노트북: blog
만든 날짜: 2020-10-06 오후 4:45
URL: https://continuous-development.tistory.com/55?category=793392

R

[R] R을 활용한 상관분석과 회귀분석 - 1

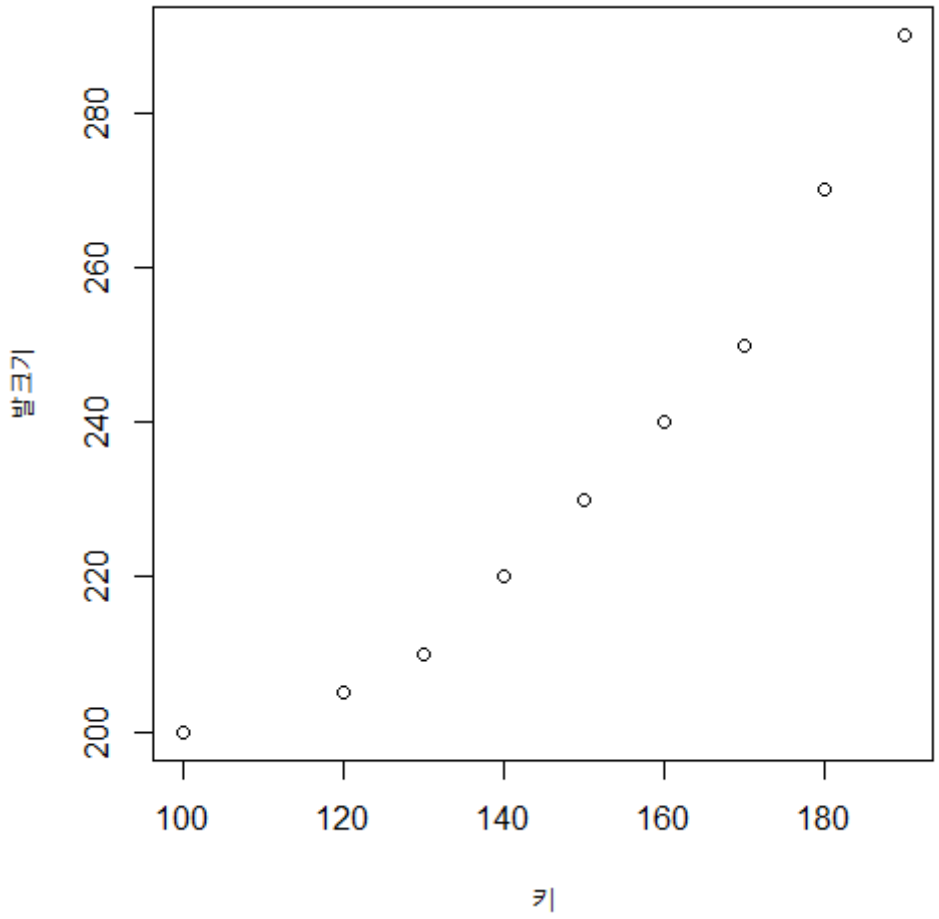
2020. 8. 6. 17:50 수정 삭제 공개

#상관분석

상관분석이란 하나의 변수와 다른 변수와의 밀접한 관련성이 있는지 분석하는 기법이다.
여기서는 상관분석을 통해 나온 상관계수와 그래프를 그리는 것까지 보여드릴 예정입니다.

```r
#지도 학습(문제와 답을 주고 그것을 통해 학습하는 방법)
# - 분류모델(classification)
# -- 알고리즘 ( KNN, SVN, D-TREE, Random Forest etc … )

# - 예측모델(prediction, estimation)
# -- 알고리즘(regression): logistic regression 예측알고리즘보다는 분류쪽 알고리즘으로 보고 있다.


#비지도 학습(문제만 있고 답이 없어 문제를 통해 학습하는 방법)
# -- 군집분석(clustering)
# -- 연관규칙(Association rule)
# -- 연속규칙(Sequence rule)


# 1. 단순 회귀 분석
# 상관분석 vs 회귀분석


# 상관분석 : 하나의 변수와 다른 변수와의 밀접한 관련성을 분석하는 기법
# cor()

# 회귀분석 : 두 변수간에 원인과 결과의 인과 관계가 있는지를 분석하는 기법
# lm()

height <- c(100,120,130,140,150,160,170,180,190)
foot   <- c(200,205,210,220,230,240,250,270,290)

plot(height, foot,
     xlab = '키',
     ylab = '발크기')
```

키와 발의 상관관계를 분석하기 위해 간단하게 데이터를 써서 넣었다. 그걸 plot차트로 시각화했다.



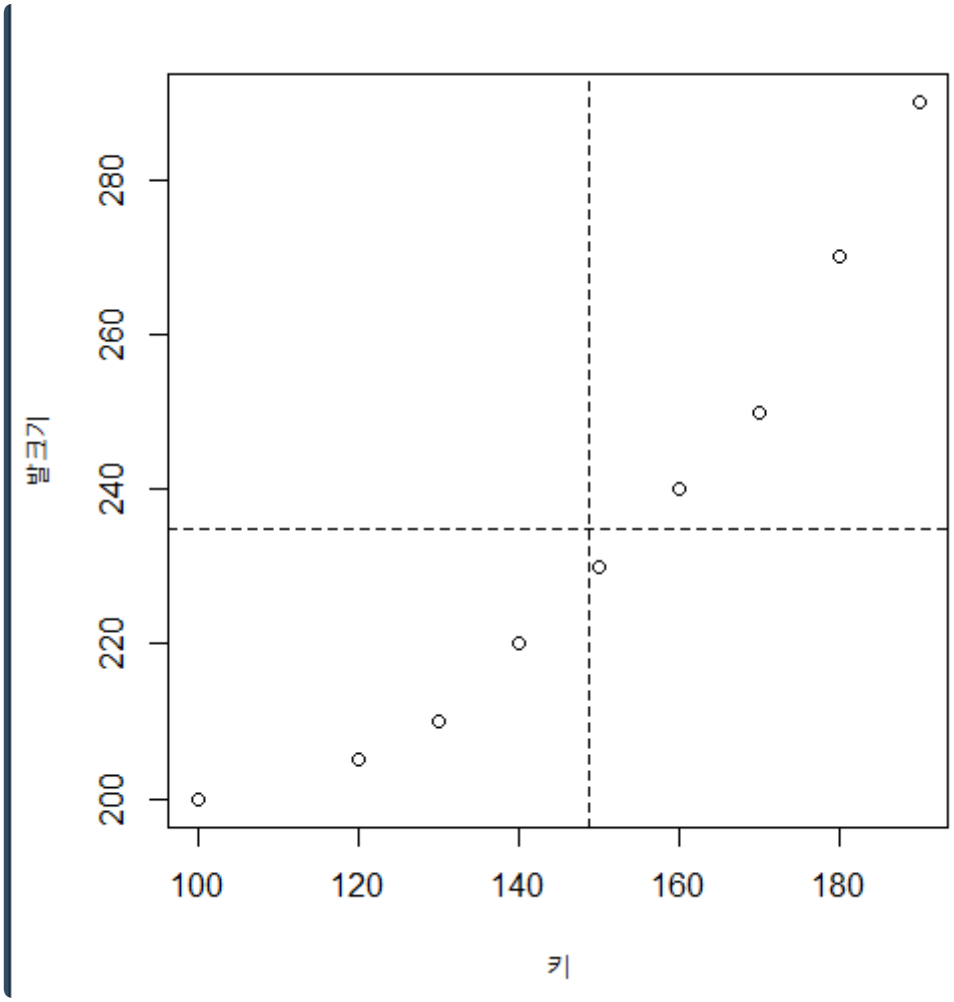
cor 이라는 명령어는 value 사이의 상관계수를 구하는 함수이다.

```
> cor(height,foot)
[1] 0.9599395
```

이때 상관계수가 1에 가까울수록 상관관계가 높음을 나타낸다.

```
37  abline(h=mean(foot), lty=2) #선을 긋는다.
38  abline(v=mean(height), lty=2)
39
```

abline으로 발 사이즈의 평균과 키의 평균을 선으로 그린다.



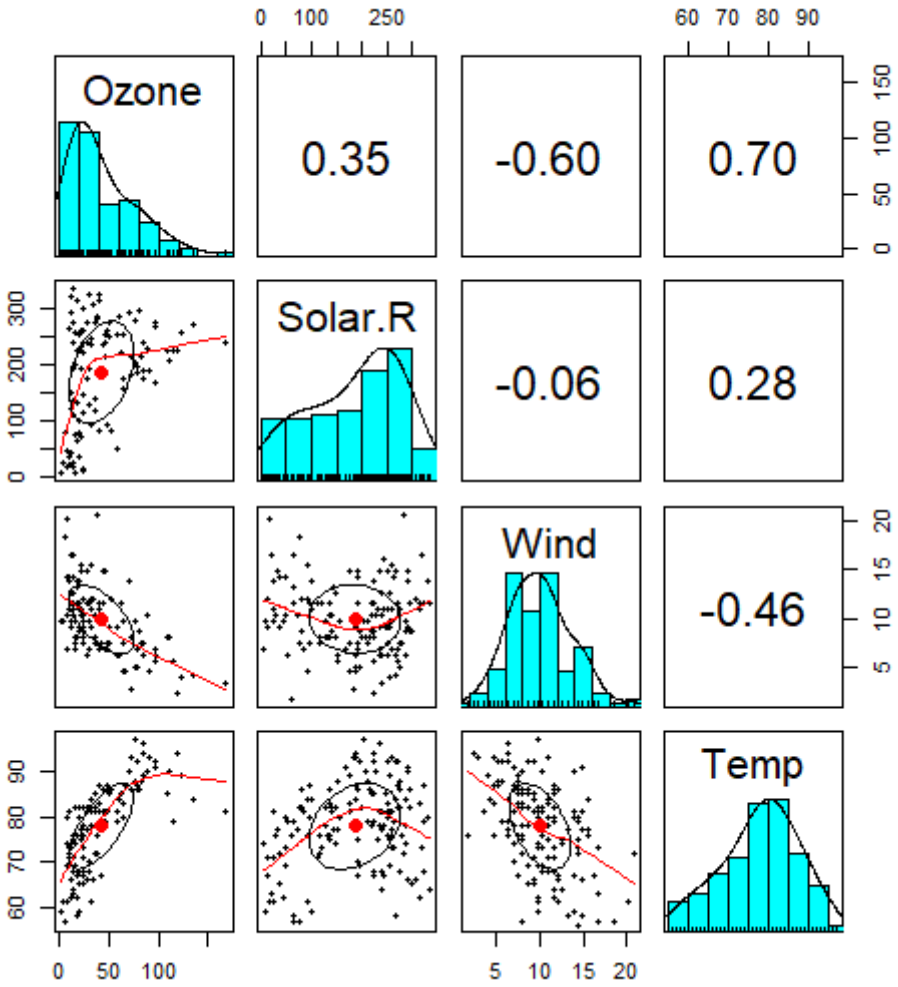
예제 airquality

Hmisc - 데이터 분석, 고급 그래픽, 유틸리티 작업, 샘플 크기 및 검정력 계산, 데이터 세트 가져오기 및 주석 달기, 결 측값 대치, 고급 테이블 작성, 변수 클러스터링, 문자열 조작, R 객체를 LaTeX로 변환하는 데 유용한 많은 기능이 있다.
psych -성격, 심리 이론 및 실험 심리학을위한 범용 툴박스. 함수는 주로 요인 분석, 주성분 분석, 군집 분석 및 신뢰도 분석을 사용하는 다변량 분석 및 척도 구성을 위한 것이지만 다른 함수는 기본적인 기술 통계를 제공해준다.

```r
airquality
str(airquality)

air01 <- airquality[ , c(1:4)]
install.packages("Hmisc")
library(Hmisc)
install.packages("psych")
library(psych)

pairs.panels(air01) #산점행렬도 그리기
```

airquality에서 몇 가지 속성 값을 가지고 와서 산점도 행렬을 그려보았다.

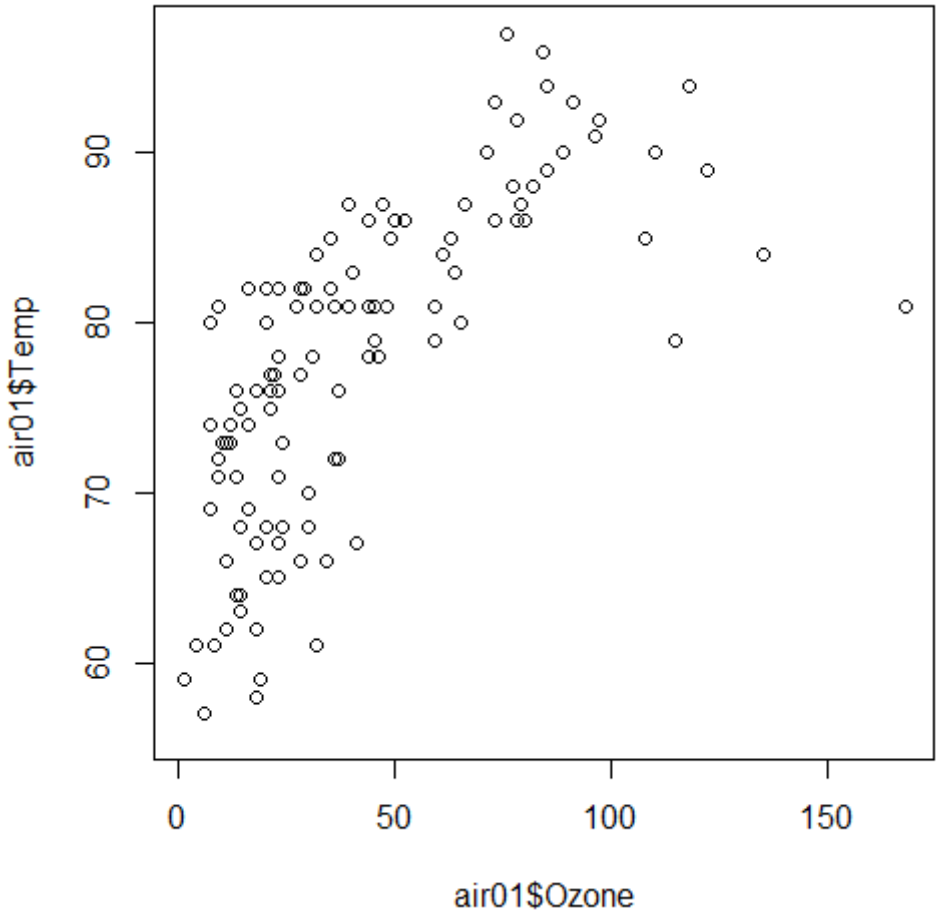
```
> air01 <- airquality[ , c(1:4)]
> air01
    Ozone Solar.R Wind Temp
1      41     190  7.4   67
2      36     118  8.0   72
3      12     149 12.6   74
4      18     313 11.5   62
5      NA      NA 14.3   56
6      28      NA 14.9   66
7      23     299  8.6   65
8      19      99 13.8   59
9       8      19 20.1   61
10     NA     194  8.6   69
11      7      NA  6.9   74
12     16     256  9.7   69
13     11     290  9.2   66
14     14     274 10.9   68
15     18      65 13.2   58
16     14     334 11.5   64
17     34     307 12.0   66
```



아래 산점행렬도를 봤을 때 Ozone이랑 Temp가 0.70의 상관계수로 가장 상관관계가 있다.

이 두가지 컬럼 값을 그래프로 두 가지 변수를 그래프로 그려본다.

```
plot(air01$Ozone,air01$Temp)
```



상관계수를 구하기 위해서는 결측값이 존재해서는 안된다.
이렇게 결측값이 있으면 cor을 쓸 수가 없다.

```
> cor(air01$Ozone,air01$Temp)
[1] NA
> cor(air01)
        Ozone Solar.R       Wind       Temp
Ozone       1      NA         NA         NA
Solar.R    NA       1         NA         NA
Wind       NA      NA  1.0000000 -0.4579879
Temp       NA      NA -0.4579879  1.0000000
> summary(air01)
     Ozone           Solar.R          Wind             Temp      
 Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00  
 1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:72.00  
 Median : 31.50   Median :205.0   Median : 9.700   Median :79.00  
 Mean   : 42.13   Mean   :185.9   Mean   : 9.958   Mean   :77.88  
 3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:85.00  
 Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00  
 NA's   :37       NA's   :7                                       
> 
```

아래 complete.cases 함수를 통해 결측치를 제외한 나머지 값을 air02에 넣는다.

```
61
62  # 모든 행에 대해서 출력
63  air01[!complete.cases(air01),]
64
65  # 결측치를 뺀 나머지값들을 air02에 저장
66  air02<-air01[complete.cases(air01),]
67
```

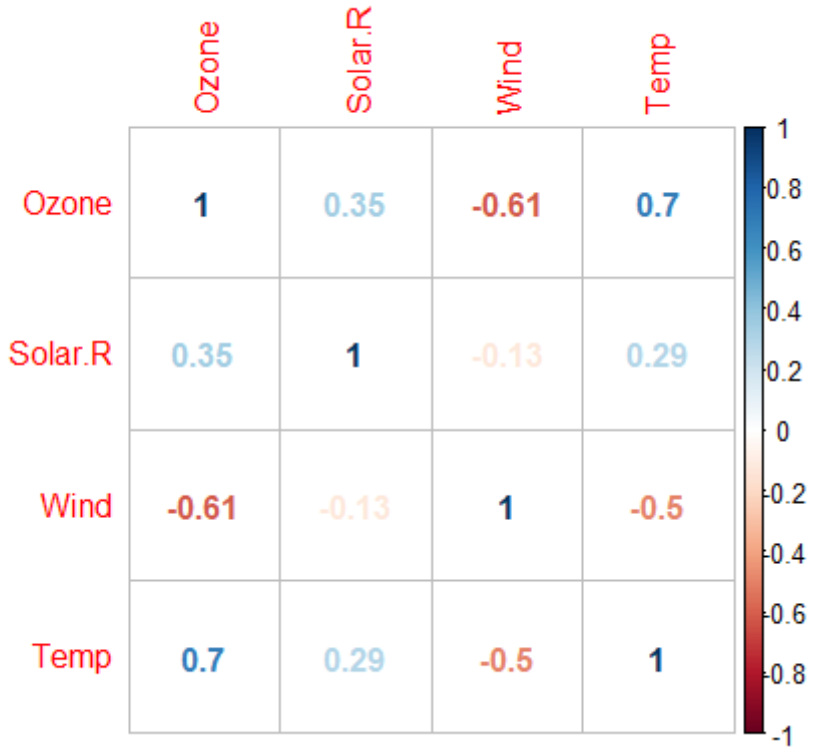
값을 확인해본다.

```
> str(air02)
'data.frame':   111 obs. of  4 variables:
 $ Ozone  : int  41 36 12 18 23 19 8 16 11 14 ...
 $ Solar.R: int  190 118 149 313 299 99 19 256 290 274 ...
 $ Wind   : num  7.4 8 12.6 11.5 8.6 13.8 20.1 9.7 9.2 10.9 ...
 $ Temp   : int  67 72 74 62 65 59 61 69 66 68 ...
> # -1 ≤ r ≤ 1 사이의 값을 가진다.
> cor(air02)
             Ozone    Solar.R       Wind       Temp
Ozone    1.0000000  0.3483417 -0.6124966  0.6985414
Solar.R  0.3483417  1.0000000 -0.1271835  0.2940876
Wind    -0.6124966 -0.1271835  1.0000000 -0.4971897
Temp     0.6985414  0.2940876 -0.4971897  1.0000000
```

#상관 계수를 시각화 하는 작업

```
83  # 상관계수를 시각화를 통해서 표현해 본다면?
84  # method - circle, square, ellipse, shade, color, pie
85  corrplot(air.cor, method = "number") #air.cor를 숫자로 표현해준다.
86
87  corrplot(air.cor, method = "circle")
88
89
```

method의 종류에 따라 다양한 방식으로 출력이 된다.



예제

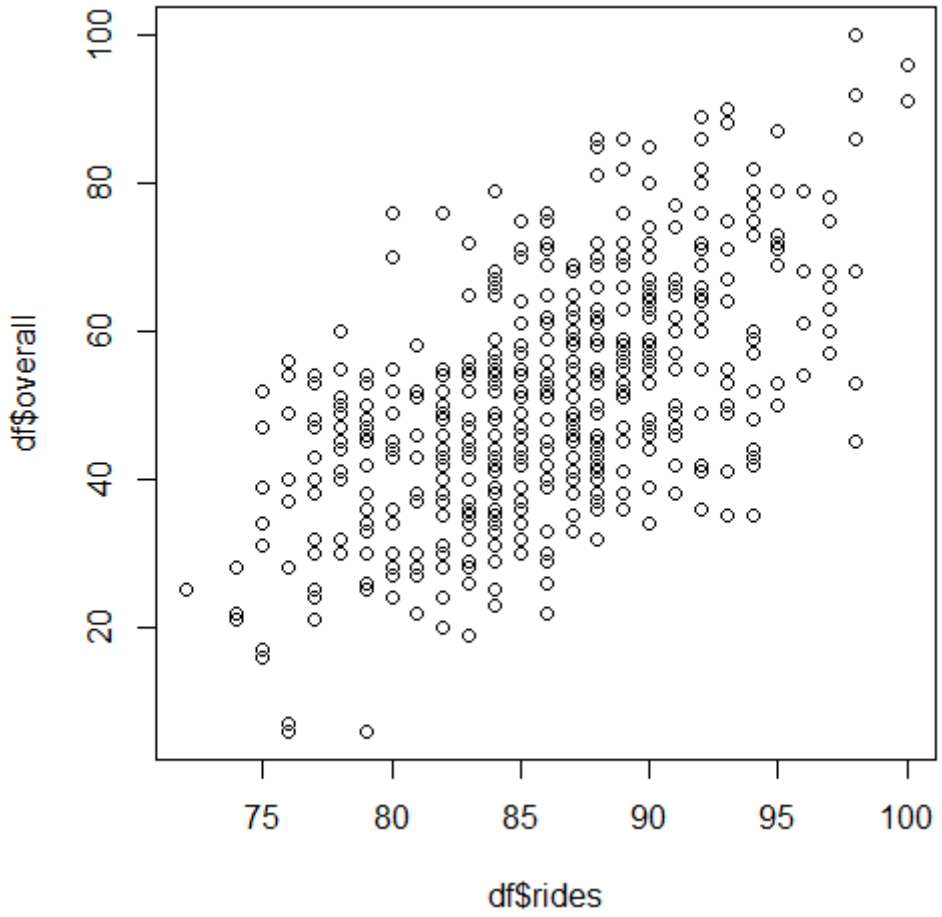
```
#---실습

df ← read.csv("http://goo.gl/HKnl74")
str(df)

#속성별 결측 값 확인
colSums(is.na(df))
```

데이터를 인터넷으로 받고 값을 확인해봤다.

```
> str(df)
'data.frame':   500 obs. of  8 variables:
 $ weekend  : chr  "yes" "yes" "no" "yes" ...
 $ num.child: int  0 2 1 0 4 5 1 0 0 3 ...
 $ distance : num  114.6 27 63.3 25.9 54.7 ...
 $ rides    : int  87 87 85 88 84 81 77 82 90 88 ...
 $ games    : int  73 78 80 72 87 79 73 70 88 86 ...
 $ wait     : int  60 76 70 66 74 48 58 70 79 55 ...
 $ clean    : int  89 87 88 89 87 79 85 83 95 88 ...
 $ overall  : int  47 65 61 37 68 27 40 30 58 36 ...
> #속성별 결측 값 확인
> colSums(is.na(df))
  weekend num.child  distance     rides     games      wait     clean   overall
        0         0         0         0         0         0         0         0
```

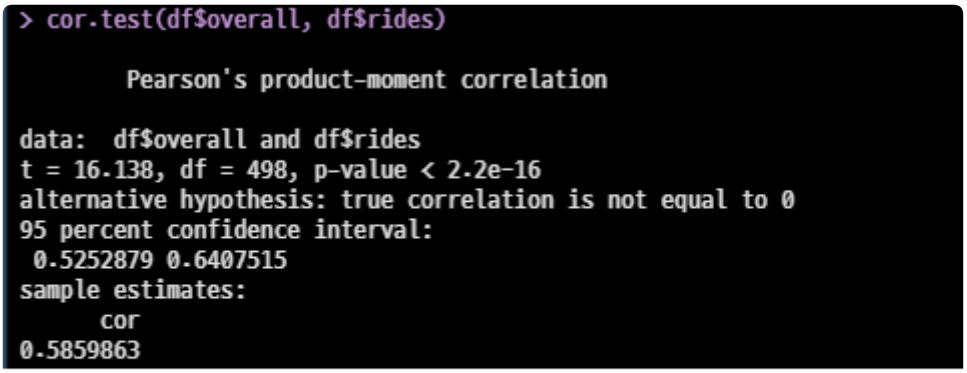
```
 98  #놀이기구의 만족도가 높으면 전체 만족도 또한 높지않을까 예상해보자
 99  plot(df$overall ~ df$rides)
100  cor(df$overall, df$rides)
101
```

```
> cor(df$overall, df$rides)
[1] 0.5859863
```



cor.test( )

cor.test는 상관계수 검정을 하는 함수로서 상관 계수 검정 Correlation Test을 수행하여 상관 계수의 통계적 유의성을 판단할 수 있다.

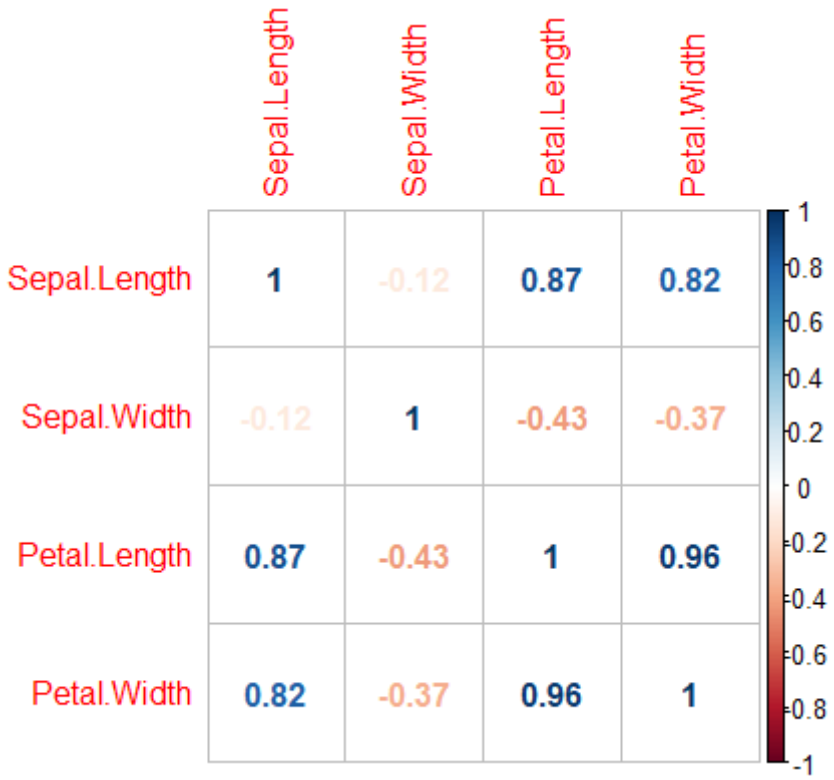


```
> cor.test(df$overall, df$rides)

        Pearson's product-moment correlation

data:  df$overall and df$rides
t = 16.138, df = 498, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5252879 0.6407515
sample estimates:
      cor
0.5859863
```

지금 귀무가설에 대해 대립가설로 검정을 진행한다고 가정했을 때
지금 결과에서는 95프로의 신뢰구간이 0.52589~ 0.6407515 정도가 되고 상관계수 값이 이 안에 들어온다면 대립 가설을 채택한다.
우리가 구한 cor(상관계수) 은 0.5859863이다.
p-value 의 유의 수준이 0.05이다. 2.2의 -16승이다. 이 결과를 보고 귀무가설이 잘못되고 대립 가설을 채택해야 된다고 생각해야 한다.
t는 검정 통계량, p-value는 **유의 확률이다.**
우리가 봐야될것은 p-value(유의 확률)를 봐야 한다.
p-value - 내가 현재 구한 통계 값이 얼마나 자주 나올 것인가를 나타낸다.

실습

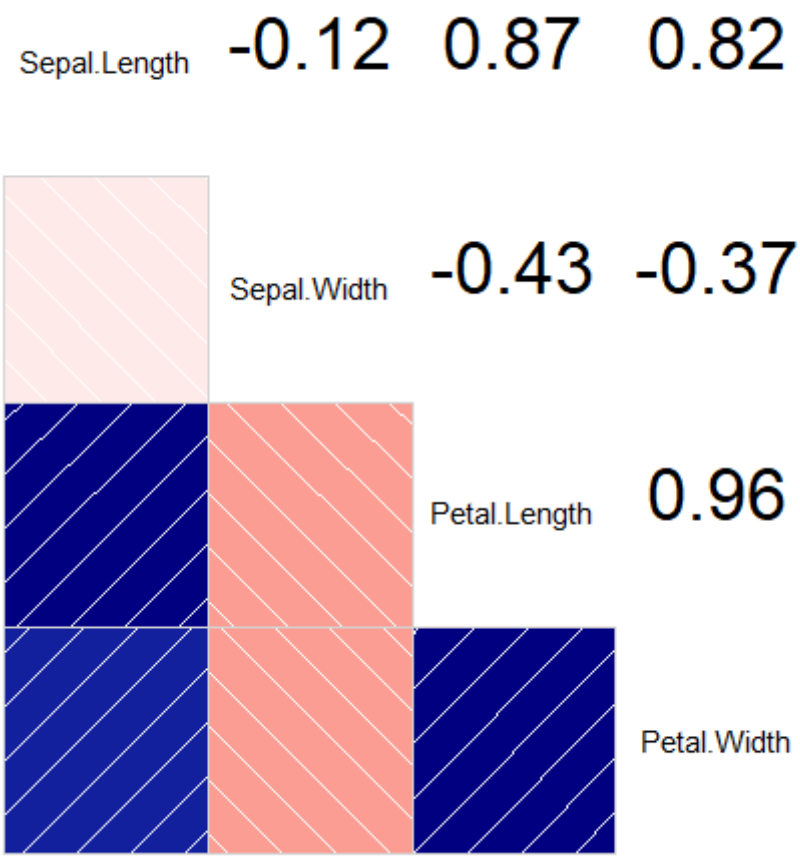
```r
#-- 실습 iris
iris
str(iris)
colSums(is.na(iris))


# 가설
# 꽃받침의 길이가 길수록 꽃잎의 넓이도 크다
# Sepal : 꽃받침
# Petal : 꽃잎
iris2 <- iris[1:4]
iris.cor <- cor(iris2)

# 상관계수를 시각화를 통해서 표현해 본다면?
# method - circle, square, ellipse, shade, color, pie
corrplot(iris.cor, method = "number") #air.cor를 숫자로 표현해준다.
plot(iris$overall ~ iris$rides)
```




```r
#symnuf 를이용해서 상관관계를 볼수있다. 여기서는 B가 제일 크고 +도 크다 라는 것을 말해준다.
symnum(iris.cor)
```

```
> symnum(iris.cor)
             S.L S.W P.L P.W
Sepal.Length 1
Sepal.Width      1
Petal.Length +   .   1
Petal.Width  +   .   B   1
attr(,"legend")
[1] 0 ‘ ’ 0.3 ‘.’ 0.6 ‘,’ 0.8 ‘+’ 0.9 ‘*’ 0.95 ‘B’ 1
```

```r
install.packages("corrgram")
library(corrgram)

corrgram(cor(iris[ ,1:4]), type='corr',
         upper.panel = panel.conf)
```



'**R**' 카테고리의 다른 글□

[R] R을 활용한 크롤링 - 로또 1등 당첨 배출점 크롤링 하기□

[R] R에서 교차검증을 위한 데이터 셋 분리방법 3가지□

[R] R을 활용한 상관분석과 회귀분석 - 1□

[R] R을 통한 텍스트마이닝에서 워드클라우드 까지□

[R] R로 하는 비정형 데이터 처리 (facebook 데이터를 통한 긍정/부정 나누기)□

[R] R에서 Database 사용하기 / DB 기본적인 구문 사용하기□

cor.test R cor R cor.test R 상관분석 R을 활용한 상관분석 상관계수 상관분석

꾸까꾸

혼자 끄적끄적하는 블로그 입니다.