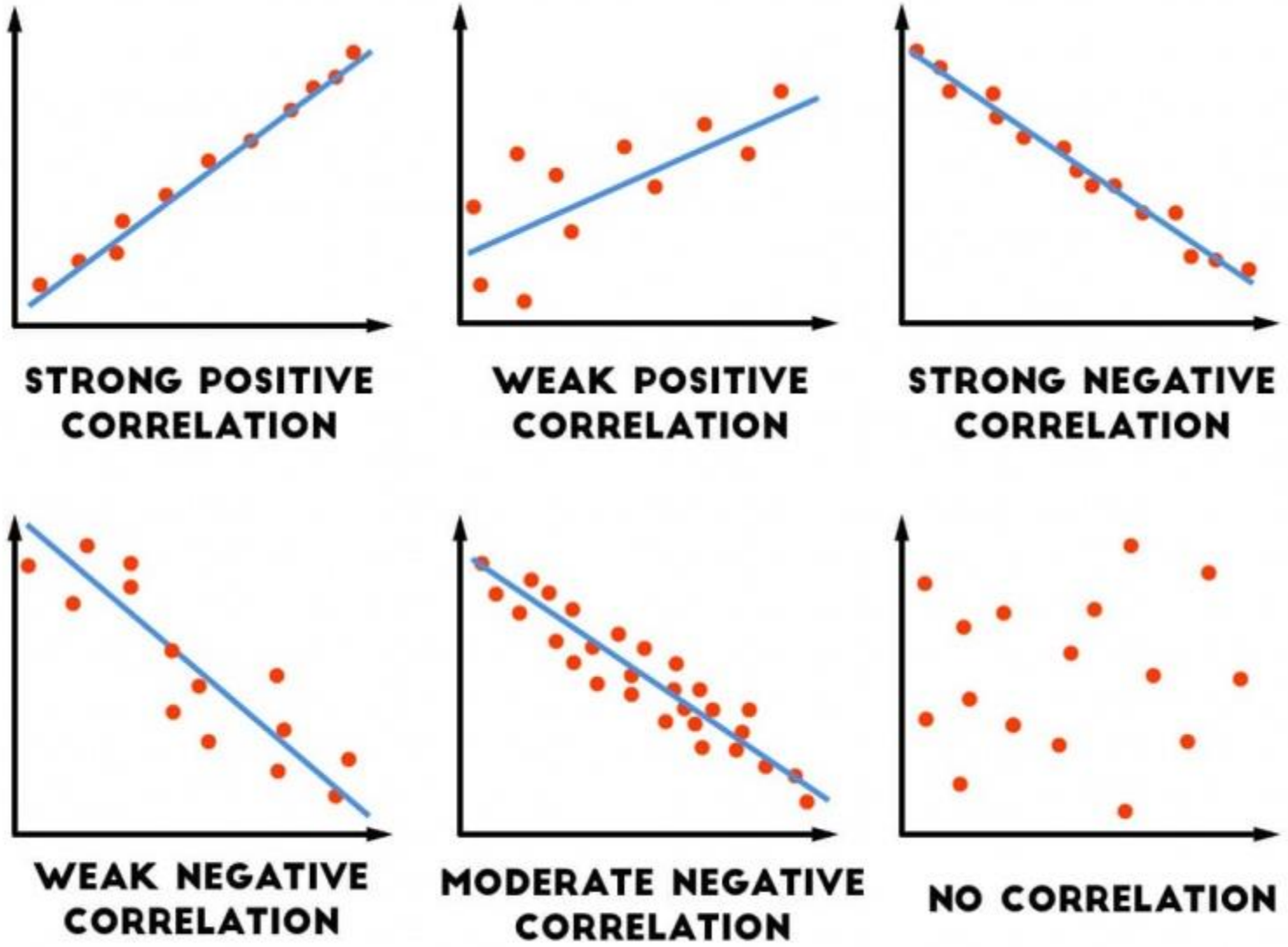
- 고양이의 몸무게와 심장무게 간의 대략적인 관계를 산점도(scatter plot)로 확인

```
plot(cats$Hwt ~ cats$Bwt,
     main="Body Weight and Heart Weight of Cats",
     col="forestgreen", pch=19,
     xlab="Body Weight (kg)",
     ylab="Heart Weight (g)")
```



- **상관계수:** *correlation coefficient*
  - 산점도로는 변수 간의 관계 패턴을 쉽게 이해할 수 있지만,
    - 선형관계의 강도를 객관적으로 파악할 수 없음
  - 상관계수:
    - 음의 상관관계: 상관계수가 –일 때
    - 양의 상관관계: 상관계수가 +일 때
    - -1에서 +1 사이의 값을 가지며
    - 0에 가까울수록 두 변수 간의 선형관계가 없음을 의미

12. 상관관계와 상관분석



상관계수의 종류

- 피어슨 상관계수: *Pearson's* correlation
  - 상관 분석에서 기본적으로 사용하는 상관계수
  - 정규성의 가정을 필요로 함
- 스피어만 상관계수: *Spearman's* correlation
  - 변수값 대신 순위로 바꿔서 사용하는 상관계수
  - 순위(rank) 데이터를 바탕으로 계산하므로 이상점에 덜 민감
- 켄달 상관계수: *Kendall's* correlation
  - 두 변수들 간의 순위를 비교하여 계산하는 상관계수
  - 샘플 사이즈가 작거나 데이터의 동률이 많을 때 유용함

12. 상관관계와 상관분석

- 고양이의 몸무게와 심장무게 간의 상관계수 계산

```
> cor(cats$Bwt, cats$Hwt)
[1] 0.8041274

> with(cats, cor(Bwt, Hwt))
[1] 0.8041274

> cor(cats$Bwt, cats$Hwt, method="pearson")
[1] 0.8041274

> cor(cats$Bwt, cats$Hwt, method="spearman")
[1] 0.7908427

> cor(cats$Bwt, cats$Hwt, method="kendall")
[1] 0.6079403
```

12. 상관관계와 상관분석

- 상관계수에 대한 유의성 검증: 모집단에서의 상관계수가 0이라는 귀무가설 검정

```
> with(cats, cor.test(Bwt, Hwt))

        Pearson's product-moment correlation

data:  Bwt and Hwt
t = 16.119, df = 142, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7375682 0.8552122
sample estimates:
      cor
0.8041274

> with(cats, cor.test(Bwt, Hwt, alternative="greater", conf.level=0.99))

> with(cats, cor.test(~ Bwt + Hwt))
```

12. 상관관계와 상관분석

- **포뮬러 형식을 이용: 암컷 고양이에 대해서만 상관계수의 유의성 검정**

```
> cor.test(~ Bwt + Hwt, data=cats)

> cor.test(~ Bwt + Hwt, data=cats, subset=(Sex=="F"))

        Pearson's product-moment correlation

data:  Bwt and Hwt
t = 4.2152, df = 45, p-value = 0.0001186
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2890452 0.7106399
sample estimates:
      cor
0.5320497
```

12. 상관관계와 상관분석

- **데이터셋에 세 개 이상의 벡터가 있을 경우: iris 데이터셋으로 상관계수 행렬 생성**

```
> str(iris)
> cor(iris[, -5])
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```

```
> iris.cor <- cor(iris[, -5])
> class(iris.cor)
[1] "matrix" "array“

> str(iris.cor)
num [1:4, 1:4] 1 -0.118 0.872 0.818 -0.118 ...
 - attr(*, "dimnames")=List of 2
  ..$ : chr [1:4] "Sepal.Length" "Sepal.Width" "Petal.Length" "Petal.Width"
  ..$ : chr [1:4] "Sepal.Length" "Sepal.Width" "Petal.Length" "Petal.Width“

> iris.cor["Petal.Width", "Petal.Length"]
[1] 0.9628654
```

- **세 개 이상의 변수 간의 상관계수 유의성 검정**: psych **패키지의** corr.test() **함수 이용**

```
> library(psych)
> corr.test(iris[, -5])
Call:corr.test(x = iris[, -5])
Correlation matrix
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length         1.00       -0.12         0.87        0.82
Sepal.Width         -0.12        1.00        -0.43       -0.37
Petal.Length         0.87       -0.43         1.00        0.96
Petal.Width          0.82       -0.37         0.96        1.00
Sample Size
[1] 150
Probability values (Entries above the diagonal are adjusted for multiple tests.)
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length         0.00        0.15            0           0
Sepal.Width          0.15        0.00            0           0
Petal.Length         0.00        0.00            0           0
Petal.Width          0.00        0.00            0           0

 To see confidence intervals of the correlations, print with the short=FALSE option
```

12. 상관관계와 상관분석

- **상관계수의 95% 신뢰구간 출력: print() 함수 이용**

```
> print(corr.test(iris[, -5]), short=FALSE)
Call:corr.test(x = iris[-5])

……(중략)

 Confidence intervals based upon normal theory.  To get bootstrapped values, try cor.ci
             raw.lower raw.r raw.upper raw.p lower.adj upper.adj
Spl.L-Spl.W     -0.27 -0.12      0.04  0.15     -0.27      0.04
Spl.L-Ptl.L      0.83  0.87      0.91  0.00      0.81      0.91
Spl.L-Ptl.W      0.76  0.82      0.86  0.00      0.74      0.88
Spl.W-Ptl.L     -0.55 -0.43     -0.29  0.00     -0.58     -0.25
Spl.W-Ptl.W     -0.50 -0.37     -0.22  0.00     -0.51     -0.20
Ptl.L-Ptl.W      0.95  0.96      0.97  0.00      0.94      0.98
```

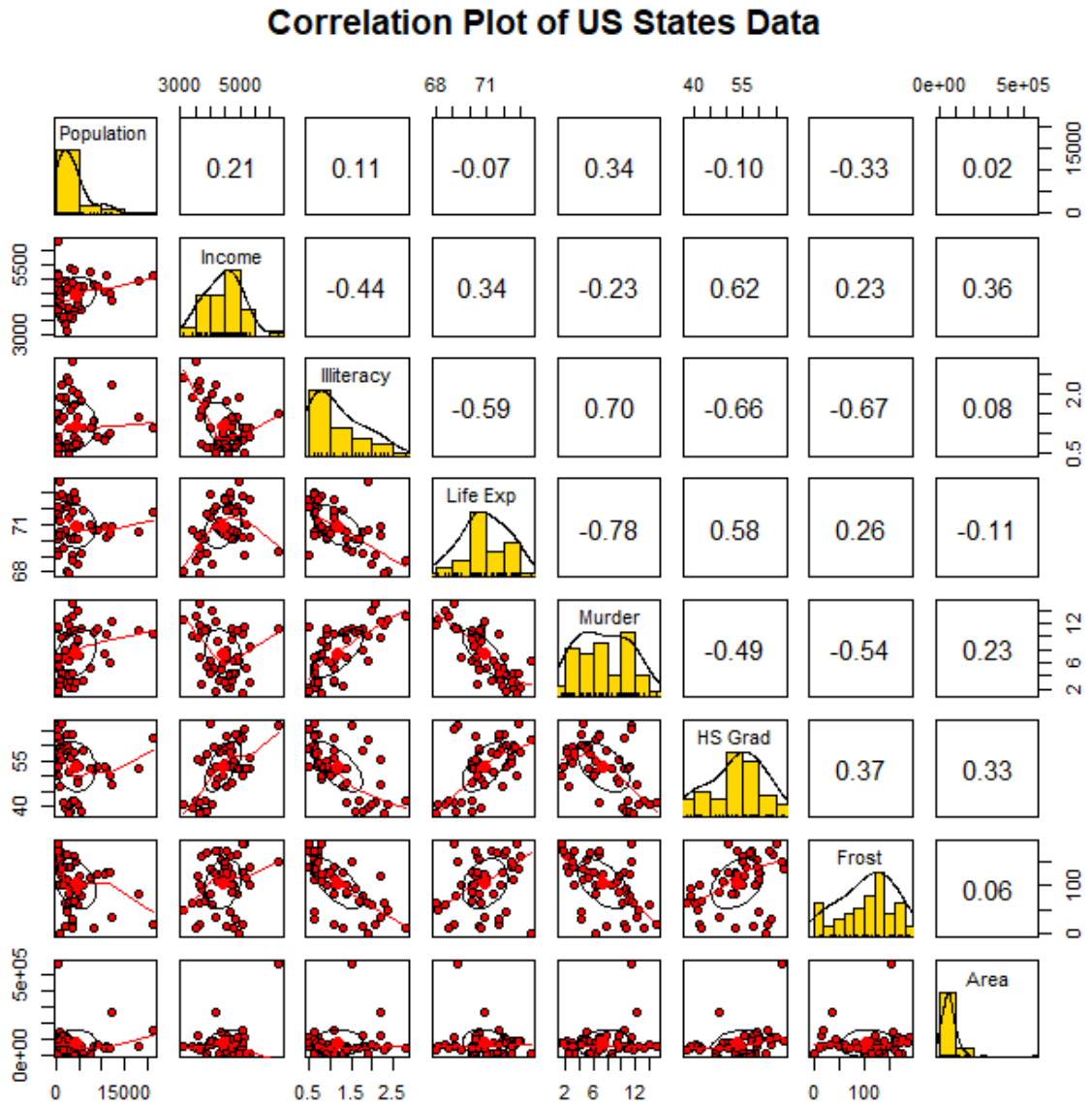
12. 상관관계와 상관분석

- state.x77 데이터셋의 변수 간 상관관계와 산점도 행렬 그래프

```
> old.op <- options(digits=2)
> cor(state.x77)
Population     Income  Illiteracy    Life Exp     Murder     HS Grad      Frost        Area
Population  1.00000000  0.2082276  0.10762237 -0.06805195  0.3436428 -0.09848975 -0.3321525  0.02254384
Income      0.20822756  1.0000000 -0.43707519  0.34025534 -0.2300776  0.61993232  0.2262822  0.36331544
Illiteracy  0.10762237 -0.4370752  1.00000000 -0.58847793  0.7029752 -0.65718861 -0.6719470  0.07726113
Life Exp   -0.06805195  0.3402553 -0.58847793  1.00000000 -0.7808458  0.58221620  0.2620680 -0.10733194
Murder      0.34364275 -0.2300776  0.70297520 -0.78084575  1.0000000 -0.48797102 -0.5388834  0.22839021
HS Grad    -0.09848975  0.6199323 -0.65718861  0.58221620 -0.4879710  1.00000000  0.3667797  0.33354187
Frost      -0.33215245  0.2262822 -0.67194697  0.26206801 -0.5388834  0.36677970  1.0000000  0.05922910
Area        0.02254384  0.3633154  0.07726113 -0.10733194  0.2283902  0.33354187  0.0592291  1.00000000
> options(old.op)
```

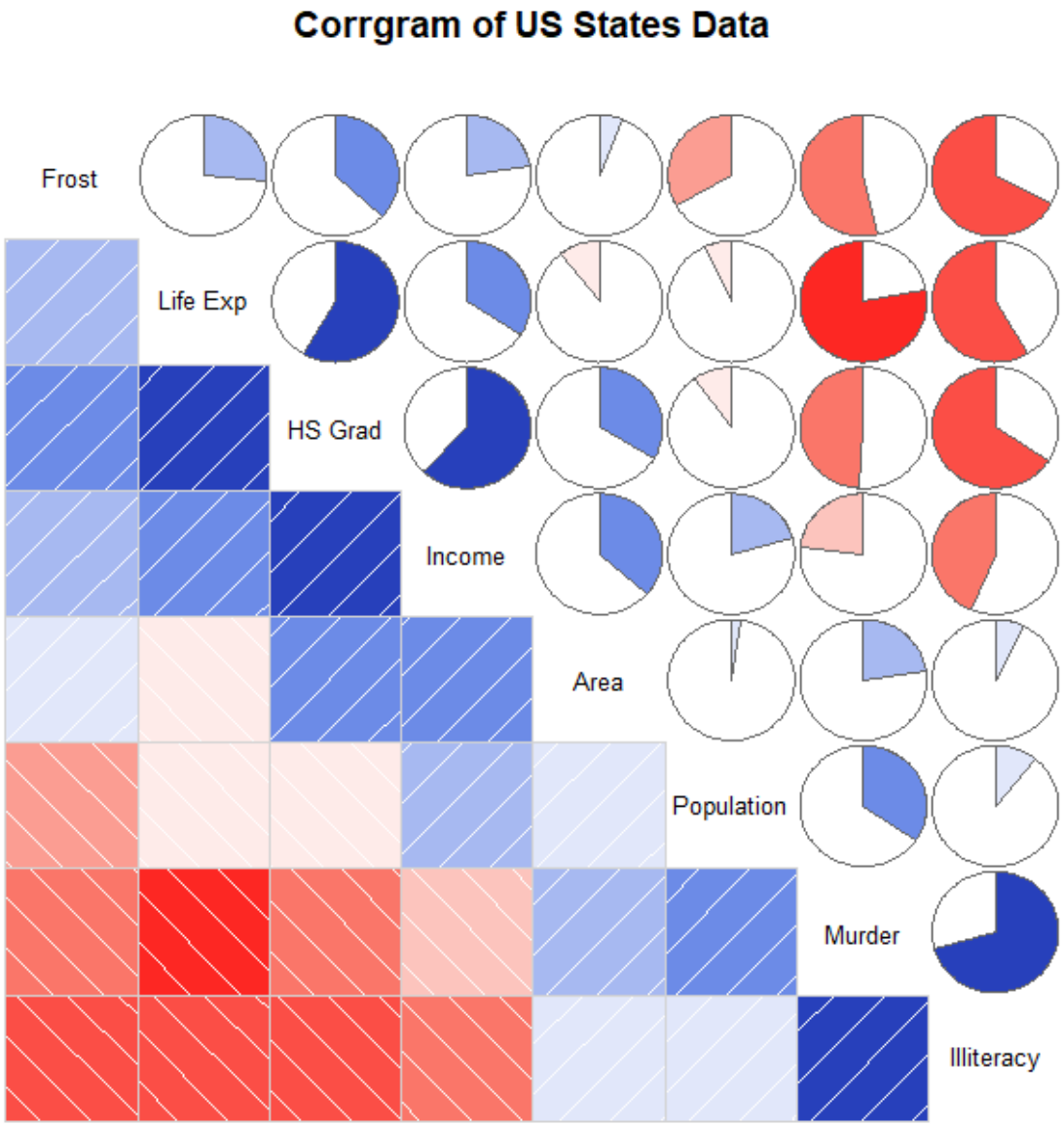
12. 상관관계와 상관분석

```
library(psych)
pairs.panels(state.x77,
main="Correlation Plot of US States Data",
             bg="red",
             pch=21,
             hist.col="gold")
```



12. 상관관계와 상관분석

```
library(corrgram)
corrgram(state.x77,
main="Corrgram of US States Data",
         order=TRUE,
         lower.panel=panel.shade,
         upper.panel=panel.pie,
         text.panel=panel.txt)
```



12. 상관관계와 상관분석

```
library(corrgram)
cols <- colorRampPalette(c("darkgoldenrod4", "burlywood1", "darkkhaki", "darkgreen"))
corrgram(state.x77,
main="Corrgram of US States Data",
         order=FALSE,
         col.regions=cols,
         lower.panel=panel.pie,
         upper.panel=panel.conf,
         text.panel=panel.txt)
```

Correlation does not imply *causation*!

- **편상관관계**: *partial* correlation
  - 두 변수 간의 관계를 분석할 때는 다른 변수의 영향을 주의 깊게 살펴봐야 함
    - 직장인의 연봉과 혈압 간의 관계: 양의 상관관계가 존재
    - 제 3의 변수는 나이:
    - 연봉과 혈압간의 관계를 분석하기 위해서는 나이 변수를 통제해야 함
  - **편상관계수**: *partial correlation coefficient*
    - 두 변수 간의 순수한 상관관계를 파악하기 위한 지표

12. 상관관계와 상관분석

- mtcars 데이터셋을 이용한 편상관분석

```
> colnames(mtcars)
[1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"   "am"   "gear" "carb“

> mtcars2 <- mtcars[, c("mpg", "cyl", "hp", "wt")]
> cor(mtcars2)
           mpg        cyl         hp         wt
mpg  1.0000000 -0.8521620 -0.7761684 -0.8676594
cyl -0.8521620  1.0000000  0.8324475  0.7824958
hp  -0.7761684  0.8324475  1.0000000  0.6587479
wt  -0.8676594  0.7824958  0.6587479  1.0000000
```

12. 상관관계와 상관분석

- 실린더 개수와 무게의 영향을 통제한 연비와 마력 간의 편상관계수 구하기

```
> cor(mtcars2[, c(1, 3)])
            mpg         hp
mpg  1.0000000 -0.7761684
hp  -0.7761684  1.0000000

> library(ggm)
> pcor(c("mpg", "hp", "cyl", "wt"), cov(mtcars2))
[1] -0.2758932

> pcor(c(1, 3, 2, 4), cov(mtcars2))
[1] -0.2758932
```

- **편상관계수에 대한 유의성 검정**

```
> pcor.test(pcor(c(1, 3, 2, 4), cov(mtcars2)), q=2, n=nrow(mtcars2))
$tval
[1] -1.518838

$df
[1] 28

$pvalue
[1] 0.1400152
```

```
> library(ppcor)
> pcor(mtcars2)
$estimate
……(중략)
$p.value
……(중략)
$statistic
……(중략)
$n
[1] 32
$gp
[1] 2
$method
[1] "pearson"

> pcor.test(mtcars2["mpg"], mtcars2["hp"], mtcars2[c("cyl", "wt")])
    estimate   p.value statistic  n gp  Method
1 -0.2758932 0.1400152 -1.518838 32  2 pearson
```

- 편상관계수를 이용한 숨겨진 관계 찾기
  - 변수 A와 변수 B 간에 기대되는 상관관계가 나타나지 않으면
    - 변수 A는 다른 변수 C와 양의 상관관계를 갖고
    - 동시에 변수 C가 변수 B와 음의 상관관계를 갖고 있을 수 있음
  - 예) 와인냉장고에 대한 '구매필요성'과 '구매의향' 간에 상관관계가 거의 없다.
    - 숨은 변수: '소득'을 고려하면
      - 소득과 구매필요성 간에는 음의 상관관계
      - 소득과 구매의향 간에는 양의 상관관계
    - 편상관분석을 이용하여 소득의 영향을 통제하면 상관관계를 확인할 수 있음

Any Questions?