3장 Correlation Analysis & Visualization

CONTENTS

3.1 서론

- 다중선형회귀분석을 비롯한 여러 가지 분석을 수행하기에 앞서 여러 변수들 간(종속변수와 예측변수, 예측변수들 간)의 상관관계를 파악하는 것은 매우 중요하다.
- 예를 들어, 예측변수들 간의 높은 상관은 회귀모형의 성능을 떨어뜨리는 요인으로 작용할 수 있으므로 미리 제거될 필요가 있다. 또한, 반응변수에 미치는 예측변수의 영향을 파악하고 이 를 모형구축에 활용할 수 있다.
- 이 장에서는 변수들 간의 다양한(모수적, 비모수적) 상관계수를 소개하고, R에서 이를 시각화하는 방법을 소개한다.

3.2 상관계수

- 표본상관계수(이하 상관계수)는 두 변수 간의 선형적인(또는 직선적인) 관계를 나타내는 측도이다. 상관계수는 -1과 1 사이의 값을 가지며, 1 또는 -1에 가까울수록 선형적인 관계가 강하며, 0에 가까울수록 그 관계가 약하다고 할 수 있다.
- 상관계수의 종류에는 피어슨 상관계수, 스피어만 상관계수, 켄달의 타우 등이 있다. 이 가운데 피어슨 상관계수는 모수적, 나머지는 비모수적 상관계수로 구분된다.
- cor() 함수는 상관분석을 수행한다. cor() 함수의 일반 형식은 다음과 같다.

```
cor(x, y=NULL, use="everything", method=c("pearson", "kendall",
    "spearman"))     # 디폴트는 "pearson"임
```

3.2 상관계수

(a) **피어슨 상관계수**

- 피어슨(Pearson) 상관계수 r은 두 개의 데이터 셋 $\{x_1, x_2, \dots, x_n\}$과 $\{y_1, y_2, \dots, y_n\}$으로부터 다음과 같이 정의된다.

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

참 고

Karl Pearson(1895)이 제안한 피어슨 상관계수는 '교차적률(product-moment) 상관계수' 또는 '이변량 상관'으로도 불린다. 모집단 버전은 다음과 같다.

$$\rho_{X,Y} = \frac{Cov(X,\ Y)}{\sigma_X \sigma_Y} = \frac{E(X-\mu_X)(Y-\mu_Y)}{\sigma_X \sigma_Y}.$$

3.2 상관계수

예제 1 longley 자료에 대해 상관분석을 수행한다.

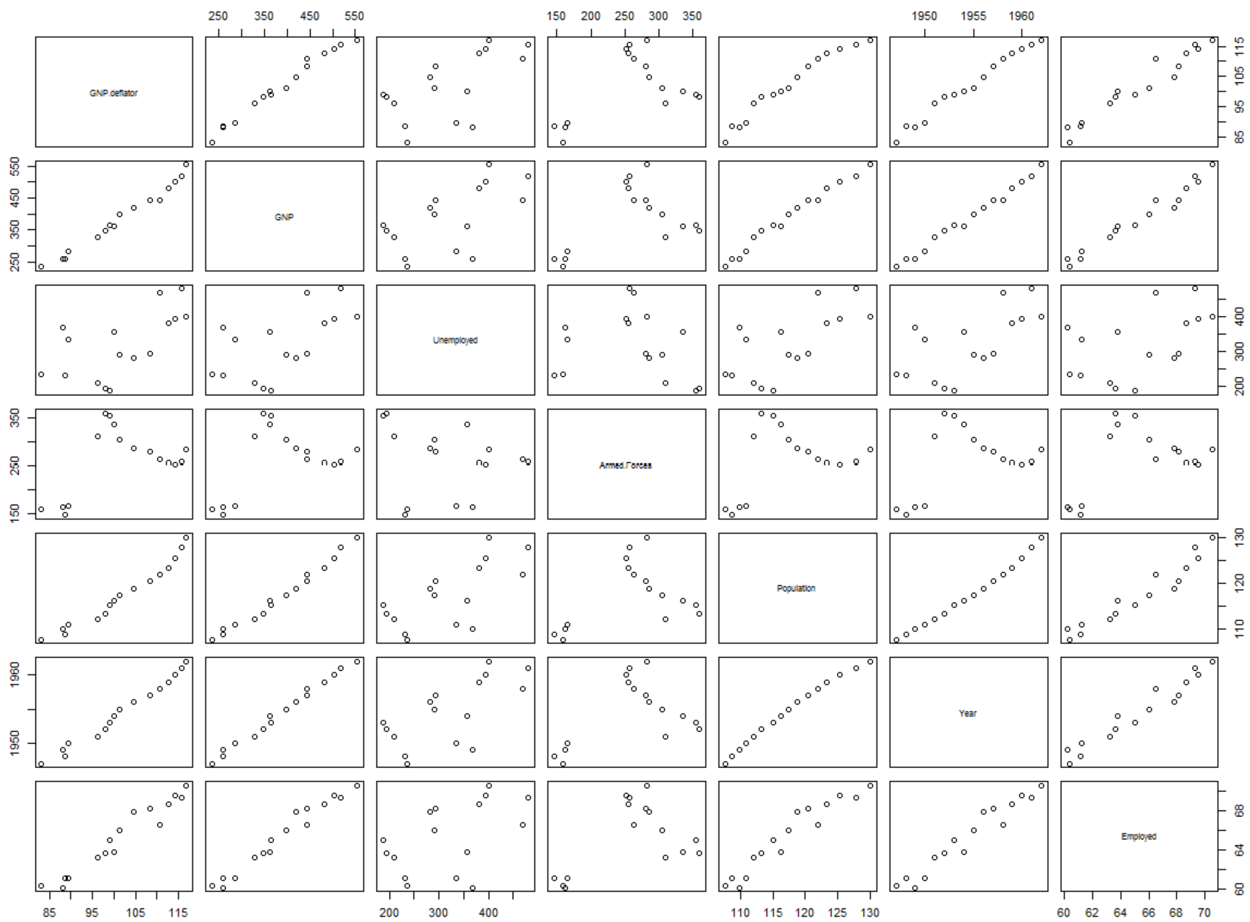
```
> data(longley)
> str(longley)
'data.frame':  16 obs. of  7 variables:
 $ GNP.deflator : num  83 88.5 88.2 89.5 96.2 ...
 $ GNP          : num  234 259 258 285 329 ...
 $ Unemployed   : num  236 232 368 335 210 ...
 $ Armed.Forces : num  159 146 162 165 310 ...
 $ Population   : num  108 109 110 111 112 ...
 $ Year         : int  1947 1948 1949 1950 1951 1952 1953 1954 ...
 $ Employed     : num  60.3 61.1 60.2 61.2 63.2 ...
```

3.2 상관계수

```
> cor(longley)
             GNP.deflator   GNP Unemployed Armed.Forces Population  Year Employed
GNP.deflator        1.000 0.991      0.621        0.465      0.979 0.991    0.971
GNP                 0.992 1.000      0.604        0.446      0.991 0.995    0.984
Unemployed          0.621 0.604      1.000       -0.177      0.687 0.668    0.502
Armed.Forces        0.465 0.446     -0.177        1.000      0.364 0.417    0.457
Population          0.979 0.991      0.687        0.364      1.000 0.994    0.960
Year                0.991 0.995      0.668        0.417      0.994 1.000    0.971
Employed            0.971 0.984      0.502        0.457      0.960 0.971    1.000
```

```
> pairs(longley)
```

3.2 상관계수


GNP.deflator
GNP
Unemployed
Armed.Forces
Population
Year
Employed

3.2 상관계수

(b) **스피어만 상관계수**

- 스피어만(Spearman) 상관계수 r_s는 일종의 비모수적 상관계수로 다음과 같이 구해진다. 먼저 두 개의 데이터 셋(원자료)을 각각 순위(rank) 자료로 전환한 뒤, 전환된 자료로부터 피어슨 상관계수를 구한 것으로 정의된다.

- 즉, 전환된 순위자료를 각각 $\{r_1, r_2, \dots, r_n\}$과 $\{s_1, s_2, \dots, s_n\}$이라고 할 때,

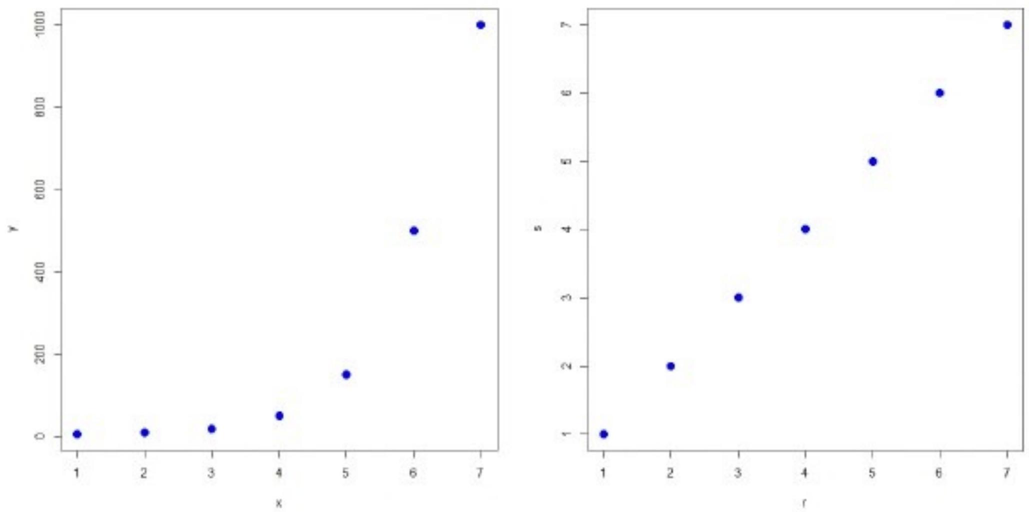
$$r_s = \frac{\sum_{i=1}^{n}(r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\sum_{i=1}^{n}(r_i - \bar{r})^2}\sqrt{\sum_{i=1}^{n}(s_i - \bar{s})^2}}$$

으로 정의된다.

3.2 상관계수

- 스피어만 상관계수는 원자료 대신 순위자료를 사용하므로 피어슨 상관계수보다 이상치 자료에 대해 덜 민감하게 반응한다. 따라서 스피어만 상관계수는 이상치가 포함된 자료에 대해 피어슨 상관계수보다 선호될 수 있다.
- 자료분석과정에서 두 상관계수의 차이가 클 경우에는 두 변수 간의 비선형성을 의심해 볼 필요가 있다. 예를 들어, 다음의 [그림 3.1]과 같이 비선형성이 강한 자료에 대해 스피어만 상관계수는 피어슨 상관계수에 비해 훨씬 큰 값을 제공한다.

3.2 상관계수



(a) 원자료(r=0.84) (b) 순위자료(r_s=1)

[그림 3.1] 피어슨과 스피어만 상관계수

3.2 상관계수

- 위의 longley 자료에 대해 스피어만 상관계수를 구하면 다음과 같다.

```
> cor(longley, method="spearman")
             GNP.deflator   GNP Unemployed Armed.Forces Population  Year Employed
GNP.deflator        1.000 0.997      0.664        0.220      0.997 0.997    0.982
GNP                 0.997 1.000      0.638        0.223      0.994 0.994    0.985
Unemployed          0.664 0.638      1.000       -0.341      0.685 0.685    0.564
Armed.Forces        0.220 0.223     -0.341        1.000      0.226 0.226    0.226
Population          0.997 0.994      0.685        0.226      1.000 1.000    0.976
Year                0.997 0.994      0.685        0.226      1.000 1.000    0.976
Employed            0.982 0.985      0.564        0.226      0.976 0.976    1.000
```

3.2 상관계수

(c) 켄달의 타우

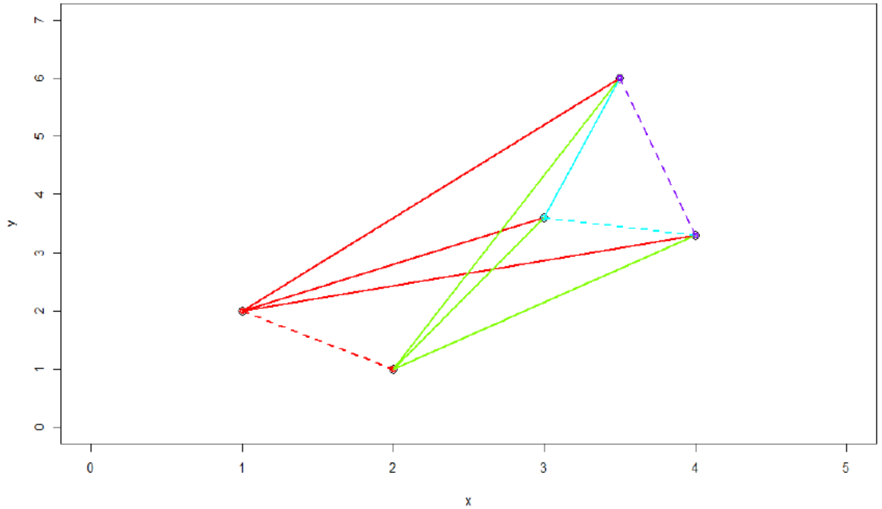
- 켄달의 타우(Kendall's tau) τ 역시 일종의 비모수적 상관계수로 다음과 같이 정의된다. 원자료 셋으로부터 부합인 쌍(concordant pairs)의 수를 P, 비부합인 쌍(disconcordant pairs)의 수를 Q라고 할 때

$$\tau = \frac{P - Q}{P + Q}$$

  으로 정의된다([그림 3.2]). 여기서 P와 Q는 각각 다음의 그림에서 기울기가 양인 직선과 음인 직선의 수를 의미한다.

3.2 상관계수

- 즉, 기울기가 양(또는 음)인 직선의 수 P(또는 Q)는 x값이 증가할 때, y값도 따라 증가하는(또는 감소하는) 점의 쌍이 몇 개인지를 나타내는 즉, 부합인(또는 비부합인) 직선의 수가 몇 개인가를 나타내는 값으로 이해될 수 있다.



[그림 3.2] 켄달의 타우($r = 0.4: P = 7, Q = 3$)

3.2 상관계수

- 위의 longley 자료에 대해 켄달의 타우(τ)를 구하면 다음과 같다.

```
> cor(longley, method="kendall")
             GNP.deflator   GNP Unemployed Armed.Forces Population  Year Employed
GNP.deflator       1.0000 0.983      0.450       0.0333      0.983 0.983    0.916
GNP                0.9833 1.000      0.433       0.0500      0.966 0.966    0.933
Unemployed         0.4500 0.433      1.000      -0.2166      0.466 0.466    0.366
Armed.Forces       0.0333 0.050     -0.216       1.0000      0.050 0.050    0.050
Population         0.9833 0.966      0.466       0.0500      1.000 1.000    0.900
Year               0.9833 0.966      0.466       0.0500      1.000 1.000    0.900
Employed           0.9166 0.933      0.366       0.0500      0.900 0.900    1.000
```

3.2 상관계수

(d) **상관계수에 대한 검정**

- 앞서 소개한 세 가지 종류의 상관계수에 대한 검정은 cor.test() 함수를 이용한다. 이 함수의 일반 형식은 다음과 같다.

```
cor.test(x, y, alternative=c("two.sided", "less", "greater"),
         method=c("pearson", "kendall", "spearman"),
         exact=NULL, conf.level=0.95, continuity=FALSE, …)
cor.test(formula, data, subset, na.action, …)
```

3.2 상관계수

- 이 가운데 피어슨 상관계수에 대한 검정은 다음의 $t-$검정을 이용한다. 모집단이 이변량 정규분포를 따른다는 가정하에서 다음의 가설

$$H_0 : \rho = 0, \qquad H_1 : \rho \neq 0$$

에 대한 검정은 표본상관계수를 이용한 다음의 검정통계량

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

이 귀무가설하에서 자유도가 $(n-2)$인 $t-$분포를 따르는 사실에 기초한다.

3.2 상관계수

예제 2

cats 자료에 대해 상관분석을 수행한다. cats 자료는 144마리 성인 고양이의 몸무게(kg)와 심장의 무게(g)를 측정한 자료이다.

```
> library("MASS")
> data(cats)
> 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 ...
```

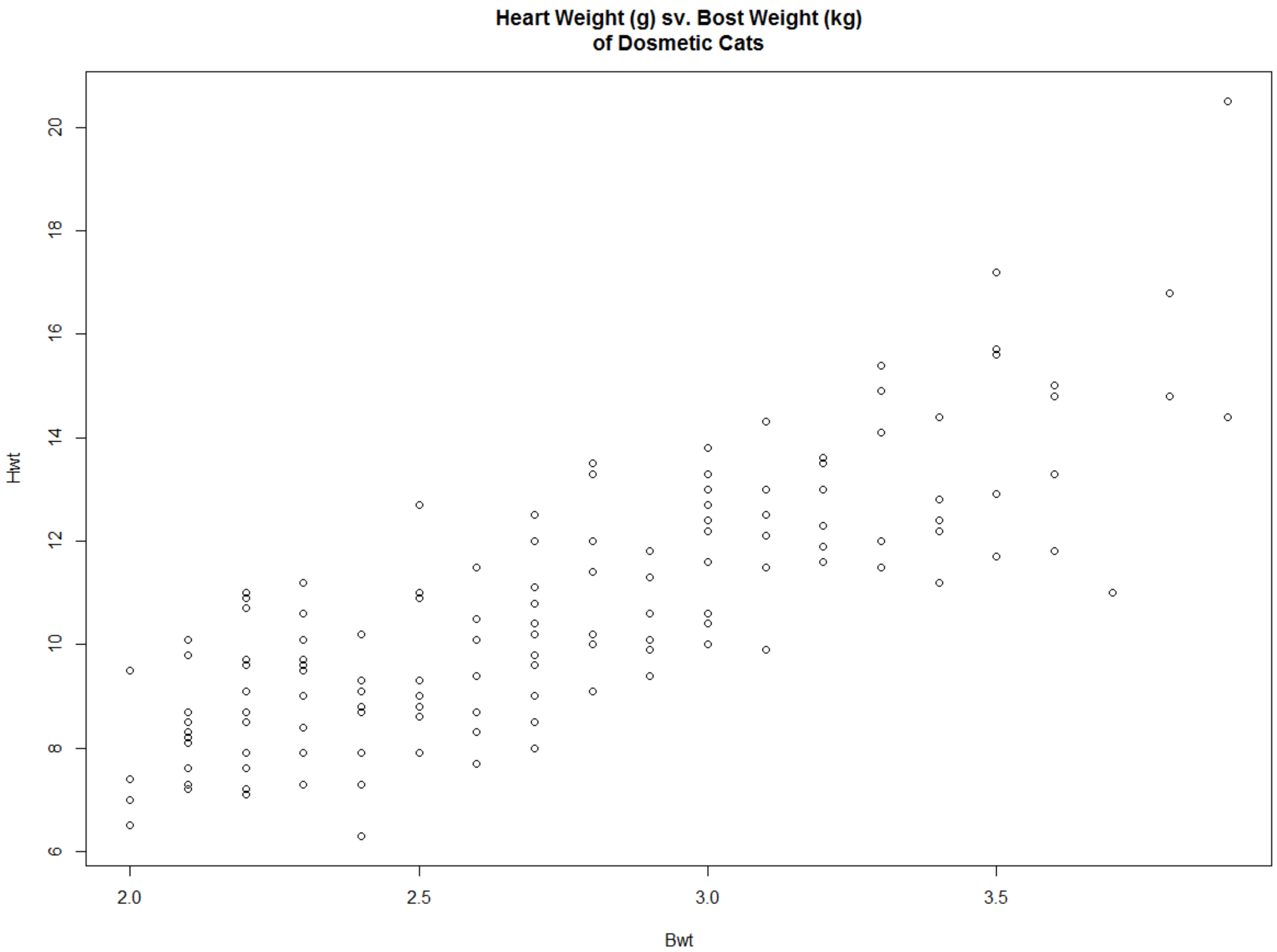
3.2 상관계수

```
> summary(cats)
 Sex             Bwt                 Hwt
 F:47    Min.     :2.000    Min.    : 6.30
 M:97    1st Qu. :2.300    1st Qu.: 8.95
         Median   :2.700    Median  :10.10
         Mean     :2.724    Mean    :10.63
         3rd Qu. :3.025    3rd Qu.:12.12
         Max.     :3.900    Max.    :20.50
```

- 위의 자료는 결측값을 가지는 변수가 없으며, 산점도는 다음과 같다.

```
> with(cats, plot(Bwt, Hwt))    # with(cats, plot(Hwt ~ Bwt))과 동일
> title(main="Heart Weight (g) sv. Bost Weight (kg)\nof Dosmetic
Cats")
```

3.2 상관계수



Heart Weight (g) sv. Bost Weight (kg)
of Dosmetic Cats
Hwt
Bwt
6
8
10
12
14
16
18
20
2.0
2.5
3.0
3.5

3.2 상관계수

- cor() 함수를 통해 두 변수 간의 피어슨 상관계수와 결정계수를 구하면 다음과 같다.

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

```
> with(cats, cor(Bwt, Hwt))^2     # 단순회귀에서는 결정계수=(상관계수의 제곱)임
[1] 0.6466209
```

3.2 상관계수

- cor.test() 함수를 통해 상관계수에 대한 검정을 수행한다.

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

       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
```

3.2 상관계수

- cor.test() 함수는 다음과 같이 공식을 사용할 수도 있다. 이 경우에는 subset= 옵션을 지원한다.

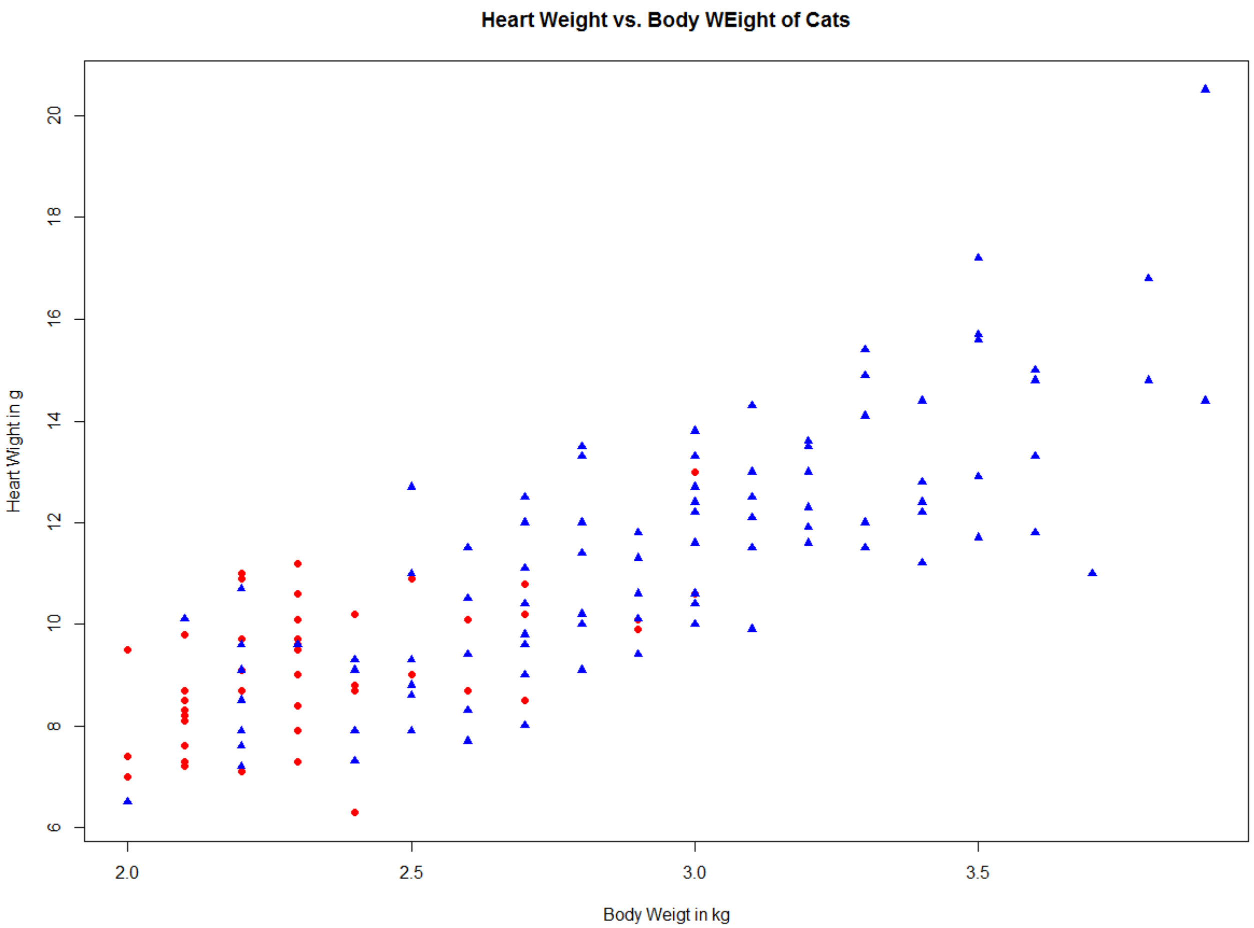
```
> with(cats, cor.test(Bwt, Hwt), subset=(Sex=="F"))

        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
```

3.2 상관계수

- 이 결과(cor ≈ 0.53)는 전체자료의 결과(cor ≈ 0.80)와 큰 차이를 보인다. 그 차이를 산점도를 통해 확인해 보면 다음과 같다.

```
      main="Heart Weight vs. Body  WEight of Cats"))
> with(cats, points(Bwt[Sex=="F"],  Hwt[Sex=="F"], pch=16,
      col="red"))
> with(cats, points(Bwt[Sex=="M"], Hwt[Sex=="M"], pch=17,
      col="blue"))
```

3.2 상관계수



Heart Weight vs. Body WEight of Cats
Heart Wight in g
6
8
10
12
14
16
18
20
2.0
2.5
3.0
3.5
Body Weigt in kg

3.3 R을 이용한 시각화

3.3.1 {psych} 패키지

- 상관행렬의 시각화를 위해 R 패키지 {psych}의 pairs.panel() 함수와 cor.plot() 함수를 이용한다.

3.3 R을 이용한 시각화

(a) pairs.panels{psych} 함수

- pairs.panels() 함수는 산점도와 함께 모든 변수들 간의 상관계수를 보여 준다. 변수의 수가 6~10개 이내인 경우 효과적이다.

- pairs.panels()의 일반형식과 주요 옵션은 다음과 같다.

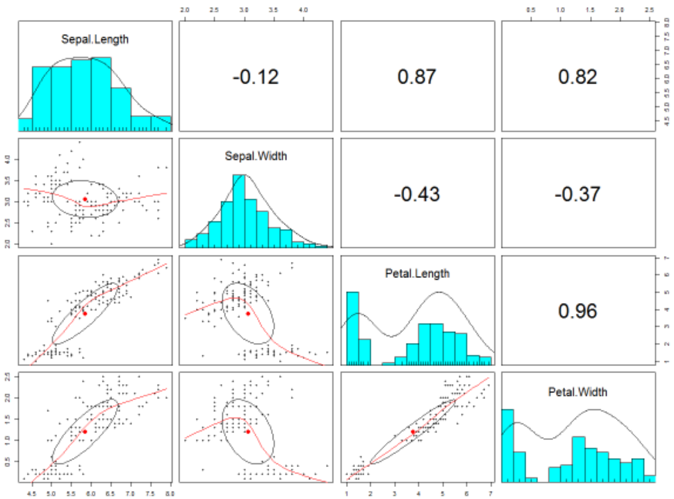
```
pairs.panels(x, smooth=TRUE, scale=TRUE, density=TRUE,
             ellipses=TRUE, lm=FALSE, digits=2, method="pearson",
             pch=20, cor=TRUE, jiggle=FALSE, factor=2,
             hist.col="cyan", show.points=TRUE, rug=TRUE,
             breaks=, cex.cor=, …)
```

3.3 R을 이용한 시각화

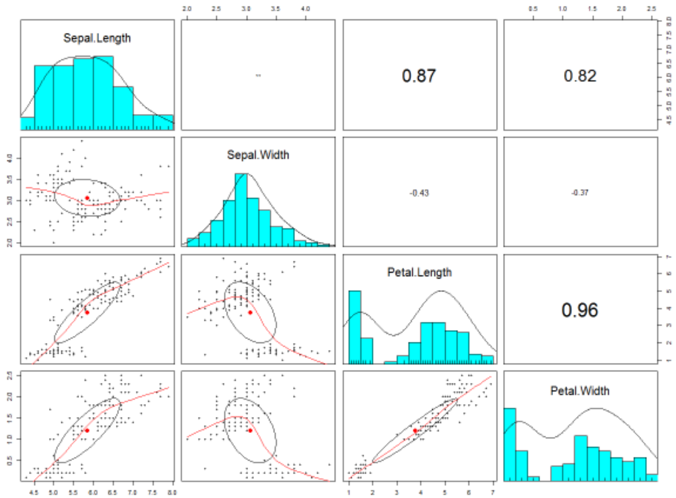
- x는 데이터프레임 또는 행렬
- smooth= loess 평활곡선그림
- scale= 상관계수값을 절댓값의 크기에 비례하게 폰트를 사용함
- method= “pearson”, “spearman”, “kendall”
- pch= 점의 형태 지정. 0~25까지 가능
- jiggle= 플롯에 흐트림(jitter) 적용 여부
- factor= 흐트림에 대한 인자(1-5)

3.3 R을 이용한 시각화

```
> # pairs.panels() 함수의 적용 예: iris 자료 이용

> data(iris)

> pairs.panels(iris[1:4], scale=T)
```



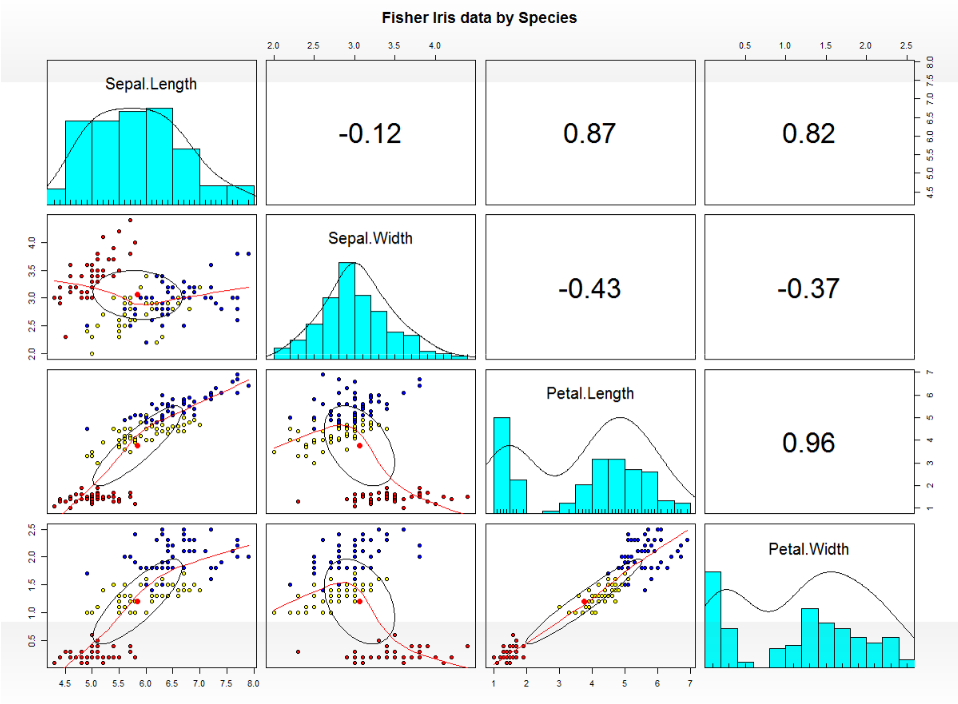
(a) scale=FALSE(디폴트)



(b) scale=TRUE

3.3 R을 이용한 시각화

```
> pairs.panels(iris[1:4], bg=c("red", "yellow",
            "blue")[iris$Species], pch=21, main="Fisher Iris
            data by Species")      # 그룹별로 색상지정
```



3.3 R을 이용한 시각화

(b) cor.plot{psych} 함수

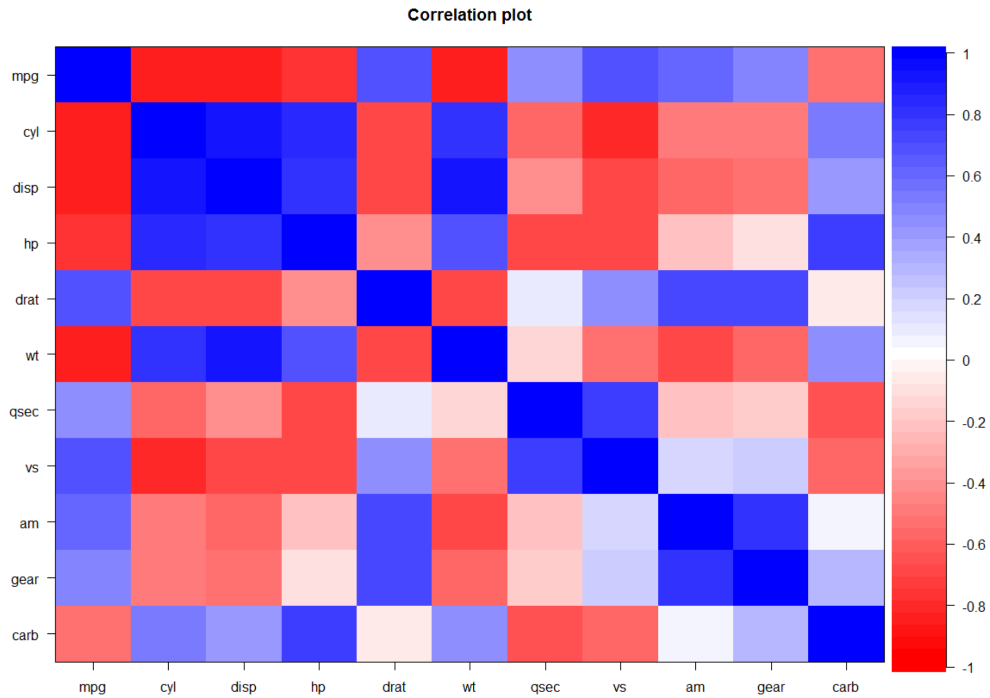
- cor.plot()는 변수의 수가 많을 경우에 유용한 방법으로 상관 또는 요인행렬에 대해 이미지 플롯을 제공한다.

```
cor.plot(r, numbers=FALSE, colors=TRUE, n=51, main=NULL, zlim=c(
         -1,1), show.legend=TRUE, labels=NULL, n.legend=10,
         keep.par=TRUE, select=NULL, pval=NULL, cuts=c(.001, .01),
         cex, MAR, upper=TRUE, diag=TRUE, …)
```

```
cor.plot.upperLowerCi(R, numbers=TRUE, cuts=c(.001, .01, .05),
                      select=NULL, main="Upper and lower confidence
                      intervals of correlations", …)
```

3.3 R을 이용한 시각화

```
> # cor.plot() 함수의 적용 예: mtcars 자료 이용
> cor.plot(cor(mtcars))
```



3.3 R을 이용한 시각화

corrplot{corrplot} 함수

- corrplot{corrplot} 함수는 상관행렬과 신뢰구간을 시각적으로 보여 준다. 일반 행렬의 시각화에도 사용될 수 있으며, 이 경우 is.corr=FALSE로 지정한다. 이 함수는 매우 다양한 옵션을 지원한다.

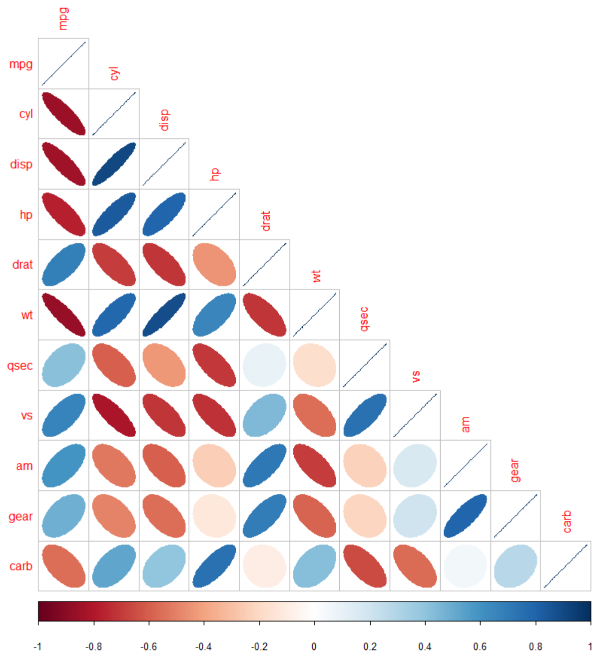
3.3 R을 이용한 시각화

```
corrplot(corr, method=, type=, add=FALSE, col=NULL, bg="white",
         title="", is.corr=TRUE, diag=TRUE, …, order=c("original",
         "AOE", "FPC", "hclust", "alphabet"),
         hclust.method=c("complete", "ward", "single", "average",
         "mcquitty", "median", "centroid"), addrect=NULL,
         rect.col="black", rect.lwd=2, …, p.mat=NULL,
         sig.level=0.05, insig=c("pch", "p-value", "blank", "n"),
         pch=4, pch.col="black", pch.cex=3, plotCI=c("n", "square",
         "circle", "rect"), lowCI.mat=NULL, uppCI.mat=NULL, …)
```

- method= "circle"(default), "square", "ellipse", "number", "shade", "color", "pie"
- type= "full"(default), "upper" or "lower" (layout type 지정)

3.3 R을 이용한 시각화

```
> # corrplot() 함수의 적용 예
> library(corrplot)
> M <- cor(mtcars)
> corrplot(M, method="ellipse", type="lower")
```



3.3 R을 이용한 시각화

(b) corrplot.mixed{corrplot} 함수

- corrplot.mixed{corrplot} 함수는 상관행렬에 대해 혼합된 방법으로 시각화를 제공한다.

```
corrplot.mixed(corr, lower="number", upper="circle",
               tl.pos=c("d", "lt", "n"), diag=c("n", "l", "u"),
               bg="white", addgrid.col="gray", …)
```

- lower=, upper= "circle"(upper default), "square", ellipse", "number"(lower default), "shade", "color", "pie"

3.3 R을 이용한 시각화

```
> # corrplot.mixed(corrplot)의 적용 예
> corrplot.mixed(M)      # corrplot.mixed(M, lower="number",
                          upper="circle")와 동일
```