데이터 통계 분석 기초 2

환경생태데이터사이언스 실습 November 19, 2019

오늘의 학습 목표

1. 상관분석
2. 단순 회귀분석

상관분석

상관분석 (Correlation Analysis)

- 상관분석[1]
  - 특정한 변수들 사이의 관계 분석
    1. 서로 관계가 있는가?
    2. 어느 정도의 관계를 갖고 있는가?
  - 관계: 변수 간 밀접한 상호작용의 결과물
    - 상관분석을 통해 각 변수에 대한 이해도 높임.
    - 변수의 변화에 대한 설명력 혹은 예측력 향상.

[1]출처: R을 활용한 기초 환경자료 분석 및 시각화

상관분석의 목적

1. 상시 측정 수질 항목 중 조류 발생량과 연관이 높은 항목은?
2. 내 몸무게와 연관이 높은 생활습관은?
3. 자동차의 제동거리와 연관 관계가 높은 요소는?

상관계수 (Correlation coefficient)

- 상관계수[2]
  - 두 변수의 관련성 정도.
  - 대표적인 방법
    1. 피어슨 (Pearson) 상관계수
    2. 스피어만 (Spearman) 상관계수
    3. 켄달 (Kendal) 상관계수
       - 일반적으로 상관계수는 피어슨 상관계수를 의미.
  - 상관계수는 -1 ~ 1 사이의 값을 갖는다.
    - 1에 가까울수록 높은 양의 상관관계.
    - -1에 가까울수록 높은 음의 상관관계.
    - 0에 가까울수록 두 변수는 독립적이며 서로 관계가 없다.
  - 상관계수는 인과관계에 대한 어떤 정보도 제공하지 않는다.

[2]출처: R을 활용한 기초 환경자료 분석 및 시각화

피어슨 상관분석의 가정

- 피어슨 상관분석의 기본 가정[3]
  1. 두 변수의 모분포는 정규분포 (정규성)
  2. 두 변수는 직선관계를 이룬다 (선형성)
- 두 변수의 비선형 상관관계 → 스피어만 상관계수
- 쌍을 이루는 두 변수의 상관관계 → 켄달 상관계수

[3]출처: R을 활용한 기초 환경자료 분석 및 시각화

상관계수 검정

- 상관계수 검정[4]
  - 상관계수의 절댓값이 어느 수준에 미쳐야 실질적으로 상관관계에 있는 것인가?
  - 절대적인 상관계수 크기는 존재하지 않음.
  - 상관계수 검정을 통해 상관계수의 통계적 유의성 검정
  - p value가 0.05 이하이면 “상관계수가 0”이라는 귀무가설 기각.
  - R에서는 cor.test를 이용하여 검정.

[4]출처: R을 활용한 기초 환경자료 분석 및 시각화

미세먼지 측정 자료를 이용한 상관관계 분석

- 미세 먼지가 높은 날에는 다른 오염물질도 높을까?
- 미세 먼지와 가장 연관성이 높은 오염물질은 무엇일까?

```
# Read Data
AirKoreaData <- read.csv("./Data/AirQuality.csv", header=T)
# Remove all rows which have at least one "NA"
AirKoreaData <- na.omit(AirKoreaData)
# Calculate correlation between PM10 and PM2.5
# Check normality
# lapply(AirKoreaData[,2:7], FUN=shapiro.test)
```

미세먼지 간 상관관계 분석

```r
# Check correlation between PM10 and PM2.5
#cor(AirKoreaData$PM10, AirKoreaData$PM2_5, method="pearson")
cor.test(AirKoreaData$PM10, AirKoreaData$PM2_5, method="pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  AirKoreaData$PM10 and AirKoreaData$PM2_5
## t = 17.427, df = 325, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6343998 0.7471706
## sample estimates:
##       cor
## 0.6950357
```

미세먼지(PM10)와 이산화질소 간 상관관계 분석

```r
# Check correlation between PM10 and NO2
#cor(AirKoreaData$PM10, AirKoreaData$NO2, method="pearson")
cor.test(AirKoreaData$PM10, AirKoreaData$NO2, method="pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  AirKoreaData$PM10 and AirKoreaData$NO2
## t = 1.8758, df = 325, p-value = 0.06158
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.005024136  0.209596814
## sample estimates:
##       cor
## 0.1034908
```

미세먼지(PM2.5)와 이산화질소 간 상관관계 분석

```r
# Check correlation between PM2.5 and NO2
#cor(AirKoreaData$PM2_5, AirKoreaData$NO2, method="pearson")
cor.test(AirKoreaData$PM2_5, AirKoreaData$NO2, method="pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  AirKoreaData$PM2_5 and AirKoreaData$NO2
## t = 4.614, df = 325, p-value = 5.692e-06
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1433418 0.3470708
## sample estimates:
##       cor
## 0.2479456
```

모든 요소간 상관관계 분석

```r
cor(AirKoreaData[,2:7], method='pearson')
```

```
##              PM10      PM2_5          O3         NO2
## PM10  1.00000000  0.6950357  0.1630690  0.1034908
## PM2_5 0.69503571  1.0000000 -0.1055996  0.2479456
## O3    0.16306901 -0.1055996  1.0000000 -0.5609327
## NO2   0.10349084  0.2479456 -0.5609327  1.0000000
## CO    0.19126634  0.5276100 -0.5737108  0.5973933
## SO4   0.08683864  0.2033191 -0.3118459  0.6250192
##                CO         SO4
## PM10   0.1912663  0.08683864
## PM2_5  0.5276100  0.20331914
## O3    -0.5737108 -0.31184589
## NO2    0.5973933  0.62501918
## CO     1.0000000  0.48658100
## SO4    0.4865810  1.00000000
```

PM2.5와 일산화탄소 간 상관관계 분석

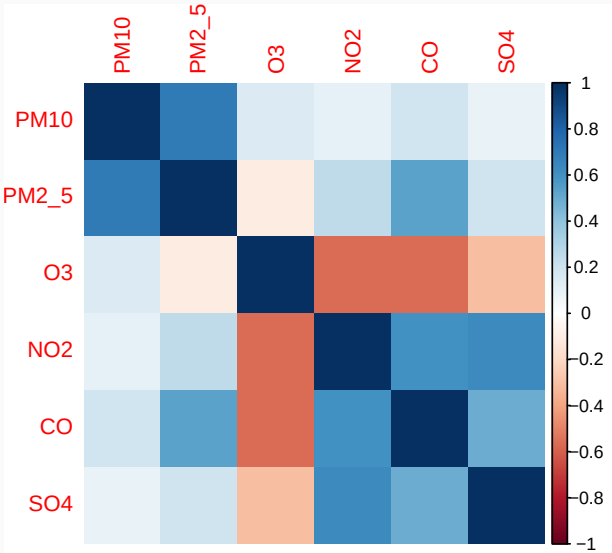
```r
cor.test(AirKoreaData$PM2_5, AirKoreaData$CO, method='pearson')
```

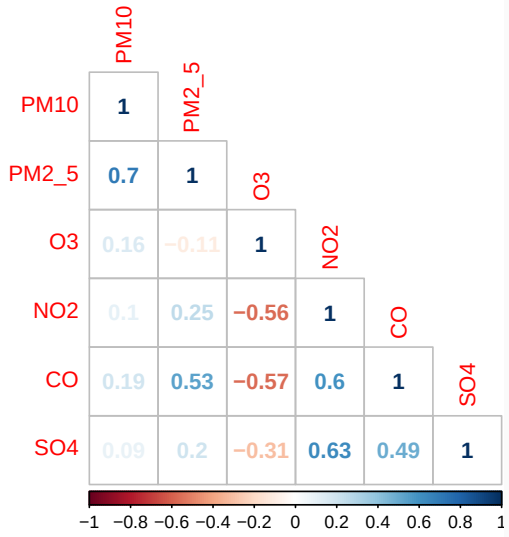
```
##
##  Pearson's product-moment correlation
##
## data:  AirKoreaData$PM2_5 and AirKoreaData$CO
## t = 11.197, df = 325, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4445927 0.6016404
## sample estimates:
##      cor
## 0.52761
```

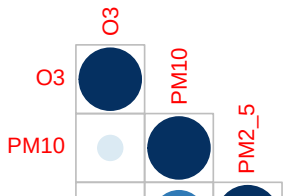
```r
# Check significance of correlations between PM2.5 and others, statis
```

corrplot 패키지를 이용한 시각화.

```
library(corrplot)
par(mfrow=c(2,2))
cor <- round(cor(AirKoreaData[,2:7], method="pearson"), 2)
corrplot(cor, method="shade", shade.col=NA)
corrplot(cor, method="number", type="lower")
corrplot(cor, method="circle", type="lower", order="hclust")
```







회귀분석

회귀분석 (Regression analysis)

- 회귀 (regression)[5]
  - 유전학자 프란시스 골턴의 유전 법칙 연구 중 나온 명칭
  - 한 세대의 유전적 형질은 그 세대의 평균으로 접근한다.
  - “평균으로의 회귀 (Regression toward mean)”
- 회귀분석
  - 한 개 또는 그 이상의 변수 (독립변수)에 대한 다른 변수 (종속변수) 사이의 관계.
  - 변수간 관계를 수학적 모형으로 산출
  - 상관분석과 다르게 변수간의 관계를 이용하여 무엇인가를 예측.
  - 단순회귀분석/다중회귀분석
  - 선형회귀분석/비선형회귀분석
  - 인과 관계를 나타내지는 않는다.
- R에서는 `lm` 함수를 이용하여 회귀분석이 가능하다.

[5]출처: R을 활용한 기초 환경자료 분석 및 시각화

회귀분석의 목적

1. 조류 발생량과 상관관계가 높은 수질 항목들로 수학적 함수를 만들 수 있을까?
2. 만들 수 있다면 향후 조류 발생량을 예측 가능한가?
3. 미세먼지에 영향을 주는 환경 요소들은 무엇일까?
4. 환경 요소들을 어떻게 조절하면 우리가 원하는 수준의 미세먼지 농도로 낮출 수 있을까?
5. 자동차 제동거리와 상관관계가 높은 항목들을 이용하여 차량 간 안전거리를 제시할 수 있을까?

회귀분석의 기본 가정

- 회귀식의 구조
  - $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i$
- 회귀분석의 기본 가정
  - 독립변수와 종속변수 간에는 통계적으로 유의한 상관관계가 존재한다.
  - 잔차 (residuals)의 평균은 0이며, 분산은 σ^2인 정규분포를 따라야 한다.
  - 잔차는 모든 i에 대해 평균과 분산이 일정하다.
  - 다중 회귀분석의 경우 각 독립변수 간에는 다중공선성 (multicollinearity)가 적어야 한다.

선형 회귀 분석과 최소 자승법 (Least square method)

- 선형 회귀는 수학적으로 잔차 즉 오차의 합이 가장 작게 만드는 식을 선택하는 것이다.
  - $\sum (Y_i - \beta_0 - \beta_1 X_i)^2$
  - RSS: Residual sum of squares (잔차 제곱합)

자동차의 주행 속도에 따른 제동 거리 예측 (데이터 호출)

```r
# Use "cars" data embedded in R
data(cars)
# Check data using head function.
head(cars)
```

```
##   speed dist
## 1     4    2
## 2     4   10
## 3     7    4
## 4     7   22
## 5     8   16
## 6     9   10
```

```r
# Data has two columns of "speed" and "dist".
```

자동차의 주행 속도에 따른 제동 거리 예측 (모형 작성 및 평가)

```r
# Create the model
TestModel <- lm(dist ~ speed, data = cars)
summary(TestModel)
```

```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791     6.7584  -2.601   0.0123 *
## speed         3.9324     0.4155   9.464 1.49e-12 ***
## ---
```

회귀분석과 결정계수

- 결정 계수 (Coefficient of determination: R^2)
  - Y_i의 총변동에 대비해 회귀모형이 얼마나 그 변동을 설명하는지 알려준다 (설명력)
  - 총제곱합 (SST: Total sum of squares) = $\sum(Y_i - \bar{Y})^2$
  - 회귀제곱합 (SSR: Sum of squares due to Regression) = $\sum(\hat{Y}_i - \bar{Y})^2$
  - $R^2 = \frac{SSR}{SST}$
  - 결정계수는 독립 변수 숫자가 커짐에 따라 늘어나는 경향이 있다.
  - $R^2_{adjusted} = 1 - \frac{RSS/(n-k-1)}{SST/(n-1)}$

자동차의 주행 속도에 따른 제동 거리 예측 (최종 모형)

- 위의 모형에 따른 주행 속도와 제동 거리간의 관계식
  - $dist = -17.759 + 3.932 \times speed + \epsilon$

자동차의 주행 속도에 따른 제동 거리 예측 (모형 확인)

```
# Check regression coefficient
coef(TestModel)
```

```
## (Intercept)       speed
##  -17.579095    3.932409
```

```
# Check the fitted values for each predictor values.
fitted(TestModel)[1:5]
```

```
##         1         2         3         4         5
## -1.849460 -1.849460  9.947766  9.947766 13.880175
```

```
# Check residuals of the model for each predictor values.
residuals(TestModel)[1:5]
```

```
##         1         2         3         4         5
##  3.849460 11.849460 -5.947766 12.052234  2.119825
```

자동차의 주행 속도에 따른 제동 거리 예측 (모형 확인2)

```r
# fitted value + residual = measured distance
(fitted(TestModel) + residuals(TestModel))[1:5] == cars$dist[1:5]
```

```
##    1    2    3    4    5
## TRUE TRUE TRUE TRUE TRUE
```

```r
# Check confidential interval of the model
confint(TestModel)
```

```
##                  2.5 %    97.5 %
## (Intercept) -31.167850 -3.990340
## speed         3.096964  4.767853
```

```r
# RSS of the model
deviance(TestModel)
```

```
## [1] 11353.52
```

자동차의 주행 속도에 따른 제동 거리 예측 (예측)

- 회귀식을 통한 예측 (predict)

```
# prediction without consideration of residuals
predict(TestModel, newdata=data.frame(speed=c(seq(50,100,10))),
        interval="confidence")
```

```
##        fit      lwr      upr
## 1 179.0413 149.8060 208.2766
## 2 218.3654 180.8489 255.8820
## 3 257.6895 211.8651 303.5139
## 4 297.0136 242.8670 351.1602
## 5 336.3377 273.8603 398.8151
## 6 375.6618 304.8480 446.4755
```

자동차의 주행 속도에 따른 제동 거리 예측 (예측)

```
# prediction considering residuals
predict(TestModel, newdata=data.frame(speed=c(seq(50,100,10))),
        interval="prediction")
```

```
##        fit      lwr      upr
## 1 179.0413 136.4865 221.5962
## 2 218.3654 169.7474 266.9834
## 3 257.6895 202.4076 312.9715
## 4 297.0136 234.6592 359.3680
## 5 336.3377 266.6266 406.0488
## 6 375.6618 298.3908 452.9328
```

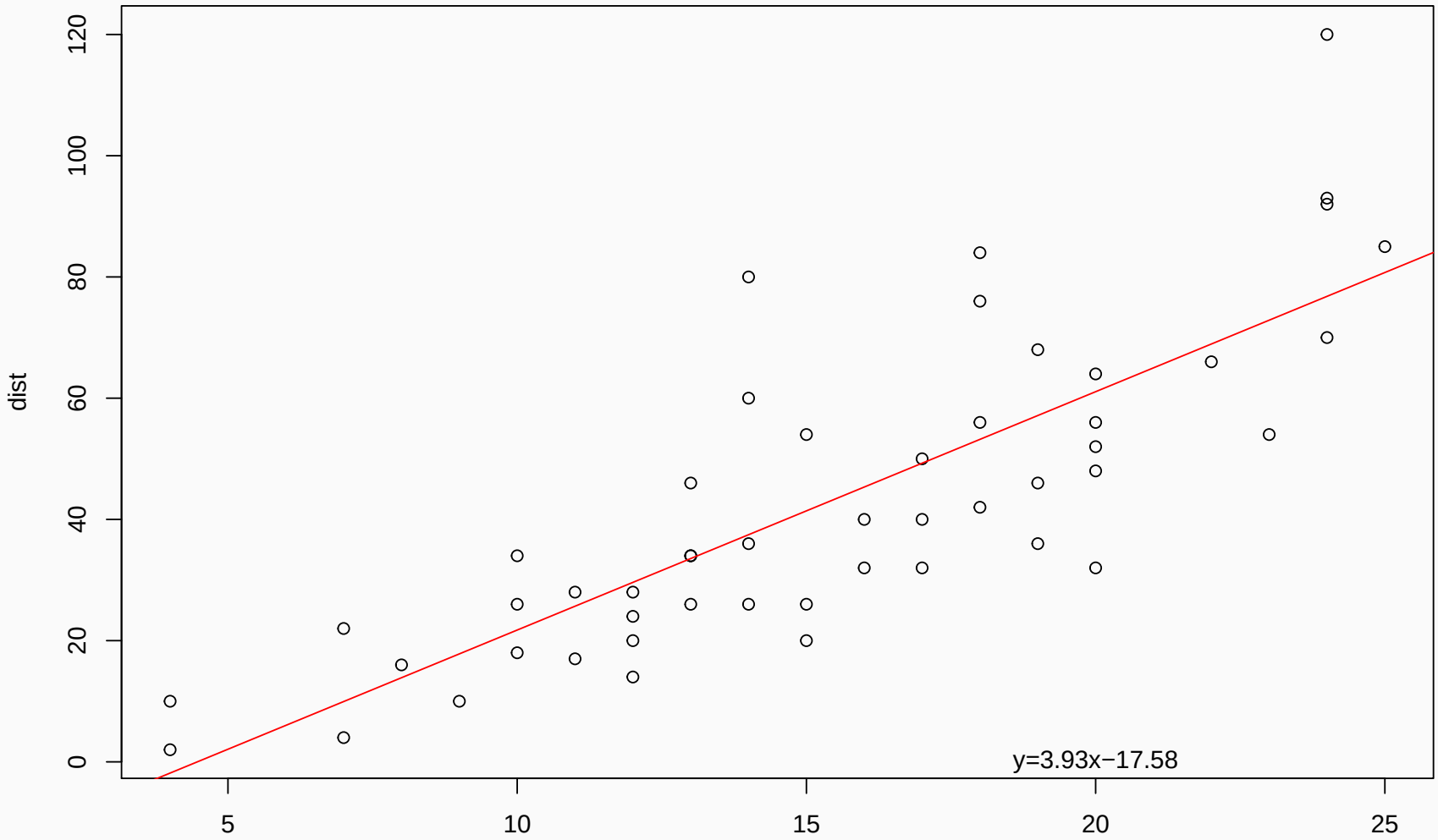
모형 가정 평가 및 이상치 확인

```r
# Check using plot of the model
# plot(TestModel)
# Check outliers using "car" package
# install.packages("car")
library(car)
outlierTest(TestModel)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##    rstudent unadjusted p-value Bonferroni p
## 49 3.184993          0.0025707      0.12853
```

시각화

```r
plot(dist ~ speed, data = cars) # Draw scatter plot
abline(TestModel, col="red") # add regression line
text(x=20, y=0, 'y=3.93x-17.58') # add text
```



시각화

```
library(ggplot2)
TestPlot <-
    ggplot(cars, aes(x=speed, y=dist)) +
    geom_point() +
    labs(x="Speed", y="Distance") +
    geom_smooth(method=lm) +
    geom_text(x=20, y=0, label = 'y = 3.93x - 17.58')
```

시각화

TestPlot