1. 회귀분석

A. 회귀분석

1) 독립변수가 한단위 증가할 때 종속변수에 미치는 영향을 측정하기 위한 통계적 예측 모형

① 단순선형회귀분석 – 1개의 독립변수를 사용

② 다중회귀분석 – 여러개의 독립변수를 사용

2) 회귀분석 프로세스

- 분석을 위한 주제 결정

  ex) 교육시간이 직원의 업무 수행에 영향을 주는가?

  식사시간이 아이의 두뇌발달에 영향을 주는가?

- 독립변수와 종속변수 선정

  독립변수 : 교육시간

  종속변수 : 업무능력

- 가설 설정

  귀무가설(H0) : 교육시간이 업무 능력 점수에 영향을 주지 않는다.

  대립가설(H1) : 교육시간이 업무 능력 점수에 영향을 준다.

- 데이터 수집
- 데이터 중에서 특이하거나 이상한 데이터의 제거
- 모델을 적용하여 데이터 분석
- 결과 해석

  p-value가 0.05보다 작으면 대립가설(H1) 채택

  결정계수가 0~1 사이의 값을 가지며 0.65~0.7 이상이어야 좋은 회귀모형이라고 할 수 있음

B. 데이터 전처리

1) 결측값 처리

```
df<-read.csv("d:/data/ozone/ozone.csv")
head(df)
#결측값 여부 확인
is.na(df)

#특정 필드의 결측값 확인
is.na(df$Ozone)

#Ozone 필드에 결측값이 있는 행
df[is.na(df$Ozone),]

#결측값의 개수
sum(is.na(df))

#특정 필드의 결측값 개수
sum(is.na(df$Ozone))

#각 샘플의 모든 필드가 NA가 아닐 때 TRUE
#샘플에 결측값이 하나라도 있으면 FALSE
complete.cases(df)

#결측값이 없는 샘플 출력
df[complete.cases(df),]

#결측값이 있는 샘플 출력
df[!complete.cases(df),]
```

```
#결측값이 있으므로 계산이 안됨
mean(df$Ozone)

#결측값을 제외하고 계산
mean(df$Ozone, na.rm=T)
#1~2번 필드의 중위수 계산
mapply(median, df[1:2], na.rm=T)

#결측값을 제외
df2<-na.omit(df)
df2

#결측값을 0으로 대체
df3<-df
df3[is.na(df)]<-0
df3

#특정한 필드만 0으로 대체
df4<-df
df4$Ozone[is.na(df4$Ozone)]<-0
df4

#결측값을 평균값으로 대체
df5<-df

m1<-mean(df[,1], na.rm=T)
m2<-mean(df[,2], na.rm=T)
df5[,1][is.na(df[,1])]<-m1
df5[,2][is.na(df[,2])]<-m2
df5
```

```
# 결측값 시각화 패키지
#install.packages('VIM')
#install.packages('mice')
library(VIM)
library(mice)

win.graph()
md.pattern(df)
#결측값이 없는 샘플 111개
#Ozone 필드에만 결측값이 있는 샘플 35개
#Solar.R 필드에만 결측값이 있는 샘플 5개
#2개 필드에 결측값이 있는 샘플 2개

## 결측값의 개수 표시
win.graph()
#prop=T 백분율로 표시, prop=F 샘플개수로 표시
aggr(df, prop = F, numbers = T)

# 결측값의 위치를 시각적으로 표현(red: 결측값, dark: 빈도수가 높은 값)
win.graph()
matrixplot(df)
```

2) 스케일링

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

#범주형 변수는 팩터 자료형으로 변환 후 스케일링 수행
df$weekend <- as.factor(df$weekend)
df$weekend

#install.packages("reshape")
library(reshape)
# melt() 필드 1개를 variable,value 로 여러 행으로 만드는 함수 (차원변경)
meltData <- melt(df[2:7])
win.graph()
boxplot(data=meltData, value~variable)

#평균 0, 표준편차 1로 만드는 작업
#스케일링: 표준편차를 1로 만드는 작업
#센터링: 평균을 0으로 만드는 작업
# 정규화된 데이터를 data.frame형태로 변경
df_scaled <- as.data.frame(scale(df[2:7])) #스케일링과 센터링
df_scaled

meltData <- melt(df_scaled)
win.graph()
boxplot(data=meltData, value~variable)
```

```
#caret 패키지(Classification And Regression Training):분류, 회귀 문제를 풀기 위한 다양한 도구 제공
#install.packages('caret')
library(caret)

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

meltData <- melt(df[2:7])
win.graph()
boxplot(data=meltData, value~variable)

#평균 0, 표준편차 1로 스케일링
prep <- preProcess(df[2:7], c("center", "scale"))
df_scaled2 <- predict(prep, df[2:7])
head(df_scaled2)

meltData <- melt(df_scaled2)
win.graph()
boxplot(data=meltData, value~variable)

#range: 0~1 정규화
prep <- preProcess(df[2:7], c("range"))
df_scaled3 <- predict(prep, df[2:7])
head(df_scaled3)

meltData <- melt(df_scaled3)
win.graph()
boxplot(data=meltData, value~variable)
```

3) 이상치 처리

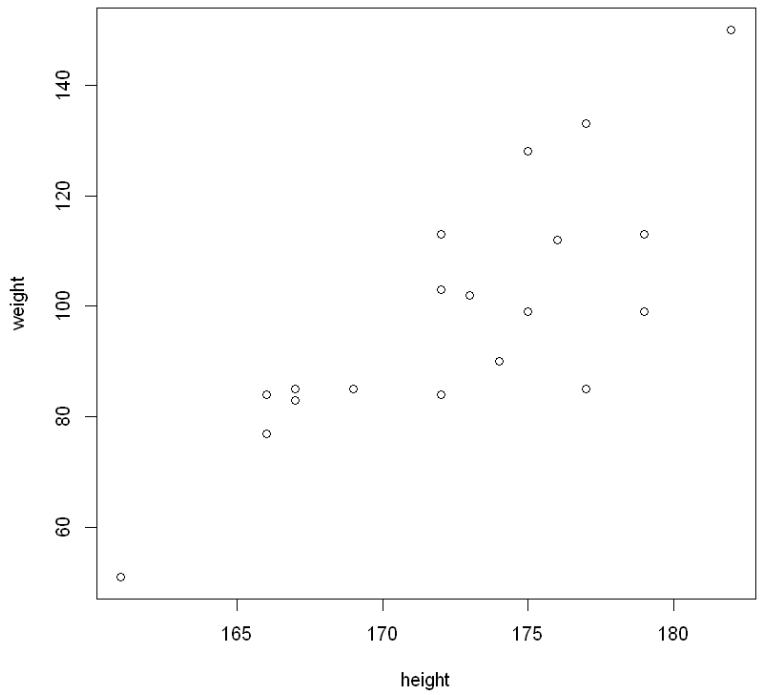
```
df<-read.csv("d:/data/rides/rides.csv")
head(df)
#install.packages('car')
library(car)
#회귀분석 모형
model<-lm(overall~num.child + distance + rides + games + wait + clean, data=df)
summary(model)
#설명력 68.27%

# 1. 아웃라이어
# 잔차가 2배 이상 크거나 2배 이하로 작은 경우
outlierTest(model)
# 이상치 데이터 발견 - 184번 샘플(Bonferonni p value가 0.05보다 작은 값)
# rstudent - Studentized Residual - 잔차를 잔차의 표준편차로 나눈 값
# unadjusted p-value : 다중 비교 문제가 있는 p-value
# 본페로니 p - 여러 개의 가설 검정을 수행할 때 다중 비교 문제로 인해 귀무가설을 기각하게 될
#   확률이 높아지는 문제를 교정한 p-value

#184번 샘플을 제거한 모형
model2<-lm(overall~num.child + distance + rides + games + wait + clean, data=df[-184,])
model2
summary(model2)
#설명력이 68.27% => 68.76%로 개선됨
```

```
#2. 영향 관측치(influential observation) : 모형의 인수들에 불균형한 영향을 미치는 관측치
# 영향 관측치를 제거하면 더 좋은 모형이 될 수 있음
# Cook's distance(레버리지와 잔차의 크기를 종합하여 영향력을 판단하는 지표)를 이용하여
#   영향 관측치를 찾을 수 있음
# 레버리지(leverage) : 실제값이 예측값에 미치는 영향을 나타낸 값
# x축: Hat-Values(큰 값은 지렛점)
# y축: Studentized Residuals(표준화 잔차) : 잔차를 표준오차로 나눈 값
win.graph()
influencePlot(model)
#184,103,227,367,373
#  2보다 큰 값, -2보다 작은 값들은 2배 이상 떨어져있는 이상치)
#레버리지와 잔차의 크기가 모두 큰 데이터들은 큰 원으로 표현(영향력이 큰 데이터들)
#184,103,227,367,373

model3=lm(overall~num.child + distance + rides + games + wait + clean, data=df[c(-184,-103,-367,-373),])
model3
summary(model3)
# 설명력 69.12%
```

C. 단순회귀분석

1) 실습1

```
#20명의 신장과 체중 데이터
height <- c(179,166,175,172,173,167,169,172,172,179,161,17
4,166,176,182,175,177,167,176,177)
weight <- c(113,84,99,103,102,83,85,113,84,99,51,90,77,112,
150,128,133,85,112,85)
plot(height,weight)
```



```
#상관계수 계산
cor(height,weight)
```

```
0.800235051021387
```

```
#기울기와 절편
slope <- cor(height, weight) * (sd(weight) / sd(height))
intercept <- mean(weight) - (slope * mean(height))
slope
intercept
```

```
#단순회귀분석 모델 생성
#체중 = 기울기x신장 + 절편
df <- data.frame(height, weight)
df

model <- lm(weight ~ height, data=df)
#절편(Intercept) -478.816
#기울기 3.347
model
```

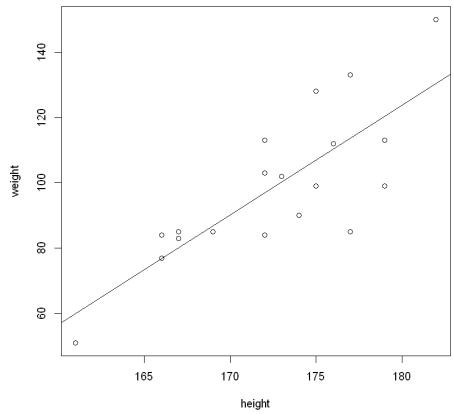
```
Call:
lm(formula = weight ~ height)

Coefficients:
(Intercept)       height
   -478.816        3.347
```

```
#키가 180인 사람의 체중 예측
model$coefficients[[2]]*180 + model$coefficients[[1]]
```

```
123.666666666667
```

```
summary(model)
```

```
plot(height,weight)
abline(model,col='red')
```




```
weight

pred<-model$coefficients[[2]]*height + model$coefficients
[[1]]
pred

sum(weight-pred) #오차의 합계는 0

err<-(weight-pred)^2

sum(err) #오차의 제곱합

sum(err/length(weight)) #평균제곱오차(MSE, mean squared erro
r)
#비용함수(cost function) : 평균제곱오차를 구하는 함수
```

```
#최적의 가중치(기울기)를 구하기 위한 계산(경사하강법, Gradient Descent)
#여기서는 전체의 값이 아닌 1개의 값만 계산
x<-height[1]
y<-weight[1]
w<-seq(-1,2.3,by=0.0001) #가중치, by 간격
#w<-seq(-1,2.3,by=0.1) #가중치, by 간격
pred<-x*w #예측값
err<-(y-pred)^2 #제곱오차
plot(err)
#기울기가 증가하면 오차가 증가하고 기울기가 감소하면 오차가 감소한다
#기울기가 0에 가까운 값이 최적의 기울기가 된다.
min(err) #최소오차
i<-which.min(err)
paste('최적의 기울기=',w[i])
```

```
#최적의 편향(절편)을 구하기 위한 계산
x<-height[1]
y<-weight[1]
w<-0.6313 #가중치
b<-seq(-3.2,3.2,by=0.0001) #편향
#b<-seq(-1,3.2,by=0.1) #편향
pred<-x*w + b #예측값
err<-(y-pred)^2 #제곱오차
plot(err)
#기울기가 증가하면 오차가 증가하고 기울기가 감소하면 오차가 감소한다
#기울기가 0에 가까운 값이 최적의 기울기가 된다.
min(err) #최소오차
i<-which.min(err)
i
paste('최적의 편향=',b[i])
```

```
#위의 계산을 통해 얻은 최적의 w,b를 적용한 회귀식
x<-height[1]
y<-weight[1]
w<- 0.6313
b<- -0.00269999999999992
pred<-x*w + b
y
pred
```

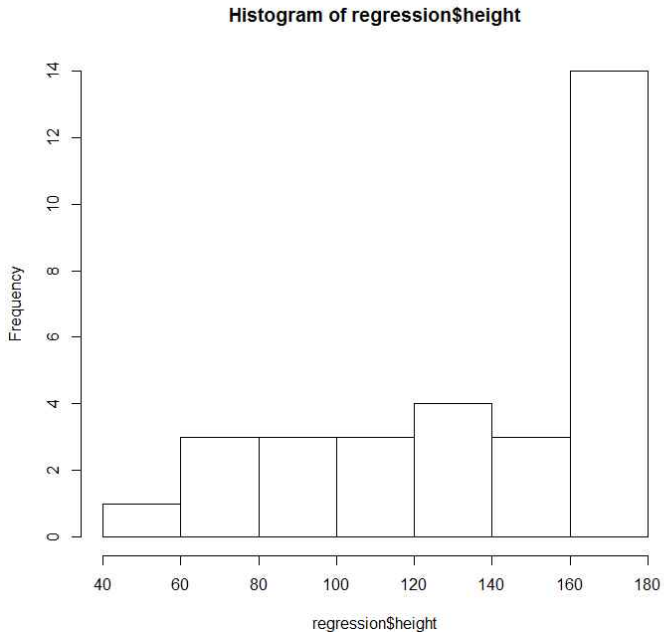
2) 단순회귀분석 실습2

```
regression<-read.csv("d:/data/regression/regression.csv",
fileEncoding='utf-8')
head(regression)
tail(regression)
```

```
          age height weight
1   0~3개월   59.1    5.9
2   3~6개월   66.7    8.0
3   6~9개월   71.4    8.9
4  9~12개월   75.0   10.1
5 12~18개월   80.1   10.9
6       2세   87.8   13.2

          age height weight
26  30~34세  171.3   71.5
27  35~39세  170.7   72.3
28  40~49세  168.6   70.6
29  50~59세  166.1   69.1
30  60~69세  164.4   65.9
31 70세이상  162.4   61.1
```

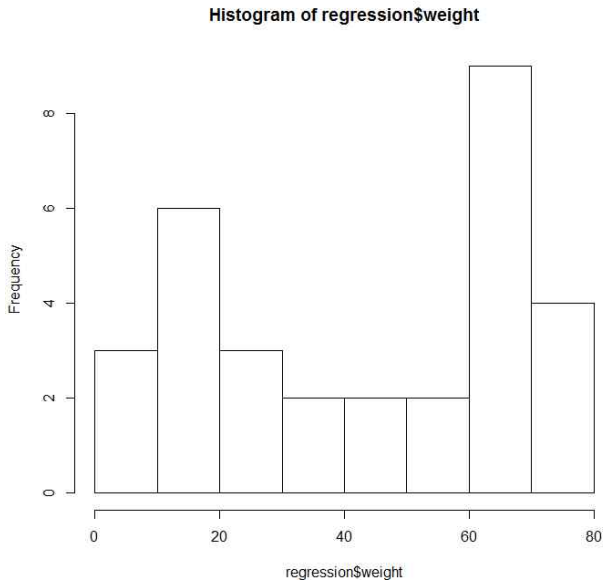
```
summary(regression)
hist(regression$height)
```

```
      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     59.1   105.7   150.7   135.8   169.8   173.8

      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     5.90   17.90   45.20   42.62   66.65   72.30
```



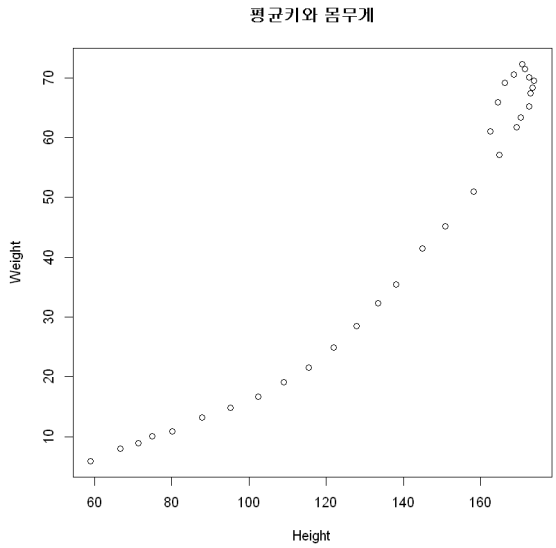
신장의 값은 **160~180**에 집중되어 있다.

```
hist(regression$weight)
```



체중의 값은 60~70, 10~20에 많이 분포되어 있다.

```
plot(regression$weight ~ regression$height, main="평균키와 몸무게", xlab="Height", ylab="Weight")
```



상관계수를 구함

-1에서 1까지의 값을 가짐, -1에 가까울수록 음의 상관관계, 1에 가까울수록 양의 상관관계를 나타냄. 0에 가까우면 두 변수는 관계가 없다는 의미

```
cor(regression$height, regression$weight)
```

```
[1] 0.9672103
```

* 키와 몸무게의 관계

독립변수 : 신장

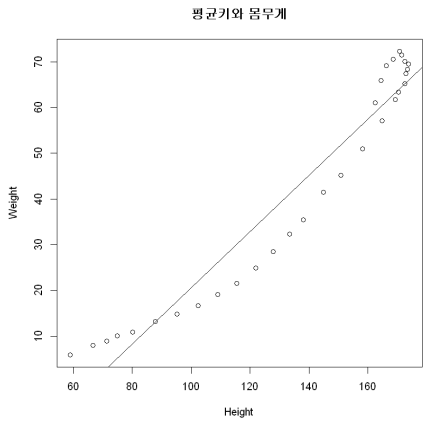
종속변수 : 체중

귀무가설 : 신장은 체중에 영향을 주지 않을 것이다.

대립가설 : 신장은 체중에 영향을 줄 것이다.

```
# lm( y ~ x ) x 독립변수, y 종속변수 (x가 한단위 증가할 때 y에게 미치는 영향)
r <- lm(regression$weight ~ regression$height)
plot(regression$weight ~ regression$height, main="평균키와 몸무게", xlab="Height", ylab="Weight")
abline(r,col='red')
```

회귀선 : 각 관측값들과 평균의 차의 제곱이 최소가 되는 직선



```
#키가 180인 사람의 체중 예측
r$coefficients[[2]]*180 + r$coefficients[[1]]
```

분석결과 요약

```
summary(r)
```

```
Call:
lm(formula = regression$height ~ regression$weight)

Residuals:
     Min       1Q   Median       3Q      Max
-20.8266  -7.9450  -0.6153   9.3139  13.4815

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       70.9481     3.6366   19.51   <2e-16 ***
regression$weight  1.5218     0.0742   20.51   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1

Residual standard error: 9.999 on 29 degrees of freedom
Multiple R-squared:  0.9355, Adjusted R-squared:  0.9333
F-statistic: 420.6 on 1 and 29 DF,  p-value: < 2.2e-16
```

#결과 해석 : R 제곱값이 0.9333이므로 93%의 설명력을 가진다. 또한 p-value가 0.05보다 작으므로 회귀분석 결과가 통계적으로 유의하다.
따라서 귀무가설을 기각하고 대립가설을 채택한다.
결론 : 키와 몸무게는 상관관계가 있다.

Call : 회귀분석에 사용된 모델 식

Residuals: 잔차, 회귀선의 값과 실제 관측 값의 차이를 각 분위수로 표시

Coefficients: 절편, 독립변수 등에 대한 회귀계수

Residual standard error: 잔차의 표준오차와 자유도

Multiple R-squared: 결정계수, 즉 추정된 회귀선이 실제 관측값을 얼마나 잘 설명하는가를 의미하는 값. 0에서 1사이의 값을 가지며 1은 실제관측 값들이 회귀선 상에 위치함을 의미함

Adjusted R-squared: 수정결정계수, 변수가 많아지면 R제곱이 무조건 높아지는 단점을 보완한 것, R제곱과 큰 차이가 나지 않을수록 좋은 모형

F-statistic: F통계량은 해당 모형이 의미가 있는지 아닌지를 알려줌. 계수 중 하나라도 0이 아닌 것이 있다면 그 모형은 유의미하다고 판단함.

3) 단순회귀분석(전기생산량과 전기사용량)

```
#월별 전기생산금액(억원)
X <- c(3.52, 2.58, 3.31, 4.07, 4.62, 3.98, 4.29, 4.83, 3.71,
4.61, 3.90, 3.20)
#월별 전기 사용량(백만kwh)
y <- c(2.48, 2.27, 2.47, 2.77, 2.98, 3.05, 3.18, 3.46, 3.03,
3.25, 2.67, 2.53)
plot(X,y)
```

```
#상관계수 계산
cor(X,y)
```

```
#단순회귀분석 모델 생성
#전기소비량 = 기울기x전기생산량 + 절편
model <- lm(y ~ X)
#절편(Intercept) 0.9196
#기울기 0.4956
model
```

```
summary(model)
#Adjusted R-squared:  0.777 (모형의 설명력 77.7%)
#p-value가 0.05보다 작으므로 통계적으로 유의한 회귀모형
```

```
#산점도
plot(X, y, main="전기생산량과 전기소비량", xlab="전기생산량",
ylab="전기소비량")
#회귀선
abline(model,col="red")
```

```
#전기생산량이 4일 때의 전기소비량 예측
model$coefficients[[2]]*4 + model$coefficients[[1]]
```

4) 단순회귀분석(오존량 예측)

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

```
# 결측값이 있는 행을 제거
df<-na.omit(df)
tail(df)
```

```
X<-df$Temp
y<-df$Ozone
X
y
```

```
#상관계수 계산
cor(X,y)
```

```
#단순회귀분석 모델 생성
#오존량 = 기울기x온도 + 절편
model <- lm(y ~ X)
#절편(Intercept) -147.646
#기울기 2.439
model
```

```
summary(model)
#Adjusted R-squared:  0.483 (모형의 설명력 48.3%)
#p-value가 0.05보다 작으므로 통계적으로 유의한 회귀모형
```

```
#산점도
plot(X, y, main="온도와 오존량", xlab="온도", ylab="오존량")
#회귀선
abline(model,col="red")
```

```
#온도가 화씨 80도일 때 오존량 예측
model$coefficients[[2]]*80 + model$coefficients[[1]]
```

5) 단순회귀분석(붓꽃품종)

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

```
#꽃받침의 너비와 꽃받침의 길이와의 관계
X<-df$SepalWidth
y<-df$SepalLength
X
y
```

```
#상관계수 계산
cor(X,y)
```

```
#단순회귀분석 모델 생성
#꽃받침길이 = 기울기x꽃받침너비 + 절편
model <- lm(y ~ X)
#절편(Intercept) 6.4812
#기울기 -0.2089
model
```

```
summary(model)
#Adjusted R-squared:  0.005286 (모형의 설명력 0.5%)
#p-value가 0.05보다 크므로 통계적으로 유의하지 않음
#결론: 꽃받침의 너비와 꽃받침의 길이는 상관관계가 없다.
```

```
#산점도
plot(X, y, xlab="SepalWidth", ylab="SepalLength")
#회귀선
abline(model,col="red")
```

```
#꽃잎의 너비와 꽃잎의 길이와의 관계
X<-df$PetalWidth
y<-df$PetalLength
X
y
```

```
#상관계수 계산
cor(X,y)
```

```
#단순회귀분석 모델 생성
#꽃잎길이 = 기울기x꽃잎너비 + 절편
model <- lm(y ~ X)
#절편(Intercept) 1.091
#기울기 2.226
model
```

```
summary(model)
#Adjusted R-squared:  0.9264 (모형의 설명력 92.6%)
#p-value가 0.05보다 작으므로 통계적으로 유의함
#결론: 꽃잎의 너비와 꽃잎의 길이는 상관관계가 있다.
```

```
#산점도
plot(X, y, xlab="PetalWidth", ylab="PetalLength")
#회귀선
abline(model,col="red")
```

6) 단순회귀분석(항공운항데이터)

```
#항공운항 데이터셋
#분석할 필드가 적은 편
#로딩 시간이 많이 걸려서 가장 레코드수가 적은 1987년 데이터로 실습
df<-read.csv("d:/data/airline/1987.csv")
head(df)
tail(df)
```

```
library(dplyr)

df<-df %>% select(Distance, DepDelay, ArrDelay)
```

```
dim(df)
```

```
#install.packages("Hmisc")
library(Hmisc)
describe(df)
```

```
# 결측값이 있는 행을 제거
df<-na.omit(df)
tail(df)
```

```
dim(df)
```

```
#운항거리와 출발지연시간과의 관계
X<-df$Distance
y<-df$DepDelay
head(X)
head(y)
```

```
#상관계수 계산
cor(X,y)
```

```
#단순회귀분석 모델 생성
#출발지연시간 = 기울기x운항거리 + 절편
model <- lm(y ~ X)
#절편(Intercept) 6.423342
#기울기 0.002612
model
```

```
summary(model)
#Adjusted R-squared:  0.003031 (모형의 설명력)
#p-value가 0.05보다 작으므로 통계적으로 유의함
```

```
#오래걸림
#산점도
plot(X, y, xlab="운항거리", ylab="출발지연시간")
#회귀선
abline(model,col="red")
```

```
#운항거리와 도착지연시간과의 관계
X<-df$Distance
y<-df$ArrDelay
head(X)
head(y)
```

```
#상관계수 계산
cor(X,y)
```

```
#단순회귀분석 모델 생성
#도착지연시간 = 기울기x운항거리 + 절편
model <- lm(y ~ X)
#절편(Intercept) 8.200779
#기울기 0.002105
model
```

```
summary(model)
#Adjusted R-squared:  0.00165 (모형의 설명력)
#p-value가 0.05보다 작으므로 통계적으로 유의함
```

```
#오래걸림
#산점도
plot(X, y, xlab="운항거리", ylab="도착지연시간")
#회귀선
abline(model,col="red")
```

7) 단순회귀분석(와인품질)

```
#와인데이터셋
#https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/
df<-read.csv("d:/data/wine/winequality-red.csv",sep=";")
head(df)
tail(df)
```

```
dim(df)
```

```
library(Hmisc)
describe(df)
```

```
#alcohol과 pH과의 관계
X<-df$alcohol
y<-df$pH
head(X)
head(y)
```

```
#상관계수 계산
cor(X,y)
```

```
#단순회귀분석 모델 생성
#pH = 기울기x알코올 + 절편
model <- lm(y ~ X)
model
```

```
summary(model)
#Adjusted R-squared:
```

```
#오래걸림
#산점도
plot(X, y, xlab="알코올", ylab="pH")
#회귀선
abline(model,col="red")
```

D. 다중회귀분석 실습

1) attitude

30개 부서에서 부서당 35명의 직원들을 대상으로 설문조사
독립변수가 2개 이상인 경우

1.분석의 대상이 되는 모든 독립변수를 넣고 회귀식을 구성
2.기여도가 낮은 변수부터 하나씩 제거
3.최종적으로 종속 변수에 대한 기여도가 높은 변수들로 구성된 회귀식을 완성

```
#R에 기본적으로 포함되는 데이터셋 목록
data()
#데이터셋에 대한 도움말
#help(데이터셋이름)

head(attitude)
tail(attitude)
```

다중회귀분석 모델 생성

```
model<-lm(rating ~ . , data=attitude)
model
```

분석결과 요약

```
summary(model)
#complaints, learning이 기여도가 높은 변수
#p-value가 0.05보다 작으므로 통계적으로 유의함
#모델의 설명력(예측의 정확성) 66%
```

```
Call:
lm(formula = rating ~ ., data = attitude)

Residuals:
     Min       1Q   Median       3Q      Max
-10.9418  -4.3555   0.3158   5.5425  11.5990

Coefficients: => 각 항목별 평가치
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.78708   11.58926   0.931 0.361634
complaints   0.61319    0.16098   3.809 0.000903 ***
privileges  -0.07305    0.13572  -0.538 0.595594
learning     0.32033    0.16852   1.901 0.069925 .
raises       0.08173    0.22148   0.369 0.715480
critical     0.03838    0.14700   0.261 0.796334
advance     -0.21706    0.17821  -1.218 0.235577
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1

Residual standard error: 7.068 on 23 degrees of freedom
Multiple R-squared:  0.7326, Adjusted R-squared:  0.6628
F-statistic:  10.5 on 6 and 23 DF,  p-value: 1.24e-05
```

p-value가 0.05보다 작으므로 통계적으로 유의함

예측의 정확성은 66%

Coefficients: 를 보면 complaints와 learning 항목만이 유의미한 것으로 평가됨

*표가 많을수록 유의하다.

```
#기여도가 낮은 항목을 제거함으로써 의미있는 회귀식을 구성하는 과정
reduced<-step(model, direction="backward")
#최종적으로 complaints와 learning 2가지 변수 외에는 제거됨
```

```
Start:  AIC=123.36
rating ~ complaints + privileges + learning + raises +
critical +
    advance => 모든 변수

             Df Sum of Sq    RSS    AIC
- critical    1      3.41 1152.4 121.45
- raises      1      6.80 1155.8 121.54
- privileges  1     14.47 1163.5 121.74
- advance     1     74.11 1223.1 123.24
<none>                    1149.0 123.36
- learning    1    180.50 1329.5 125.74
- complaints  1    724.80 1873.8 136.04

Step:  AIC=121.45
rating ~ complaints + privileges + learning + raises +
advance => critical이 제거됨
             Df Sum of Sq    RSS    AIC
- raises      1     10.61 1163.0 119.73
```

```
- privileges  1     14.16 1166.6 119.82
- advance     1     71.27 1223.7 121.25
<none>                    1152.4 121.45
- learning    1    177.74 1330.1 123.75
- complaints  1    724.70 1877.1 134.09

Step:  AIC=119.73
rating ~ complaints + privileges + learning + advance

            Df Sum of Sq    RSS    AIC
- privileges  1     16.10 1179.1 118.14
- advance     1     61.60 1224.6 119.28
<none>                    1163.0 119.73
- learning    1    197.03 1360.0 122.42
- complaints  1   1165.94 2328.9 138.56

Step:  AIC=118.14
rating ~ complaints + learning + advance

            Df Sum of Sq    RSS    AIC
- advance     1     75.54 1254.7 118.00
<none>                    1179.1 118.14
- learning    1    186.12 1365.2 120.54
- complaints  1   1259.91 2439.0 137.94

Step:  AIC=118
rating ~ complaints + learning => 최종적으로 complaints와
learning 2가지 변수 외에는 제거되었음
            Df Sum of Sq    RSS    AIC
<none>                    1254.7 118.00
- learning    1    114.73 1369.4 118.63
- complaints  1   1370.91 2625.6 138.16
```

최종 결과 확인

```
summary(reduced)

#p-value가 0.05보다 작으므로 이 회귀모델은 통계적으로 유의함.
#모델의 설명력(신뢰도,예측정확성) : 68%
```

```
Call:
lm(formula = rating ~ complaints + learning, data =
attitude)
s
Residuals:
     Min       1Q   Median       3Q      Max
-11.5568  -5.7331   0.6701   6.5341  10.3610

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.8709     7.0612   1.398    0.174
complaints    0.6435     0.1185   5.432 9.57e-06 ***
learning      0.2112     0.1344   1.571    0.128
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1

Residual standard error: 6.817 on 27 degrees of freedom
Multiple R-squared:  0.708, Adjusted R-squared:  0.6864
F-statistic: 32.74 on 2 and 27 DF,  p-value: 6.058e-08
```

68%의 신뢰도
p-value가 0.05보다 작으므로 위 예측은 통계적으로 유의하다.

2) 다중공선성

```
#다중공선성(Multicollinearity) : 독립변수끼리 강한 상관관계를 가지는 현상

#다중공선성을 파악하기 위한 수치적 지표
#VIF(Variance Inflation Factor, 분산팽창인자)
# VIFi= 1 / ( 1 - R^2i)

library(car)

#미국 미니애폴리스 지역의 총인구,백인비율,흑인비율,외국태생,
가계소득,
#빈곤,대학졸업비율을 추정한 데이터셋
df<-MplsDemo
head(df)

#독립적인 그래픽창에 그래프 출력
win.graph()
plot(df[,-1])
#독립변수들의 상관계수
cor(df[,2:7])

#install.packages('corrplot')
library(corrplot)
win.graph()
corrplot(cor(df[,2:7]), method="number")
# white 변수의 경우 다른 변수들과 상관관계가 높음(다중공선성이 의심됨)
```

```
model1<-lm(collegeGrad~.-neighborhood,data=df)
summary(model1)
#설명력은 81.86%로 좋은 모형이지만
#black(흑인비율), foreignBorn(외국태생) 변수의 회귀계수가
양수로 출력됨
#실제 현상을 잘 설명하지 못하는 모형

#white 변수를 제거한 모형
model2<-lm(collegeGrad~.-neighborhood-white,data=df)
summary(model2)
#설명력은 다소 떨어졌지만 회귀계수가 실제 현상을 잘 설명하는
것으로 보임
#black(흑인비율)이 음수로 바뀌었음, foreignBorn(외국태생)
변수는 양수이지만 유의하지 않음

#다중공선성에 대해 확인이 필요한 경우
# p-value가 유의하지 않은 경우
# 회귀계수의 부호가 예상과 다른 경우
# 데이터를 추가,제거시 회귀계수가 많이 변하는 경우

model<-lm(population~.-collegeGrad-neighborhood,data=df)
# ^{-1}  -1승  중괄호를 안써도 됨
print(paste("population의 VIF : ",(1-summary(model)$r.squar
ed)^{-1}))

#다중공선성이 매우 높은 변수
model<-lm(white~.-collegeGrad-neighborhood,data=df)
print(paste("white의 VIF : ",(1-summary(model)$r.squared)^
{-1}))
```

```
model<-lm(black~.-collegeGrad-neighborhood,data=df)
print(paste("black의 VIF : ",(1-summary(out)$r.squared)^{-
1}))

model<-lm(foreignBorn~.-collegeGrad-neighborhood,data=df)
print(paste("foreinBorn의 VIF : ",(1-summary(model)$r.squar
ed)^{-1}))

model<-lm(hhIncome~.-collegeGrad-neighborhood,data=df)
print(paste("hhIncome의 VIF : ",(1-summary(model)$r.square
d)^{-1}))

model<-lm(poverty~.-collegeGrad-neighborhood,data=df)
print(paste("poverty의 VIF : ",(1-summary(model)$r.square
d)^{-1}))

#다중공선성을 계산해주는 함수
vif(model1)
# 다중공선성이 높은 white 변수 제거

model2<-lm(collegeGrad~.-neighborhood-white,data=df)
summary(model2)
vif(model2)
# vif 수치가 많이 낮아졌고 특히 black의 수치도 많이 낮아졌음
```

3) 보스턴주택가격

종속변수

1978년 보스턴의 주택 가격

506개 타운의 주택 가격 중앙값 (단위 1,000 달러)

독립변수

CRIM: 범죄율

INDUS: 비소매상업지역 면적 비율

NOX: 일산화질소 농도

RM: 주택당 방 수

LSTAT: 인구 중 하위 계층 비율

B: 인구 중 흑인 비율

PTRATIO: 학생/교사 비율

ZN: 25,000 평방피트를 초과하는 거주 지역의 비율

CHAS: 찰스강의 경계에 위치한 경우는 1, 아니면 0

AGE: 1940년 이전에 건축된 주택의 비율

RAD: 고속도로까지의 거리

DIS: 고용지원센터의 거리

TAX: 재산세율

```
library(MASS)
head(Boston)
tail(Boston)
```

```
dim(Boston)
```

```
summary(Boston)
```

```
#산점도 행렬
pairs(Boston)
```

```
plot(medv~crim, data=Boston, main="범죄율과 주택가격과의
관계", xlab="범죄율", ylab="주택가격")
```

```
#범죄율과의 상관계수 행렬
(corrmatrix <- cor(Boston)[1,]) # 첫번째 변수
#범죄율이 높으면 주택가격이 떨어진다.
```

```
#강한 양의 상관관계, 강한 음의 상관관계
corrmatrix[corrmatrix > 0.5 | corrmatrix < -0.5]
```

```
#세율과의 상관계수 행렬
(corrmatrix <- cor(Boston)[10,])
#세율이 높으면 주택가격이 떨어진다.
```

```
#강한 양의 상관관계, 강한 음의 상관관계
corrmatrix[corrmatrix > 0.5 | corrmatrix < -0.5]
```

```
#CHAS: 찰스강의 경계에 위치한 경우는 1, 아니면 0
table(Boston$chas)
```

```
#최고가로 팔린 주택들
(seltown <- Boston[Boston$medv == max(Boston$medv),])
```

```
#최저가로 팔린 주택들
(seltown <- Boston[Boston$medv == min(Boston$medv),])
```

```
#다중회귀분석 모델 생성
(model<-lm(medv ~ . , data=Boston))
```

```
#분석결과 요약
summary(model)
#p-value가 0.05보다 작으므로 통계적으로 유의함
#모델의 설명력(예측의 정확성) 73.3%
```

```
#전진선택법과 후진제거법
#후진제거법:기여도가 낮은 항목을 제거함으로써 의미있는 회귀식을 구성하는 과정
reduced<-step(model, direction="backward")
#최종적으로 선택된 변수들 확인
```

```
#최종 결과 확인
summary(reduced)
#p-value가 0.05보다 작으므로 이 회귀모델은 통계적으로 유의함.
#모델의 설명력(신뢰도,예측정확성) : 73.4%
```

4) 주택가격예측2

```
# https://www.kaggle.com/anthonypino/price-analysis-and-linear-regression
df<-read.csv("d:/data/house_regress/data.csv")
head(df)
tail(df)
```

```
library(dplyr)
# Suburb, Address, Type, Method, SellerG, Date,
CouncilArea, Regionname필드 제거
df<-df %>% select(-Suburb, -Address, -Type, -Method,
-SellerG, -Date, -CouncilArea, -Regionname)
```

```
dim(df)
```

```
# 결측값이 있는 행을 제거
df<-na.omit(df)
tail(df)
```

```
dim(df)
```

```
summary(df)
```

```
#상관계수 행렬
(corrmatrix <- cor(df))
```

```
#강한 양의 상관관계, 강한 음의 상관관계
corrmatrix[corrmatrix > 0.5 | corrmatrix < -0.5]
```

```
#install.packages("corrplot")
library(corrplot)
corrplot(cor(df), method="circle")
```

```
#다중회귀분석 모델 생성
model<-lm(Price ~ ., data = df )
model
```

```
#분석결과 요약
summary(model)
#p-value가 0.05보다 작으므로 통계적으로 유의함
#모델의 설명력(예측의 정확성) 0.4965
```

```
#전진선택법과 후진제거법
#후진제거법:기여도가 낮은 항목을 제거함으로써 의미있는 회귀식을 구성하는 과정
reduced<-step(model, direction="backward")
#최종적으로 선택된 변수들 확인
```

```
#최종 결과 확인
summary(reduced)
#p-value가 0.05보다 작으므로 이 회귀모델은 통계적으로 유의함.
#모델의 설명력(신뢰도,예측정확성) : 73.4%
```

5) 보험료예측

```
#회귀분석(보험료 예측)
# https://www.kaggle.com/mirichoi0218/insurance/downloads/insurance.csv/1
df<-read.csv("d:/data/insurance/insurance.csv")
head(df)
tail(df)
```

```
dim(df)
```

```
summary(df)
```

```
#산점도 행렬
pairs(df)
```

```
plot(charges~age, data=df, main="나이와 보험료의 관계",
xlab="나이", ylab="보험료")
```

```
#다중회귀분석 모델 생성
(model<-lm(charges ~ . , data=df))
```

```
#분석결과 요약
summary(model)
#p-value가 0.05보다 작으므로 통계적으로 유의함
#모델의 설명력(예측의 정확성) 0.7494
```

```
#전진선택법과 후진제거법
#후진제거법:기여도가 낮은 항목을 제거함으로써 의미있는 회귀식을 구성하는 과정
```

```
reduced<-step(model, direction="backward")
#최종적으로 선택된 변수들 확인
```

```
#최종 결과 확인
summary(reduced)
#p-value가 0.05보다 작으므로 이 회귀모델은 통계적으로 유의함.
#모델의 설명력(신뢰도,예측정확성) : 0.7496
```

6) 신용카드거래

```
# 원본 데이터셋 출처

# https://www.kaggle.com/janiobachmann/credit-fraud-dealing
-with-imbalanced-datasets/data

# 2013년 9월 유럽 카드 소지자가 신용카드로 거래한 내용, 284807
건의 거래 가운데 492건의 사기거래
# 변수 v1~v28, Amount 거래 금액, Class 0/1 정상거래/사기거래
df<-read.csv("d:/data/creditcard/creditcard.csv")
head(df)
tail(df)
```

```
library(dplyr)
# Time,Class 필드 제거
df<-df %>% select(-Time,-Class)
```

```
head(df)
```

```
dim(df)
```

```
summary(df)
```

```
#다중회귀분석 모델 생성
(model<-lm(Amount ~ . , data=df))
```

```
#분석결과 요약
summary(model)
#p-value가 0.05보다 작으므로 통계적으로 유의함
#모델의 설명력(예측의 정확성) 0.9175
```

```
#전진선택법과 후진제거법
#후진제거법:기여도가 낮은 항목을 제거함으로써 의미있는 회귀식을 구성하는 과정
reduced<-step(model, direction="backward")
#최종적으로 선택된 변수들 확인(V11 변수가 제거됨)
```

```
#최종 결과 확인
summary(reduced)
#p-value가 0.05보다 작으므로 이 회귀모델은 통계적으로 유의함.
#모델의 설명력(신뢰도,예측정확성) : 0.7496
```

7) 난방효율성

```
#에너지 효율성 데이터셋
#건축 구조의 기본 요소인 건물 표면적, 벽과 지붕 면적, 높이, 사각지대, 건물 외형의 간결성,
#건물의 난방과 냉방 효율의 관계 등을 조사한 데이터
#18가지의 건축 특성을 지닌 12가지의 건축 속성, 총 768채의 주택 조사

# X1 : 상대적 크기
# X2 : 건축 표면적
# X3 : 벽체 면적
# X4 : 지붕 면적
# X5 : 전체 높이
# X6 : 건물의 방위
# X7 : 유리창 면적
# X8 : 유리창 면적의 분산
# Y1 : 난방 하중
# Y2 : 냉방 하중
```

```
df<-read.csv("d:/data/energy/ENB2012_data.csv")
head(df)
tail(df)
```

```
library(dplyr)
# 필드 제거
df<-df %>% select(-Y2)
```

```
dim(df)
```

```
summary(df)
```

```
#상관계수 행렬
(corrmatrix <- cor(df))
```

```
#강한 양의 상관관계, 강한 음의 상관관계
corrmatrix[corrmatrix > 0.5 | corrmatrix < -0.5]
```

```
library(corrplot)
corrplot(cor(df), method="circle")
```

```
#다중회귀분석 모델 생성
(model<-lm(Y1 ~ . , data=df))
```

```
#분석결과 요약
summary(model)
#p-value가 0.05보다 작으므로 통계적으로 유의함
#모델의 설명력(예측의 정확성) 0.9154
```

```
#전진선택법과 후진제거법
#후진제거법:기여도가 낮은 항목을 제거함으로써 의미있는 회귀식을
구성하는 과정
reduced<-step(model, direction="backward")
#최종적으로 선택된 변수들 확인
```

```
#최종 결과 확인
summary(reduced)
#p-value가 0.05보다 작으므로 이 회귀모델은 통계적으로 유의함.
#모델의 설명력(신뢰도,예측정확성) : 0.9155
```

8) 냉방효율성

```
#에너지 효율성 데이터셋
#건축 구조의 기본 요소인 건물 표면적, 벽과 지붕 면적, 높이, 사각지대, 건물 외형의 간결성,
#건물의 난방과 냉방 효율의 관계 등을 조사한 데이터
#18가지의 건축 특성을 지닌 12가지의 건축 속성, 총 768채의 주택 조사

# X1 : 상대적 크기
# X2 : 건축 표면적
# X3 : 벽체 면적
# X4 : 지붕 면적
# X5 : 전체 높이
# X6 : 건물의 방위
# X7 : 유리창 면적
# X8 : 유리창 면적의 분산
# Y1 : 난방 하중
# Y2 : 냉방 하중
```

```
df<-read.csv("d:/data/energy/ENB2012_data.csv")
head(df)
tail(df)
```

```
library(dplyr)
# 필드 제거
df<-df %>% select(-Y1)
```

```
head(df)
```

```
dim(df)
```

```
summary(df)
```

```
cor(df)
```

```
#상관계수 행렬
(corrmatrix <- cor(df))
```

```
#강한 양의 상관관계, 강한 음의 상관관계
corrmatrix[corrmatrix > 0.5 | corrmatrix < -0.5]
```

```
library(corrplot)
corrplot(cor(df), method="circle")
```

```
#다중회귀분석 모델 생성
(model<-lm(Y2 ~ . , data=df))
```

```
#분석결과 요약
summary(model)
#p-value가 0.05보다 작으므로 통계적으로 유의함
#모델의 설명력(예측의 정확성) 0.8868
```

```
#전진선택법과 후진제거법
#후진제거법:기여도가 낮은 항목을 제거함으로써 의미있는 회귀식을
구성하는 과정
reduced<-step(model, direction="backward")
#최종적으로 선택된 변수들 확인
```

```
#최종 결과 확인
summary(reduced)
```

```
#p-value가 0.05보다 작으므로 이 회귀모델은 통계적으로 유의함.
#모델의 설명력(신뢰도,예측정확성) : 0.8868
```

```
# 해석(t value 기준)
# X4(지붕 면적), X6(건물의 방위), X8(유리창 면적의 분산)은 유의하지 않음
# X1(상대적 크기) : 건물이 크면 냉방비용이 감소된다.
# X2(건축 표면적) : 건축 표면적이 크면 냉방비용이 감소된다.
# X3(벽체 면적) : 벽체 면적이 크면 냉방비용이 증가한다.
# X5(전체 높이) : 건물 높이가 높으면 냉방비용이 증가한다.
# X7(유리창 면적) : 유리창 면적이 크면 냉방비용이 증가한다.
```

9) 놀이동산만족도

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

```
library(dplyr)
# 필드 제거
df<-df %>% select(-weekend)
```

```
head(df)
```

```
dim(df)
```

```
summary(df)
```

```
cor(df)
```

```
#상관계수 행렬
(corrmatrix <- cor(df))
```

```
#강한 양의 상관관계, 강한 음의 상관관계
corrmatrix[corrmatrix > 0.5 | corrmatrix < -0.5]
```

```
library(corrplot)
corrplot(cor(df), method="circle")
```

```
#다중회귀분석 모델 생성
(model<-lm(overall ~ . , data=df))
```

```
#분석결과 요약
summary(model)
#p-value가 0.05보다 작으므로 통계적으로 유의함
#모델의 설명력(예측의 정확성) 0.6789
```

```
#전진선택법과 후진제거법
#후진제거법:기여도가 낮은 항목을 제거함으로써 의미있는 회귀식을 구성하는 과정
reduced<-step(model, direction="backward")
#최종적으로 선택된 변수들 확인
```

```
#최종 결과 확인
summary(reduced)
#p-value가 0.05보다 작으므로 이 회귀모델은 통계적으로 유의함.
#모델의 설명력(신뢰도,예측정확성) : 0.6789
```

10) 회귀분석 모형 저장, 불러오기

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

model<-lm(overall~num.child + distance + rides + games +
wait + clean, data=df)
summary(model)

save(model, file="d:/data/R/rides_regress.model")

rm(list=ls())  #현재 작업중인 모든 변수들을 제거

load("d:/data/R/rides_regress.model")

ls()

summary(model)
```