1. 상관 분석

A. 상관계수

1) 피어슨 상관계수(Pearson correlation coefficient)

두 변수간의 관련성을 구하기 위해 보편적으로 이용됨

r = X와 Y가 함께 변하는 정도 / X와 Y가 따로 변하는 정도

결과의 해석

r 값은 X 와 Y 가 완전히 동일하면 +1, 전혀 다르면 0, 반대방향으로 완전히 동일하면 −1 을 가진다.

결정계수 (coefficient of determination) 는 r^2 로 계산하며 이것은 X 로부터 Y 를 예측할 수 있는 정도를 의미한다.

일반적으로

r이 -1.0과 -0.7 사이이면, 강한 음적 선형관계,
r이 -0.7과 -0.3 사이이면, 뚜렷한 음적 선형관계,
r이 -0.3과 -0.1 사이이면, 약한 음적 선형관계,
r이 -0.1과 +0.1 사이이면, 거의 무시될 수 있는 선형관계,
r이 +0.1과 +0.3 사이이면, 약한 양적 선형관계,
r이 +0.3과 +0.7 사이이면, 뚜렷한 양적 선형관계,
r이 +0.7과 +1.0 사이이면, 강한 양적 선형관계

로 해석한다.

2) Spearman 상관계수(Spearman correlation coefficient)

데이터가 서열척도인 경우 즉 자료의 값 대신 순위를 이용하는 경우의 상관계수로서, 데이터를 작은 것부터 차례로 순위를 매겨 서열 순서로 바꾼 뒤 순위를 이용해 상관계수를 구한다. 두 변수 간의 연관 관계가 있는지 없는지를 밝혀주며 자료에 이상점이 있거나 표본크기가 작을 때 유용하다. 스피어만 상관계수는 -1과 1 사이의 값을 가지는데 두 변수 안의 순위가 완전히 일치하면 +1이고, 두 변수의 순위가 완전히 반대이면 -1이 된다. 예를 들어 수학 잘하는 학생이 영어를 잘하는 것과 상관있는지 없는지를 알아보는데 쓰일 수 있다.

```
x<-c(0,1,4,9)
y<-c(1,2,3,4)
z<-c(0,5,7,9)
mean(x)
mean(y)
mean(z)
```

```
3.5
2.5
5.25
```

```
cor(x,y,method="pearson") #기본값
cor(x,y,method="spearman")

cor(y,z,method="pearson")
cor(y,z,method="spearman")

cor(x,z,method="pearson")
cor(x,z,method="spearman")
```

3) 담배값의 인상이 흡연에 미치는 영향을 분석

담배값 인상 전의 월별 매출액 자료와 인상 후의 월별 매출액 자료를 조사하여 상관분석을 하면 담배값의 인상이 흡연에 미치는 영향을 분석할 수 있다.

귀무가설 : 담배값 인상과 흡연과 상관관계가 없다
대립가설 : 담배값 인상과 흡연과 상관관계가 있다

```
x<-c(70,72,62,64,71,76,0,65,74,72)
y<-c(70,74,65,68,72,74,61,66,76,75)
cor.test(x,y,method="pearson")
```

```
  Pearson's product-moment correlation

data:  x and y
t = 3.4455, df = 8, p-value = 0.008752
alternative hypothesis: true correlation is not equal to
0
95 percent confidence interval:
 0.2791495 0.9434286
sample estimates:
      cor
0.7729264
```

p-value가 0.05보다 작으므로 대립가설을 채택한다. 즉, 담배값 인상과 흡연과 상관관계가 있다.
cor 상관계수가 0.77 강한 양의 상관관계

4) 공분산과 상관계수

```
df<-read.csv("d:/data/rides/rides.csv")
head(df)
```

```
#산점도
plot(df$overall~df$rides) # y ~ X
# rides와 overall은 양의 상관관계가 있는 것으로 보임
```

```
#공분산(Covariance): 두 변수의 상관정도를 나타내는 값, 두 변수가 같은 방향으로 움직이는 정도
#  x의 편차와 y의 편차를 곱한 값의 평균값
# X 증가 => y 증가 => 양수
# X 증가 => y 감소 => 음수
# 공분산이 0이면 두 변수는 선형관계가 없음

cov(df$overall, df$rides)
# 양수이므로 양의 상관관계임
# 공분산은 증가, 감소 방향을 이해할 수는 있으나 어느 정도의 상관관계인지 파악하기는 어려움
```

```
cov(1:5, 2:6)  # x,y가 같은 방향으로 증가하므로 양수
cov(1:5, rep(3,5))  # x의 변화에 y가 영향을 받지 않으므로 0
```

```
cov(1:5, 5:1)  # x,y의 증가 방향이 다르므로 음수
cov(c(10,20,30,40,50), 5:1)
# 공분산은 변수의 단위에 크게 영향을 받는 단점이 있음.
# 이것을 보완하기 위해 공분산을 표준화시킨 상관계수를 사용함

#상관계수 :  X와 Y가 함께 변하는 정도 / X와 Y가 각각 변하는 정도
# 공분산을 표준편차의 곱으로 나눈 값(-1 ~ 1)
# +1 : 완벽한 양의 상관관계, -1 : 완벽한 음의 상관관계
# 0 : 선형관계가 없음

# 피어슨 상관계수: 일반적으로 사용되는 방법, 숫자형-숫자형 변수, 정규분포인 경우 정확한 결과를 얻을 수 있음, 이상치에 민감함
# 스피어만 상관계수: 서열척도의 경우 사용
#    직선관계가 아니어도 상관관계가 있으면 1에 가까운 값을 갖게 됨

# 상관분석은 선형관계를 설명할 수는 있으나 인과관계(원인과 결과)를 설명하기는 어려움
# 원인과 결과를 설명하려면 회귀분석을 사용해야 함
```

```
cor(1:5, 5:1)
cor(c(10,20,30,40,50), 5:1)
```

```
#피어슨 상관계수
cor(df$overall, df$rides, method='pearson')
#  use='complete.obs'  결측값을 제외하고 계산하는 옵션
cor(df$overall, df$rides, use='complete.obs',
method='pearson')
```

```
#상관계수 검정: 상관계수의 통계적 유의성 판단
#통계적으로 유의하다는 것은 관찰된 현상이 전적으로 우연에 의해
벌어졌을 가능성이 낮다는 의미
#  귀무가설: 상관계수가 0이다
#  대립가설: 상관계수가 0이 아니다
cor.test(df$overall, df$rides, method = "pearson",
conf.level = 0.95)
#cor.test(iris$Sepal.Length, iris$Petal.Length, method =
"pearson", conf.level = 0.95)
#p-value가 0.05 이하이므로 귀무가설 기각
# 결론: 두 변수는 선형적으로 상관관계가 있음
# 95% 신뢰구간: 0.5252879 0.6407515
# 피어슨 상관계수: 0.5859863
```

```
head(df[,4:8])
```

```
#산점도 행렬
plot(df[,4:8])
```

```
#추세선(회귀선) 그리기
pairs(df[,4:8], panel=panel.smooth)
```

```
#install.packages("PerformanceAnalytics")
library(PerformanceAnalytics)
chart.Correlation(df[,4:8], histogram=TRUE, pch=19)
#산점도, 히스토그램, 상관계수가 함께 출력됨
# rides와 games 0.46
# rides와 clean 0.79
```

```
#결측값이 있는 경우
#df <- na.omit(df)
#상관계수 행렬
cor(df[,4:8])
```

```
#상관계수 플롯
#install.packages('corrplot')
library(corrplot)
X<-cor(df[,4:8])
corrplot(X) #원의 크기로 표시됨
```

```
#숫자로 출력됨
corrplot(X, method="number")
```

```
# method: circle,square,ellipse,number,shade,color,pie
corrplot.mixed(X, lower='ellipse',upper='circle')
```

```
#계층적 군집의 결과에 따라 사각형 표시, addrect 군집개수
#hclust ; hierarchical clustering order(계층적 군집순서)로 정렬
corrplot(X,order="hclust",addrect=3)
```