R

전산통계

김 승 환

swkim4610@inha.ac.kr

하이테크 401호

강의계획서

1주: Why R?, R Install, R Studio Install, R Documentation

2주: Data type, R Programming

3주: Data Manipulation

4주: Database Connection

5주: Graph

6주: Descriptive Statistics

9주: Correlation & Cross Tabulation

10주: Simple Linear Regression

11주: Multiple Linear Regression

12주: Logistic Regression

13주: Decision Tree

14주: Word Cloud

15주: 숙제 발표

1주: Why R, R/Rstudio Install

- Why R?
- R 설치(Windows, Linux, Mac 등) www.r-project.org , www.rstudio.com
  Linux에서 해보는 것도 바람직함(같은 오픈 소프트웨어이므로 ...)
- R 시작
- Rstudio 사용법(4개 화면의 용도)
- 간단한 분석 해보기

  : 예제로 타이타닉호(http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets)에 관련된 파일을 읽고 간단한 분석을 해보자.

```
setwd("C:₩₩Users₩₩SWKim₩₩Downloads")
titanic<-read.csv("titanic3.csv", header=TRUE)
table(titanic$pclass, titanic$survived)
install.packages("gmodels")
library(gmodels)
CrossTable(titanic$pclass, titanic$survived
           , prop.t=FALSE, expected=TRUE, chisq =TRUE)
```

2주: Data type, R Programming

- a<-1 의 의미, a는 객체이다.
- 변수이름 규칙, <-과 =, 대문자 소문자 구분
- R은 함수로 이루어져 있다.

  foo(1,2) # foo는 함수명, 1, 2는 인자(argument)라고 부른다.
- Scalar와 Vector

  R의 기본 형은 벡터이다. 스칼라는 길이가 1인 벡터를 의미한다. 즉, R에서 스칼라는 존재하지 않는다.
- NA(Not Available)과 NULL

  NA는 결측값을 의미한다. 즉, 어떤 이유로든 이 값을 관측하지 못함을 의미하는 표현이다.

  NULL은 이 변수가 초기화 되지 않았음을 나타낼 때 사용한다.
- 숫자와 문자열, Boolean

  1, 1.5, "hello", 'hello', TRUE, FALSE, &, |, !
- Factor: Categorical Data를 표현할 때 사용한다.

  sex<-factor(c("m","m","f"),c("m","f"))

  이와 같이 factor로 표현하면, 정해진 factor이외에 다른 값이 선언되는 것을 막을 수 있다.

2주: Data type, R Programming

- vector: 1차원 배열 형태의 자료형이다.

  x<-c(1,2,3,4,5), y<-c(1:10), y[1], y[-1], y[c(1,3,5)] 과 같은 형식이 가능하다.

  vector는 집합의 형태로 union(x,y), intersect(x,y), setdiff(x,y)의 연산이 가능하다.

  "a" %in% c("a", "b", "c") 의 결과는 TRUE이다.

  "d" %in% c("a", "b", "c") 의 결과는 FALSE이다.

  x<-c(1,2,3,4,5), x+1, 10-x 같은 연산도 가능하다.

  이와 같은 연산은 자바나 C에 비해 코딩량을 현격하게 줄일 수 있어 가독성, 디버깅이 용이한 프로그래밍을 할 수 있다는 장점이 있다. 하지만, Run Time은 늘어날 수 밖에 없다.

- Sequence: seq 함수를 이용해 1:10, 5:1의 표현이 가능하다.

  seq(1:10), seq(3,7,2), seq(3,7,3), seq(5,1,-2)

- Replication: rep함수를 이용해 반복된 수열을 만들 수 있다.

  rep(1:3,times=5), rep(1:3,each=5), rep(1:3,each=5,times=3)

2주: Data type, R Programming

- List: 벡터와 다르게 서로 다른 형(즉, 문자와 숫자)을 혼합하여 저장이 가능하고
  중간에 데이터를 삽입하거나 삭제하기에 유리한 자료구조이다. 하지만, 저장공간이나 처리속도는 느리다.
  y<-list("foo", 70), x<-list(name="foo", height=70) 과 같은 형태도 가능함
  리스트를 생성할 때, 이름을 지정하면 x$name, x$height와 같은 방식으로 리스트내 데이터에 접근이 가능하다.
  이름을 생략하면, x[[1]], x[[2]]의 형태로 접근해야 한다.
  x[1]과 x[[1]]의 차이: x[1]은 1번째 요소의 전체를 의미하고, x[[1]]은 1번째 요소의 값을 의미한다.

- Matrix: Vector의 확장된 형이다. 즉, 벡터를 하나의 Column으로 보면 몇 개의 Column를 순차적으로
  붙인 형태이다.
  x<-matrix (c(1, 2, 3, 4, 5, 6, 7, 8, 9) , nrow =3)는 행의 개수가 3인 벡터를 만들어 행렬을 만든다.
  x[1,1], x[2,3], x[1:2,3], x[1:2,], x[,3], x[-c(1:2),] 와 같은 접근이 가능하다.

- Matrix Operation: 선형대수학에서 배우는 행렬 연산이 가능하다.
  x*2, x/2, x+y, x * y, x / y 는 대응되는 원소 간의 연산(elementwise operation)이 이루어진다.
  x %*% y는 대응되는 원소의 곱이 아닌 행렬의 곱셈이 이루어진다.
  t(x): Transpose of x, solve(x): Inverse of x

2주: Data type, R Programming

- Array: Matrix는 2차원이다. 행렬을 k차원으로 확장하면 array로 표현가능하다. 하지만, 잘 사용하지 않는다.

- Data Frame: R에서 가장 중요한 자료형이다. 우리는 일반적으로 Excel Sheet 형태의 자료를 많이 취급하는데 이러한 형식의 자료는 data frame으로 표현한다.
  data frame은 k개의 변수와 n개의 관측값 으로 이루어지는 것이 일반적인 형태이다.
  여기서, 변수 각각은 벡터이므로 벡터를 모은 것이 data frame 이며, 숫자 벡터, 문자 벡터를 하나의 데이터 프레임으로 표현 가능하다는 점이 행렬과 다른 점이다.
  d<-data.frame(x=c(1:5), y=c("A","B","C","D","E"))
  일반적으로 데이터 프레임은 많은 량의 자료와 변수를 취급하기 때문에 str(d), head(d), View(d) 와 같은 명령을 통해 데이터프레임의 구조를 파악한다.
  dx, dy 같은 명령으로 데이터 프레임 안의 변수에 접근이 가능하다.

- 데이터 타입 판별: R에서는 여러가지 자료형식이 있기 때문에 특정 객체의 변수가 어떤 자료형인지 모를 경우가 있다.
  class(d) 라는 명령은 d 객체가 어떤 클래스에서 나왔는지 알려주는 함수이다.
  is.data.frame(d), is.numeric(x), is.factor(y)

2주: Data type, R Programming

- 데이터 타입 변환: R은 함수로 구성되어 있으므로 함수에 입력될 변수의 타입이 정해져 있다.
  내가 가지고 있는 객체가 원하는 타입이 아닐 경우, 원하는 타입으로 변환하여 사용할 수 있다.
  x <- c("m", "f"), as.factor(x)는 x 객체를 factor 형으로 변환하는 것이다.

- if 문: if (x %% 2==0) {print("even")} else {print("odd")} 의 형식이다.

- for 문: for (i in 1:10) {print (i)} 의 형식이다.

- while 문: i<-0, while(i<10) {print(i) 의 형식이다.
  i<-i+1}

- 사용자 함수 정의: R은 함수로 구성되어 있다. R에서는 기본제공함수와 각종 패키지를 통해 구한 함수,
  내가 만든 함수를 쓸 수 있는데 아래와 같이 함수를 만들 수 있다.

```
lag<-function(x,lagx) {
  n<-length(x)
  for(i in 1:n-1) lagx[i+1]<-x[i]
  return(lagx)
}
x<-c(1:10)
y<-NULL
lag(x,y)
```

2주: Data type, R Programming

- 객체 삭제: rm(x)는 객체 x를 메모리에서 제거하라는 명령이다. R은 쉽고 강력한 툴이지만 R의 치명적인 단점은 데이터를 메모리에서 처리한다는 점이다. 빅데이터 시대에 데이터를 메모리에 로딩한다는 것은 빅데이터를 처리할 수 없다는 의미이다.
  반면, SAS는 데이터를 디스크에 저장한 상태로 처리하는 방식을 택해 빅데이터의 처리에 적합한 구조이다.
  그러므로, 데이터의 크기가 커지면 R은 급격하게 성능이 저하된다.
  R의 성능저하를 막으려면 메모리 관리를 잘해야 하는데 가장 쉬운 메모리 관리는 메모리에 올라와 있는 불필요한 객체를 없애는 것이다.
  또 다른 방법은 데이터베이스를 같이 활용하는 방법, 그리고 여러 대의 컴퓨터로 병렬처리하는 방법이 있다.

- ls() : 메모리에 있는 모든 객체를 보여준다.
  rm(list=ls()) → 메모리에 있는 모든 객체를 삭제한다.

숙제 2: nCr 사용자 함수 만들기

3주: Data Manipulation

- 자료분석에서 데이터는 내 입맛에 맞게 존재하지 않으므로 데이터의 가공은 필수적이다.
  또한, 데이터 가공을 빨리 끝낼 수록 분석시간이 많이 확보되므로 데이터 가공 속도는 지식과 노련미의 결정체라 할 수 있다.(프로와 아마추어의 차이)

- 예제 데이터 불러오기
  data(iris), data(mtcars)

- CSV 파일 입력: 아래는 D:₩a.csv 파일을 읽어 a라는 데이터 프레임 객체에 저장하는 명령이다.
  a<-read.csv("d:/a.csv", header=TRUE)

- CSV 파일 출력: 아래는 iris 객체를 iris.csv라는 폴더에 저장하는 명령이다.
  write.csv (iris, " iris.csv", row.names=FALSE)

- Project, Working Directory의 사용
  본격적으로 R을 사용하려면 프로젝트별로 파일을 관리할 필요가 있다.
  이를 위해 RStudio에서는 R Working Directory를 지정하고 그 밑에 프로젝트 폴더를 만들어 사용한다.
  현재, working dir.을 확인하는 방법은 getwd(), working dir.을 바꾸는 방법은 setwd("<path>") 명령을 쓴다.

- save(), load(): 메모리에 있는 객체를 디스크에 저장하기 위해서는 save(x,file="x.RData") 명령을 사용한다.
  불러올 때에는 load("x.RData") 로 불러올 수 있다.

3주: Data Manipulation

- rbind(), cbind(): data를 합치는 함수로 결과는 행렬 혹은 데이터 프레임이다.

  rbind (c(1, 2, 3) , c(4, 5, 6)), cbind (c(1, 2, 3) , c(4, 5, 6))

- apply(): 행방향, 열방향으로 특정 함수를 적용시키는 함수이다.

  아래의 명령처럼 행방향은 1, 열방향은 2를 주면 되고 함수로는 sum, mean, min, max 등을 사용할 수 있다.

```
> d <- matrix (1:9 , ncol =3)
> d
     [,1] [,2] [,3]
[1,]    1    4    7
[2,]    2    5    8
[3,]    3    6    9
> apply(d,1,sum) # 행방향으로 합을 구함
[1] 12 15 18
> apply(d,2,sum) # 열방향으로 합을 구함
[1]  6 15 24
> apply ( iris, 2, min )
Sepal.Length  Sepal.Width Petal.Length  Petal.Width      Species
       "4.3"         "2.0"         "1.0"         "0.1"       "setosa"
> apply ( iris, 2, max )
Sepal.Length  Sepal.Width Petal.Length  Petal.Width      Species
       "7.9"         "4.4"         "6.9"         "2.5"    "virginica"
```

3주: Data Manipulation

- lapply() : 벡터나 리스트에 대해 특정 함수를 적용한다. 결과도 리스트로 제공된다. l은 리스트를 의미한다. 리스트는 unlist() 함수로 벡터로 변환가능하다.

```
> x<-list(x1=c(1:5), x2=c(5:10))
> lapply(x,sum)
$x1
[1] 15

$x2
[1] 45

lapply(iris[,1:4], mean)
$Sepal.Length
[1] 5.843333

$Sepal.Width
[1] 3.057333

$Petal.Length
[1] 3.758

$Petal.Width
[1] 1.199333
```

3주: Data Manipulation

▸ sapply() : 벡터나 리스트, 데이터프레임에 대해 특정 함수를 적용한다. 결과는 행렬이나 벡터로 제공된다.

```
> sapply(iris[,1:4], mean)
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
    5.843333     3.057333     3.758000     1.199333

> sapply(iris,class)  #data frame에 소속된 멤버의 class를 출력한다.
Sepal.Length  Sepal.Width Petal.Length  Petal.Width      Species
   "numeric"    "numeric"    "numeric"    "numeric"     "factor"
```

▸ tapply() : 그룹별 처리를 위한 함수이다. 아래의 예는 iris 종류에 따른 Sepal.Length의 평균이다.

```
> tapply ( iris $ Sepal.Length , iris $ Species , mean )
    setosa versicolor  virginica
     5.006      5.936      6.588
```

3주: Data Manipulation

- doBy Package: summaryBy(), orderBy(), splitBy(), sampleBy()로 이루어져 있다.
  doBy 패키지를 사용하기 위해서는 install.packages("doBy") 명령으로 패키지를 다운 받아야 하고,
  library(doBy) 명령을 통해 메모리로 로딩한 이후에 사용하여야 한다.
  summaryBy(Sepal.Length~Sepcies, iris)
  summary(iris), quantile(iris$Sepal.Length)

  orderBy(~Sepal.Width, iris) # Sepal.Width 기준으로 모든 자료를 Sort한다.
  orderBy(~ Species+Sepal.Width , iris)
  sample(1:10,5), sample(1:10,5,replace=TRUE) #Sampling 방법을 제공한다.
  sampleBy (~Species , frac =0.1 , data = iris) # 종에 따라 10%씩 샘플링한다.

- split(): 데이터를 분리하는 함수이다.
  x<-split(iris,iris$Species) # 종에 따라 데이터프레임을 분리해서 리스트에 저장한다.
  x[[1]], x[[2]], x[[3]]은 각각 분리된 데이터이다.
  x1<-as.data.frame(x[[1]]) # setosa 만의 데이터 프레임을 얻는다.

- subset(): 데이터 프레임의 일부를 얻는다.
  subset (iris , Species == "setosa")

3주: Data Manipulation

- merge(): 두개의 데이터프레임을 공통기준으로 묶는다.

```
> x <- data.frame ( name =c("a", "b", "c"), math =c(1, 2, 3))
> y <- data.frame ( name =c("c", "b", "a"), english =c(4, 5, 6))
> merge (x, y, all= TRUE )
  name math english
1    a    1       6
2    b    2       5
3    c    3       4
```

- sort(): 정렬 함수로 sort (x, decreasing = TRUE)의 형식을 사용한다.

- order(): 순위함수, order(x) or order(-x)로 명령하여 Ascending, Descending order를 조정한다.

  newdata <- mtcars[order(mtcars$mpg),]는 order함수를 이용하여 데이터 프레임을 소트하는 예이다.

- with(): 분석시에 하나의 데이터 프레임을 지속적으로 핸들링하는데 매번 iris$Sepal.Length 처럼 쓰기가 불편하다. 이 때, with(data=iris, <R Code>) 의 형식으로 쓰면 R Code에서 iris$를 생략할 수 있다.

```
> summary(aov(iris$Sepal.Length~iris$Species))
> with(data=iris,summary(aov(Sepal.Length~Species)))
```

3주: Data Manipulation

- sqldf Package: R의 데이터프레임을 데이터베이스의 테이블처럼 SQL 문을 쓸 수 있도록 해주는 함수이다.
  R에서 제공하는 Data Manipulation 기능은 많고 강력하지만, 함수가 너무 많은 단점도 있다.
  SQL은 데이터를 다루는 가장 많이 쓰이는 언어이므로 SQL에 익숙한 사람은 sqldf 패키지가 대안이 된다.
  install.packages("sqldf"), library(sqldf) 명령으로 sqldf를 메모리에 올려놓자.
  sqldf("select distinct species from iris") 의 형식으로 SQL 문을 사용할 수 있다.
  중요한 사실은 SQL에서는 대소문자의 구별이 없고 Sepal.Length와 같이 "." 이 있는 필드명을 처리할 수 없다.
  처음부터 변수명에 "."을 안쓰는 것도 방법이다.

- RODBC Package: R에서 ODBC를 통해 외부 데이터베이스에 접속하여 SQL을 사용할 수 있도록 해주는 패키지이다.

4주: Database Connection

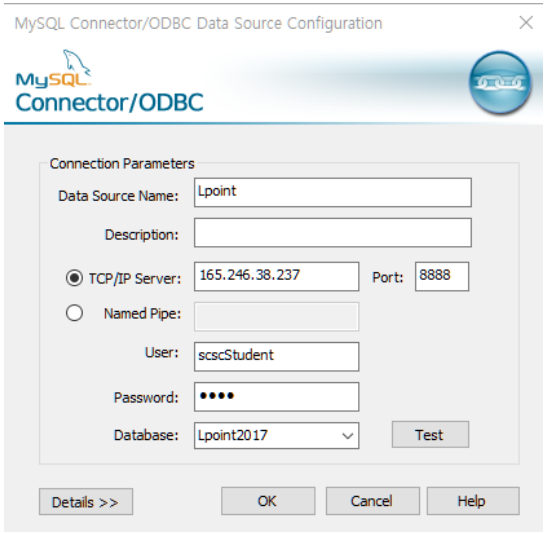
- 실습 내용(실습에는 원격지에 데이터베이스를 접속해 데이터를 분석하는 방법을 실습한다.)
  1. R에서 외부 DB Access
  2. Data Exploring
  3. RFM Analysis
  4. Customer Segmentation

Lpoint2017 데이터베이스 구조

```
use Lpoint2017;
create table Cust (
    ID           char(5),
    GENDER       char(1),
    AGE_PRD      char(5),
    HOM_PST_NO   char(3)
);
load data local infile '/home/swkim/DB/Lpoint/2017/cust.txt' into table Cust
fields terminated by ',';

create table Tr1 (
    ID       char(11),
    RCT_NO       char(6),
    BIZ_UNIT     char(3),
    PD_S_C       char(4),
    BR_C         char(4),
    DE_DT        char(8),
    DE_HR        INT,
    BUY_AM       INT,
    BUY_CT       INT
);
```

4주: Database Connection

1. ODBC 설정 및 R에서 원격 접속 Test




```
library(RODBC)
library(sqldf)
Conn<-odbcConnect("Lpoint",uid='scscStudent',pwd='1234')
Cust<-sqlQuery(Conn,"select count(*) from Cust")
```

4주: Database Connection

2. Data Exploring & Manipulation

(1) 고객수를 구하라(성별, 연령대, 지역별)

SELECT GENDER, count(*) from Cust group by GENDER

SELECT GENDER, AGE_PRD, count(*) from Cust group by GENDER, AGE_PRD")

(2) 인당 백화점, 마트, 슈퍼, 편의점, Drug Store의 평균 지출금액을 구하시오.

(3) 인당 쇼핑 외 업종의 평균지출금액을 구하시오.

숙제 3: 위의 내용에서 (1) ~ (4)를 수행한 결과를 제출하세요

4주: Database Connection

```
library(RODBC)
library(sqldf)
library(gmodels)

setwd("D:₩₩Lecture₩₩LPoint")

Conn<-odbcConnect("Lpoint",uid='scscStudent',pwd='1234')
temp<-sqlQuery(Conn,"select * from ZipCode")
sqlQuery(Conn,"select count(*) from Cust")
sqlQuery(Conn,"select GENDER, count(*) from Cust group by GENDER")
sqlQuery(Conn,"select GENDER, AGE_PRD, count(*) from Cust group by GENDER, AGE_PRD")
sqlQuery(Conn,"select HOM_PST_NO, count(*) from Cust group by HOM_PST_NO")

Cust<-sqlQuery(Conn,"select a.ID, a.GENDER, a.AGE_PRD, a.HOM_PST_NO, b.CITY, b.GU from Cust a
                    left outer join ZipCode b on a.HOM_PST_NO=b.HOM_PST_NO order by a.ID")

temp<-sort(table(Cust$CITY))
barplot(temp)

A01<-sqlQuery(Conn,"select ID, count(*) as cntA01 from Tr1 where BIZ_UNIT='A01' group by ID ")
A02<-sqlQuery(Conn,"select ID, count(*) as cntA02 from Tr1 where BIZ_UNIT='A02' group by ID ")
A03<-sqlQuery(Conn,"select ID, count(*) as cntA03 from Tr1 where BIZ_UNIT='A03' group by ID ")
A04<-sqlQuery(Conn,"select ID, count(*) as cntA04 from Tr1 where BIZ_UNIT='A04' group by ID ")
A05<-sqlQuery(Conn,"select ID, count(*) as cntA05 from Tr1 where BIZ_UNIT='A05' group by ID ")
Cust[is.na(Cust)]  <- 0
Cust$cntA35<-Cust$cntA03+Cust$cntA04+Cust$cntA05
```

4주: Database Connection

```
Cust$lcntA01<-ifelse(Cust$cntA01==0,0,log(Cust$cntA01))
Cust$lcntA02<-ifelse(Cust$cntA02==0,0,log(Cust$cntA02))
Cust$lcntA35<-ifelse(Cust$cntA35==0,0,log(Cust$cntA35))

keeps<-c("ID","GENDER","AGE_PRD","HOM_PST_NO","lcntA01","lcntA02","lcntA35")
Cust1<-Cust[keeps]
hist(Cust1$lcntA01)
range01 <- function(x){(x-min(x))/(max(x)-min(x))}

Cust2<-range01(Cust1[,c(-1,-2,-3,-4)])

library(caret)
nearZeroVar(Cust2, saveMetrics = TRUE)
segments <- kmeans(Cust2, 4)
segments$size
segments$centers
Cust$segment<-segments$cluster

aggregate(data = Cust, cntA01 ~ segment, mean)
aggregate(data = Cust, cntA02 ~ segment, mean)
aggregate(data = Cust, cntA03 ~ segment, mean)
aggregate(data = Cust, cntA04 ~ segment, mean)
aggregate(data = Cust, cntA05 ~ segment, mean)

CrossTable(x = Cust$GENDER, y = Cust$segment, prop.t=FALSE, expected=TRUE, chisq =TRUE)
CrossTable(x = Cust$AGE_PRD, y = Cust$segment, prop.t=FALSE, expected=TRUE, chisq =TRUE)
CrossTable(x = Cust$CITY, y = Cust$segment, prop.t=FALSE, expected=TRUE, chisq =TRUE)
```

5주: Graph

- SAS에서 데이터 마이닝의 단계를 SEMMA로 표현한다.

  S: Sampling, E: Explore, M: Manipulation, M: Modeling, A: Assessment 이다.

  즉, 데이터마이닝은 Data Manipulation, Data Exploring이 중요하다.

  그래프는 Data Exploring에 중요한 도구이다.

- Scatterplot: 산점도

```
install.packages("mlbench")
library(mlbench)
data(Ozone)
head(Ozone)
?Ozone
with(data=Ozone,plot(V8,V9))
with(data=Ozone,plot(V8,V9), xlab="Temperature measured at Sandburg",
                             ylab="Temperature measured at El Monte", main="Ozone")
```

- 산점도에서 포인트의 종류, 크기를 바꿔보자.

5주: Graph

- 꺽은선 그래프: cars 데이터를 이용하여 속도별 제동거리 꺽은선 그래프 그리기

```
data(cars)
head(cars)
res<-tapply ( cars $dist , cars $speed , mean )
plot(res, type="o", cex=0.5, xlab="speed", ylab="dist")
```

- 그래프 배열: 여러 개의 그래프를 한꺼번에 그리고 싶을 때 사용한다.

```
par1<-par(mfrow=c(1,2))
with(Ozone,plot(V8,V9, xlab="Sandburg Temp.", ylab="El Monte Temp."))
with(Ozone,plot(V10,V12, xlab=" height (feet) at LAX", ylab=" temperature (degrees F) at LAX"))
par(par1)
```

- points() : 이미 그려진 plot 위에 점을 추가한다.

```
data(iris)
with(iris,plot(Sepal.Width,Sepal.Length,cex=0.5,pch=20,xlab="width",ylab="length",main="Iris Data"))
with(iris,points(Petal.Width,Petal.Length,cex=0.5,pch="+",col="#FF0000"))
```

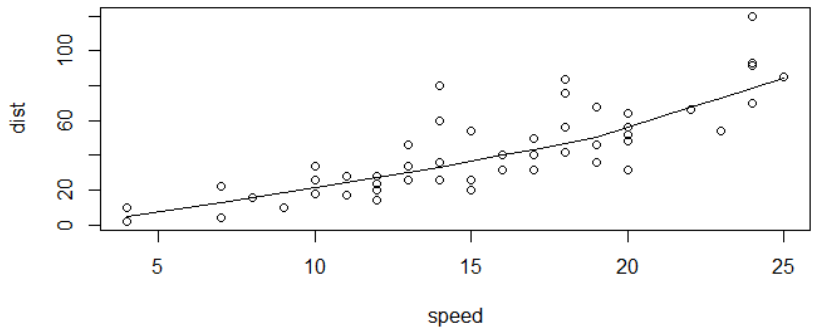
5주: Graph

- lines() : 이미 그려진 plot 위에 선을 추가한다.

```
x <- seq (0, 2*pi , 0.1)
y <- sin(x)
plot (x, y, cex=.5 , col ="red ")
lines (x, y)
```

```
data(cars)
head(cars)
plot(cars)
lowess(cars)
lout<-lowess(cars)
lines(lout)
```

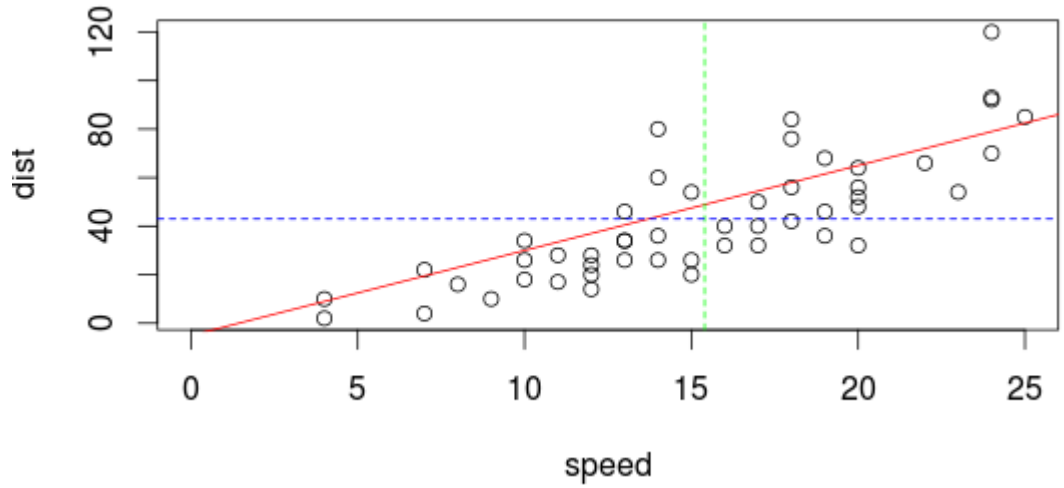
Local Polynomial Regression Fitting



5주: Graph

- abline() : 직선 그래프를 그린다.

```
plot (cars , xlim =c(0, 25) )
abline (a=-5, b=3.5 , col ="red ") # 절편이 -50이고, 기울기가 3.5인 직선을 그린다.
abline (h= mean ( cars $ dist ), lty =2, col =" blue ")
abline (v= mean ( cars $ speed ), lty =2, col =" green ")
```



5주: Graph

- curve() : 곡선 그래프를 그린다.

```
curve(sin , 0, 2*pi)
curve(x^2,-1,1)
curve(exp(-0.5*x^2)/sqrt(2*pi), -3, 3)
```

- 기타 실습

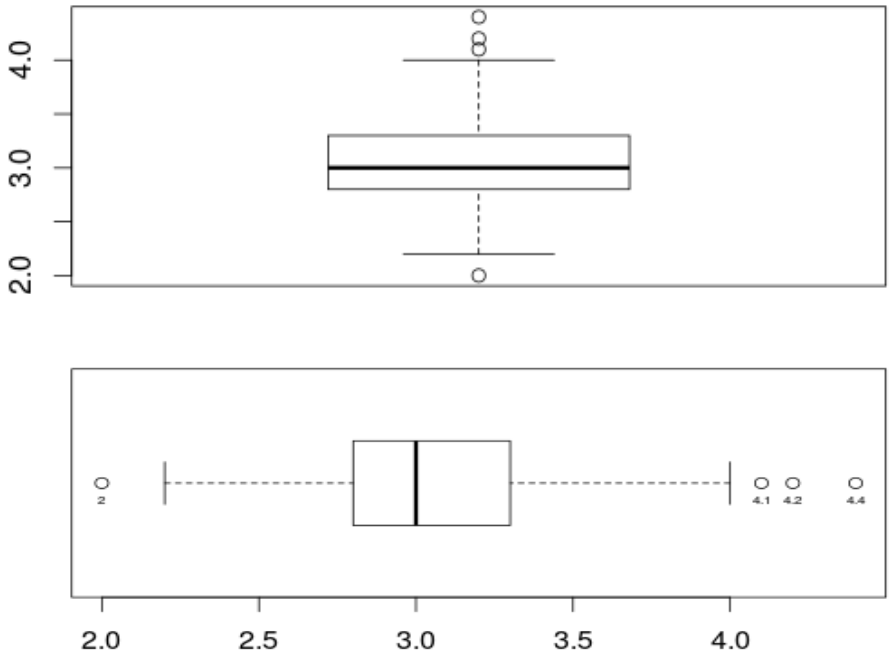
```
plot (cars , cex=.5)
text ( cars $speed , cars $dist , pos =4, cex =.5)

plot (cars , cex=.5)
identify ( cars $speed , cars $ dist )

plot ( iris $ Sepal.Width , iris $ Sepal.Length , cex =.5 , pch =20, xlab =" width ", ylab =" length ")
points ( iris $ Petal.Width , iris $ Petal.Length , cex =.5, pch="+", col="#FF0000")
legend ("topright", legend =c("Sepal", "Petal"), pch=c(20 , 43) , cex=.8 , col=c("black", "red"),
          bg="gray")
```

5주: Graph

- Boxplot() : 상자그림을 그린다.

```
boxplot ( iris $ Sepal.Width )
boxstats <- boxplot ( iris $ Sepal.Width )
boxstats
boxstats <- boxplot ( iris $ Sepal.Width , horizontal = TRUE )
text ( boxstats $out , rep (1, NROW ( boxstats $out )), labels = boxstats $out , pos =1, cex=.5)
```

```
> boxstats
$ stats
[ ,1]
[1 ,] 2.2    # Lower Whisker
[2 ,] 2.8    # 25% Quantile
[3 ,] 3.0    # Median
[4 ,] 3.3    # 75% Quantile
[5 ,] 4.0    # Upper Whisker
```



5주: Graph

- hist() : 히스토그램을 그린다.

```
hist ( iris $ Sepal.Width )
hist ( iris $ Sepal.Width , freq = FALSE )
```

- density() : 밀도그림을 그린다.

```
hist ( iris $ Sepal.Width )
hist ( iris $ Sepal.Width , freq = FALSE )
```

- barplot() : 막대그림을 그린다.

```
barplot ( tapply ( iris $ Sepal.Width , iris $ Species , mean ))
```

- pie() : 파이그림을 그린다.

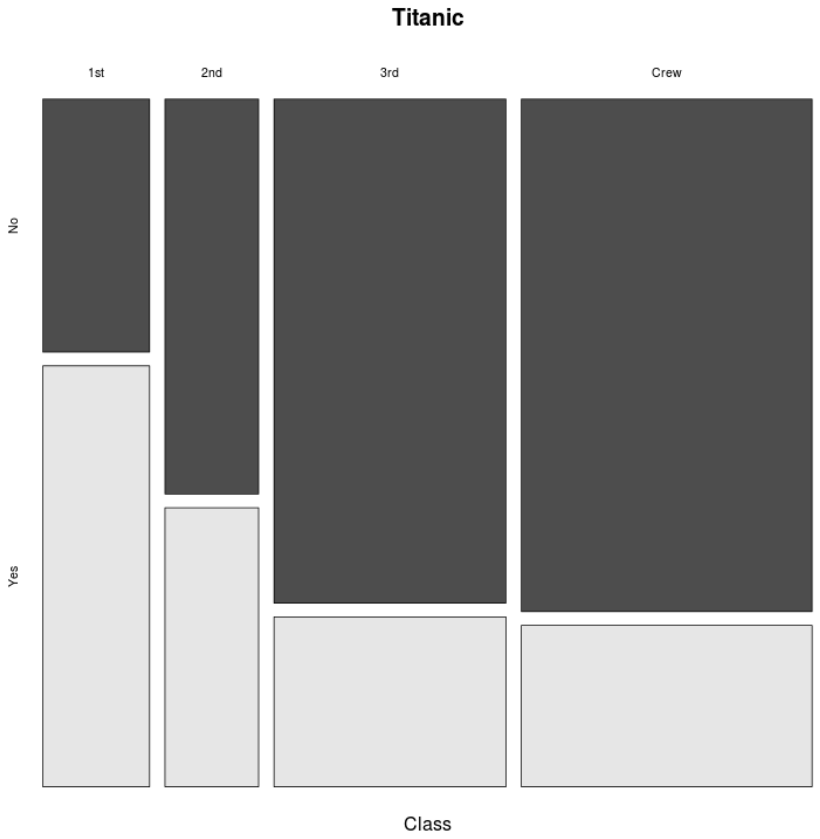
```
x<-c(0,1,3,1,2,0,4,6,3,2,2,1,0,1,2)
f<-table(x)
pie(f,labels=c(0,1,2,3,4,6), col=rainbow(length(f)),main="# of Children")
```

5주: Graph

- mosaicplot() : 모자이크 그림을 그린다.

```
mosaicplot (Freq ~ Class + Survived , data = Titanic , color = TRUE )
```

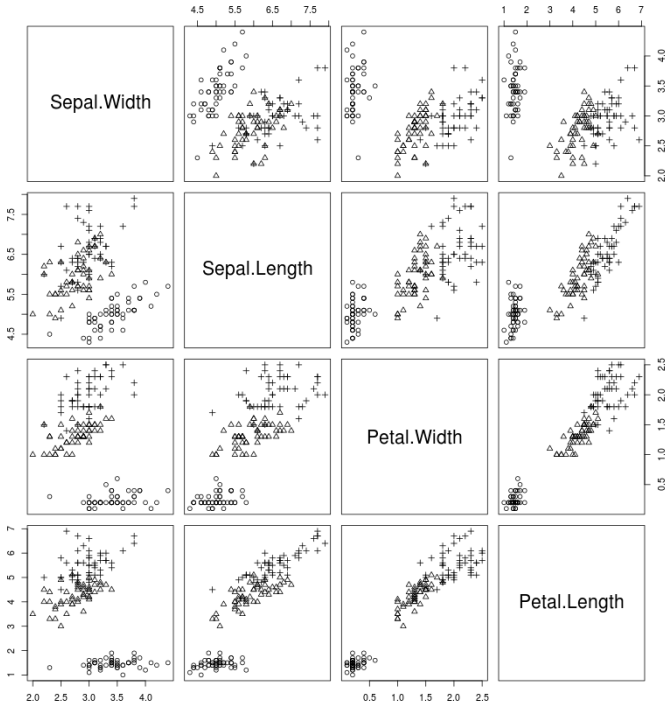
	Class	Sex	Age	Survived	Freq
1	1st	Male	Child	No	0
2	2nd	Male	Child	No	0
3	3rd	Male	Child	No	35
4	Crew	Male	Child	No	0
5	1st	Female	Child	No	0
6	2nd	Female	Child	No	0
7	3rd	Female	Child	No	17
8	Crew	Female	Child	No	0
9	1st	Male	Adult	No	118
10	2nd	Male	Adult	No	154
11	3rd	Male	Adult	No	387
12	Crew	Male	Adult	No	670
13	1st	Female	Adult	No	4
14	2nd	Female	Adult	No	13
15	3rd	Female	Adult	No	89
16	Crew	Female	Adult	No	3



5주: Graph

- pairs() : 산점도 행렬을 그린다.

```
pairs(~Sepal.Width+Sepal.Length+Petal.Width+Petal.Length, data=iris,
pch=c(1:3)[iris$Species])
```



6주: Random Number and Simulation

▸ Random Number와 Distribution

난수: 규칙성이 존재하지 않는 수열 값으로 이전 값으로부터 다음값의 예측이 불가능하다.

Quasi-Random Number(Pseudo Random Number): 컴퓨터에 의해 생성되는 난수열로 실제로는 규칙성이 존재하지만, 규칙성이 없는 것처럼 보이기 때문에 난수의 대용으로 시뮬레이션에서 많이 사용되는 방법이다.

• Random number generation

 Manual method: coin flipping, dice rolling, card shuffling and roulette wheels

We can get “truly” random numbers but these methods were too slow for general use.

 Method of using computer

1) Mid-Square method

For example, if we are generating 4-digit numbers and arrive at 5232

5232**2 = 27,373,824, we get 3738

3738**2 = 13,972,644, we get 9726

9726**2 = 94,595,080, we get 5950

and so on. This numbers are not “truly” random, we called this “Pseudo-random” or “Quasi-Random”.

6주: Random Number and Simulation

❒ Multiplicative generator

The second widely used generator is the multiplicative generator

$X_i = aX_{i-1} (\mathrm{mod}\ m)$ which is a particular case of the (2.2.2) with c=0.

Generally, a full period cannot be achieve here, but a maximal period can provided when

$m = 2^{\beta}, a = 8r \pm 3$.

The procedure for generating pseudorandom numbers on a binary computer can be written as:

1) Choose any odd number as a starting values X_0. **X0=iyr +100*(imonth-1+12*(iday-1+31*(ihour+24* (imin+60 *isec)))**

2) Choose a close to $2^{\beta/2}$ (if $\beta = 35, a = 2^{17} + 3$ is a good selection).

3) Compute X_1.

4) Calculate $U_1 = \frac{X_1}{2^{\beta}}$ to obtain a uniformly distributed variable.

Ex: U(0,1) Generator

beta=35; a=2**17+3; x0=331; m=2**beta;

x[1]=a*x0/m-int(a*x0/m); u[1]=x[1]/m;

Loop i=2 to rndcount:

    x[i]=a*x[i-1]/m-int(a*x[i-1]/m); u[i]=x[i]/m;

Loop end

숙제 4: 난수 발생기를 Rand()의 함수로 만들어보세요

6주: Random Number and Simulation

- 주요 분포의 **Random number generation**

❑ Bin (n,p)

Generate $U_1, \cdots, U_n \sim U(0,1)$

$RND = \#(U_i \leq p)$

Return RND

❑ $N(\mu, \sigma^2)$

Box & Muller Transformation

Generate $U_1, U_2 \sim U(0,1)$

$Z_1 = (-2\ln U_1)^{0.5} \cos 2\pi U_2,\ Z_2 = (-2\ln U_1)^{0.5} \sin 2\pi U_2$

$RND_1 = \mu + \sigma Z_1, RND_2 = \mu + \sigma Z_2$

Return

❑ $Exp(\lambda)$

Generate $U \sim U(0,1)$

$$RND = -\frac{\log(U)}{\lambda}$$

Return

6주: Random Number and Simulation

- runif(): Uniform 분포를 따르는 난수를 만든다.

```
x<-runif(100)
hist(x)
```

- set.seed(): 난수의 초기값을 특정값으로 고정시킨다.

```
set.seed(1)
x<-runif(1)
x
set.seed(1)
x<-runif(1)
x
```

- rbinom(): 이항분포를 따르는 난수를 만든다.

```
x<-rbinom(10,1,0.5) # n=1, p=0.5인 이항분포 난수 10개를 만든다.
y<-rbinom(4,1,0.3)  # 타율이 3할인 타자의 4타석 결과
z<-rbinom(10,4,0.3) # 타율이 3할인 타자의 10일동안 일별 안타수
```

6주: Random Number and Simulation

- 포아송 분포: 단위시간이나 면적 등 일정한 단위에서 발생하는 사건의 수는 평균이 m인 포아송분포를 따른다.

$$f(x) = \frac{e^{-m}m^x}{x!},\ x = 0,1,2,\ \cdots,\ E(X) = V(X) = m$$

1년에 평균 4회 사고가 일어나는 교차로에서 6회 사고가 발생할 확률은?

X: 년간 사고횟수 ~ Poisson(4), $P(X = 6|m = 4) = \frac{e^{-m}m^x}{x!} = \frac{e^{-4}4^6}{6!} = 0.1042$

```
dpois(6,4)  # 평균이 4인 포아송분포에서 확률변수가 6일 확률을 구한다.
rpois(10,4) # 평균이 4인 포아송분포에서 10년치 사고건수
```

6주: Random Number and Simulation

▸ 지수 분포: 포아송 프로세스에서 사건과 사건의 시간간격은 지수분포를 따른다.

$$f(w) = \lambda e^{-\lambda w} = \frac{1}{\theta} exp\left(-\frac{w}{\theta}\right), \mathrm{E(X)} = \theta, \mathrm{V(X)} = \theta^2$$

수명이 10년인 냉장고가 20년 이상 고장이 없을 확률, 5년 이내 고장확률은?

X: 고장까지의 수명~Exp(10)

$$P(X > 20) = \int_{20}^{\infty} \frac{1}{10} \exp\left(-\frac{x}{10}\right) dx = \frac{1}{10} \cdot -10 \exp\left(-\frac{x}{10}\right)_{20}^{\infty} = \exp\left(-\frac{20}{10}\right) = \exp(-2) = 0.135$$

$$P(X < 5) = \int_{0}^{5} \frac{1}{10} \exp\left(-\frac{x}{10}\right) dx = \frac{1}{10} \cdot -10 \exp\left(-\frac{x}{10}\right)_{0}^{5} = 1 - \exp\left(-\frac{1}{2}\right) = 0.39$$

```
x<-rexp(100,0.1) #평균이 10인 난수 100개를 생성한다.
sum<-sum(x>20) # 이중 20 이상인 난수의 갯수
sum/100 # 냉장고 수명이 20년 이상인 냉장고 비율 Simulation 결과값
1-pexp(20,0.1) # 냉장고 수명이 20년 이상인 냉장고 비율 True 값
```

6주: Random Number and Simulation

▸ 정규 분포: 정규분포는 오차의 분포이다. (Gaussian Error Curve)

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \quad -\infty < x < \infty \qquad E(X) = \mu,\ V(X) = \sigma^2$$

```
x<-rnorm(100,0,1)
hist(x)
summary(x)
pnorm(1.645,0,1)  # 0.95
qnorm(0.975,0,1)  # 1.96
```

▸ Chi-Square 분포: N(0,1)을 따르는 확률변수의 제곱은 자유도가 1인 카이제곱분포를 따른다.
자유도가 1인 카이제곱분포를 r 번 더하면 자유도가 r 인 카이제곱분포가 된다.

$$f(x) = \frac{1}{\Gamma\left(\frac{v}{2}\right)}\left(\frac{1}{2}\right)^{v/2} x^{v/2-1}\exp\left(-\frac{x}{2}\right), \quad x > 0 \qquad E(X) = v,\ V(X) = 2v$$

```
x<-rchisq(100,20) # 자유도가 20인 카이제곱 난수 100개를 생성한다.
hist(x)
```

6주: Random Number and Simulation

- Simulation(모의실험): 실제 시스템에 가까운 수학적 모형을 정의하고 이 모형에서 관심의 대상에 대한 Behavior를 관찰하는 과정을 말한다.(Game, Flight Simulator 등)

숙제 5: Coffee shop에 들어오는 손님의 시간 간격은 평균이 10분인 지수분포를 따른다.
그리고, 이 손님이 Coffee Shop에서 머무는 시간은 평균이 30분인 카이제곱분포를 따른다고 가정할 때,
커피숍에 의자는 최소 몇 개를 준비해야 자리가 없어 서 있는 손님을 없게 할 수 있는가?

7주: Descriptive Statistics

- mean(): 평균을 구한다.
- var(): 분산을 구한다.
- sd(): 표준편차를 구한다.
- fivenum(): 5 number summary를 구한다.(최소값, Q1, 중앙값, Q3, 최대값)
- sample(): 표본추출을 수행한다.

```
x<-rnorm(10,10,3)
mean(x)
sd(x)
median(x)
fivenum(x)
sample(x,5)
sample(x,5,replace=TRUE)
```

7주: Descriptive Statistics

- 층화 추출: Stratified Sampling

  표본 추출시, 먼저 층을 나누고 각 층에서 Simple Random Sampling을 시행한다.

```
strata(c("Species"), size=c(3,3,3), method="srswor", data=iris)
strata(c("Species"), size=c(3,3,3), method="srswr", data=iris)
```

- 계통 추출: Systematic Sampling

  표본 추출시, 일정 간격으로 표본을 추출한다.

```
sampleBy(~1, frac=0.1, data=iris,systematic=TRUE)
```

- 분할표: Contingency Table

  명목형이나 순서형 자료로 표를 만든다.

```
x<-c("a", "a", "a", "b","b","c")
table(x)
d<-data.frame(x=c("1","2","2","1"),y=c("A","B","A","B"), num=c(3,5,8,7))
xtabs(num~x+y, data=d) # 이미 frequenc가 요약되어 있는 경우
```

7주: Descriptive Statistics

```
t1 <- rep(c("A","B","C"),5)
t2 <- rpois(15,4)
df <- data.frame(ques=t1,resp=t2)
head(df)
tab<-xtabs(~ques+resp,data=df) # frequency가 요약되어 있지 않은 경우
tab
margin.table(tab,1) # RowSum
margin.table(tab,2) # ColSum
margin.table(tab)   # Total
prop.table(tab,1) # 행비율 합이 100%
prop.table(tab,2) # 열비율 합이 100%
prop.table(tab)   # 모든 비율 합이 100%
```

- 독립성 검정: Contingency Table에서 (i,j) 번째 셀의 확률에 대해 P(i,j) = P(i) * P(j) 이면 두개의 명목변수는 서로 독립이다. 두개의 명목변수가 서로 독립일 때, 아래의 식은 카이제곱 분포를 따른다. 여기서, r의 행의 수, c는 열의 수이다.

$$\chi^2 = \sum\sum \frac{(O_{ij}-E_{ij})^2}{E_{ij}} \sim \chi^2(r-1)(c-1)$$

O_{ij} : **관측도수** E_{ij} : **기대도수** r : **자유도**

7주: Descriptive Statistics

```
library(MASS)
data(survey)
tab<-xtabs(~Sex+Exer, data=survey) # 성별과 운동정도의 Cross Tabulation
chisq.test(tab)                   # 성별과 운동정도는 서로 독립인가?
```

- Fisher's Exact Test: 카이제곱 검정은 대중적으로 많이 사용하는 검정이지만 오차에 대한 정규근사 그리고 그의 제곱이 카이제곱분포를 따른다는 이론에 의해 만들어진 방법이다.
  이 방법은 셀(Cell)의 기대도수가 작을 때(일반적으로 5이하), 정확성이 결여되는 단점을 가지고 있다.
  이 때에는 정규근사를 이용하지 않고 초기하분포를 이용한 Fisher's Exact Test를사용한다.
  Fisher's Exact Test는 계산량이 많은 단점을 가지고 있으므로 큰 데이터를 사용하면 상당히 오랜 시간 계산을 할 수 있다.

```
tab<-xtabs(~W.Hnd+Clap, data=survey) # 글씨쓰는 손과 박수칠때 위에 오는 손이 독립인가?
chisq.test(tab)   # 계산이 부정확할 수 있다는 경고 메세지가 나옴
fisher.test(tab)
```

7주: Descriptive Statistics

- McNemar Test: 카이제곱 검정은 두개의 명목형 변수가 서로 독립인지 알아보았다면 맥네마 검정은 쌍으로 이루어진 두개의 결과에서 결과를 변화가 없는지, 있는지를 판단하는 검정이다.
  예를 들어, 벌금부과를 시작한 후, 안전벨트 착용자의 수, 유세 후의 지지자의 수와 같이 사건 전후의 값을 비교해야 하는 경우이다. 사건 전의 결과를 Test1, 사건 후의 결과를 Test2라고 하고 아래와 같은 결과를 얻었다.

	Test2 Pos.	Test2 Neg.	Row. Total
Test1 Pos.	a	b	a+b
Test1 Neg.	c	d	c+d
Col. Total	a+c	b+d	n

사건 전후의 결과가 변화 없다면 a+b=a+c, c+d=b+d가 성립하므로 b와 c가 유사하다면 사건이후, 변화가 없다고 볼 수 있다. b, c가 유사하다는 것은 b+c에서 b가 차지하는 값이 ½ 정도라는 의미이므로 확률변수 b는 아래의 분포를 따른다.

$$b \sim Bin(b+c, 0.5), E(b) = \frac{b+c}{2}, V(b) = \frac{b+c}{4} \qquad \frac{b-\frac{b+c}{2}}{\sqrt{\frac{b+c}{4}}} \sim N(0,1)$$

$$\left[\frac{b-\frac{b+c}{2}}{\sqrt{\frac{b+c}{4}}}\right]^2 \sim \chi^2(1) \quad \therefore \chi^2_M = \frac{(b-c)^2}{b+c} \sim \chi^2(1)$$

7주: Descriptive Statistics

```
Performance <-  matrix(c(794, 86, 150, 570),
                nrow = 2,
                dimnames = list("1st Survey" = c("Approve", "Disapprove"),
                                "2nd Survey" = c("Approve", "Disapprove")))
Performance
mcnemar.test(Performance)
```

- Shapiro Wilk Test: 데이터가 정규 분포를 따르는지 알아보는 방법이다.

```
shapiro.test(rnorm(1000))
```

$$W = \frac{(\mathbf{a}'\mathbf{y})^2}{S_n^2} = \frac{(\sum_{i=1}^n a_i y_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \tag{2.19}$$

where

$$\mathbf{a}' = (a_1, \ldots, a_n) = \frac{\mathbf{m}'V^{-1}}{(\mathbf{m}'V^{-1}V^{-1}\mathbf{m})^{\frac{1}{2}}}$$

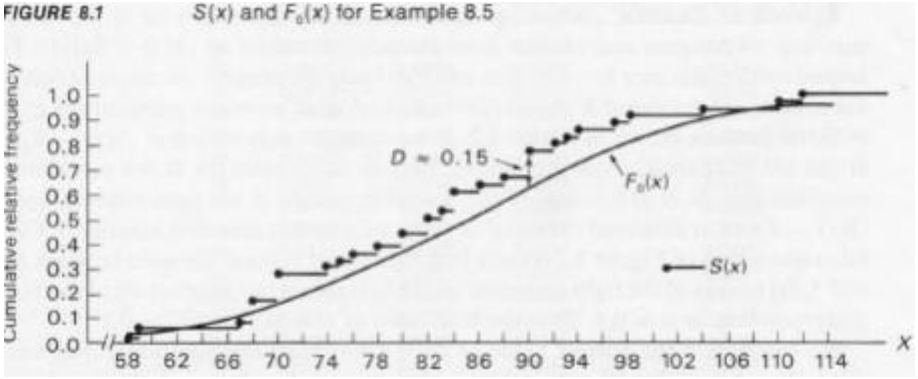
is called W **statistic**; *for the testing procedure we will use the abbreviation* **Shapiro Wilk test** *or just* W **test**.

7주: Descriptive Statistics

- Kolmogorov Smirnov Test: 데이터가 정규 분포를 따르는지 알아보는 비모수적 방법이다.

```
ks.test( rnorm(1000), "pnorm", 0, 1)
ks.test( runif(1000), "pnorm", 0, 1)
```

Kolmogorov Smirnov Goodness of Fit Test는 데이터로 부터 구한 Empirical CDF와 가정한 분포 (예를 들어 정규분포)와의 최대 차이값을 이용하여 이 값이 일정 수준이상이면 이 데이터가 정규분포에서 나오지 않았다고 생각하는 방법이다. 아래의 그림에서 부드러운 곡선이 정규분포의 CDF이고, Step Function이 데이터로 부터 구한 Empirical CDF이다. 그림에서 D=0.15는 두 선의 최대값이 0.15라는 의미이다.



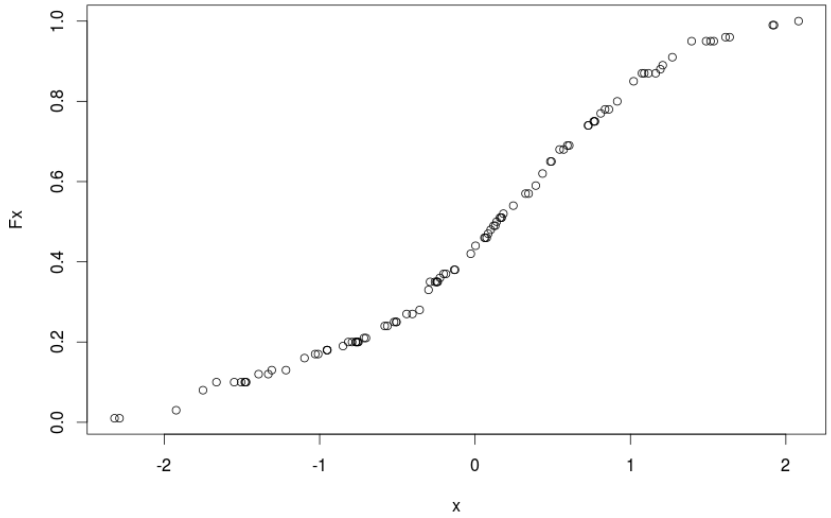
7주: Descriptive Statistics

- Empirical CDF: 각 분포의 이론적 CDF 처럼 데이터를 특정 분포에서 추출되었다고 가정하고 구한 경험적 분포함수이다.

$$\hat{F}(x) = \frac{1}{n}\sum 1_{X_i<x}$$

위의 식은 x가 주어졌을 때, 자료 중 x 보다 작은 수가 몇 % 인가를 구하는 식이다.

```
# ecdf  example :
x <- sort(rnorm(100))
Fx<-ecdf(x)
plot(Fx)
```



7주: Descriptive Statistics

- Q-Q Plot: Quantile-Quantile Plot의 약어로 주어진 자료가 특정 분포를 따르는지 판단하는 그림으로 가장 널리 사용되는 방법으로 그래프가 직선에 가까울수록 가정한 분포가 정확하다는 것을 의미한다.

```
x <- rnorm(100)
qqnorm(x)
abline(a=0, b=1)
```

Q-Q Plot 그리는 법

1. 관측값을 Sort 한다. $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$
2. Sort된 관측값의 Index를 생성한다. $\frac{i}{n+1}, i = 1, \cdots, n$
3. Index에 해당하는 이론적으로 검사하고자 하는 분포의 Inverse C.D.F 값을 구한다. 분포가 정규분포라면 $z_{k/(n+1)}$ 값이다.
4. 1.에서 구한 관측값과 3.에서 구한 역분포함수의 값을 Plotting 한다. $(z_{k/(n+1)}, x_{(k)})$

7주: Descriptive Statistics

- Pearson Correlation: 두개의 양적 자료에 대해 변수간의 선형 상관관계의 강도를 구한다.

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

$$\hat{\rho} = r = \frac{S_{xy}}{S_x S_y} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2} \sqrt{\sum (Y_i - \bar{Y})^2}}$$

```
with(data=iris,cor(Sepal.Length,Sepal.Width))
cor(iris[,1:4])
```

- Spearman rank Correlation: 이 방법은 두개의 양적 자료의 순위에 대한 선형 상관관계의 강도를 구하는 식으로 자료가 순서형이거나 자료의 수가 적은 경우, 극단값이 많은 경우에 적합하다.

$$r = 1 - \frac{6 \sum (x_i - y_i)^2}{n(n^2 - 1)}$$

```
with(data=iris,cor(Sepal.Length,Sepal.Width,method="spearman"))
```

7주: Descriptive Statistics

- 상관관계 검정: 두 변수의 모집단 상관계수가 "0"인지 여부를 가설검정한다.

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

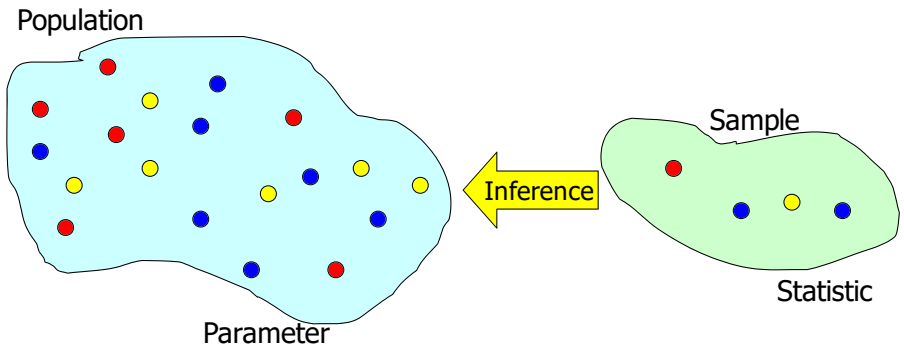
$$t = r\frac{\sqrt{n-2}}{\sqrt{1-r^2}} \sim t(n-2) \; under \; H_0$$

```
with(data=iris,cor.test(Sepal.Length,Sepal.Width))
```

숙제 6: 7장의 모든 과정을 실행하여 결과를 출력하고 설명하시오.

9주: Hypotheses Testing

- **Inferential Statistics**



우리가 수집한 Data는 표본(Sample)이다. 일반적으로 Data를 분석한다는 것은 Sample Data로 부터 Histogram이나 Bar chart를 그려 모집단의 분포를 추정하고, 표본평균, 최대, 최소값, 표준편차 등의 기술통계량을 통해 모집단의 Parameter를 추정하는 작업이다.

기술통계량으로부터 모집단의 모습을 Describe 할 수는 있지만, 우리가 구한 통계량이 어느 정도 정확한 값인지 알수는 없다. 모집단을 정확히 이해하려면 모집단의 분포를 찾아내고 분포에 맞는 Parameter를 추정해야만한다. 이 장에서는 우리가 알고 있는 통계량으로부터 모집단의 Parameter를 추정하는 문제에 대해 알아본다. 일반적으로 모수 θ를 추정의 추정량을 $\hat{\theta}$으로 표기한다.

9주: Hypotheses Testing

- 가설: 아직 확인되지 않은 명제
- 통계적 가설: 모집단에 대한 아직 확인 되지 않은 명제로 예를 들어, 두 집단의 평균은 같다 혹은 같지 않다, 모집단의 분산은 10보다 작다 등이 가설이다.
- 귀무가설과 대립가설: 연구자가 주장하고자 하는 가설을 대립가설(Alternative Hypothesis)이라 하고 대립가설을 반대 내용을 귀무가설(Null Hypothesis)이라 한다.
  일반적으로 귀무가설을 H_0, 대립가설은 H_1 으로 표시한다.
- 가설검정: 연구자가 자신이 주장하는 대립가설이 옳은지 여부를 논리적으로 증명해나가는 과정

판결 \ True	H_0: 범인(X)	H_1: 범인(O)
H_0 무죄	Right Decision	Type II Error
H_1 유죄	Type I Error	Right Decision

- 판사가 피고를 대상으로 검찰이 조사한 각종 증거와 피고의 진술, 증인의 진술 등을 토대로 범인여부를 가려 최종 판결을 한다. 여기서, 실제 범인이 아닌 사람을 유죄로 판결하는 오류를 1종 오류(Type I error)라고 하고, 실제 범인을 무죄로 판결하는 오류를 2종 오류(Type II error)라고 한다.
- 1종 오류의 최대 허용 확률을 α 로 쓰고 유의수준(Significant level)이라 부른다.
- 2종 오류의 확률은 β 로 쓰고, $1-\beta$를 검정력(Power of Test)라고 한다.

9주: Hypotheses Testing

- One Sample t-Test: 수집된 자료의 모평균에 대한 가설검정을 실시한다.

$$H_0 : \mu = 60, H_1 : \mu \neq 60$$

$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \frac{\bar{X}-60}{\sigma/\sqrt{n}} \sim N(0,1) \quad t = \frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{\bar{X}-60}{S/\sqrt{n}} \sim t(n-1)$$

```
x<-rnorm(30,60)
t.test(x,mu=60)
t.test(x,mu=30)
```

- Two Sample t-Test: 독립된 두 집단의 모평균이 서로 같은지에 대한 가설검정을 실시한다.

$$H_0 : \mu_1 = \mu_2, H_1 : \mu_1 \neq \mu_2$$

$$Z = \frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} = \frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

$$t = \frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \sim t(v), \quad v = \frac{\left(S_1^2/n_1+S_2^2/n_2\right)^2}{\frac{(S_1^2/n_1)^2}{n_1-1}+\frac{(S_2^2/n_2)^2}{n_2-1}}$$

$$t = \frac{\bar{X}_1-\bar{X}_2}{\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \sim t(v), \; v = n_1+n_2-2 \quad S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$$

9주: Hypotheses Testing

```
data(sleep)
head(sleep) # extra: 수면 증가량, group: 수면제 종류
var.test(extra~group,data=sleep)
t.test(extra~group,data=sleep,var.equal=TRUE)
```

▸ Paired Sample t-Test: Two sample t-Test에서 두 집단은 서로 독립이었다.
예를 들어, 한 집단에 어떤 Treatment를 하기 전과 후의 Data가 쌍으로 발생한다면 두 자료는 서로 독립이 아니기 때문에 Two sample t-Test를 수행할 수 없다. 이 경우, 자료의 형태는 아래와 같이 표현되므로 paired Sample t-Test는 아래의 자료에서 두 쌍의 자료의 차이에 대한 one sample t-Test가 된다.

$$\begin{matrix}(x_1,\ x_1+\delta_1)\\(x_2,\ x_2+\delta_2)\\(\vdots\qquad\vdots\quad)\\(x_n,\ x_n+\delta_n)\end{matrix}$$

$$H_0: \mu_\delta = 0,\ H_1: \mu_\delta > 0$$
$$H_0: \mu_\delta = 0,\ H_1: \mu_\delta < 0$$
$$H_0: \mu_\delta = 0,\ H_1: \mu_\delta \neq 0$$

숙제 7: 가족 경영기업의 새로운 CEO 효과(XM13-02.xls)
사장의 아들, 딸이 새로운 CEO를 하는 경우와 그렇지 않은 경우의 영업이익률을 비교하여 두 경우 회사의 영업이익률이 다르다고 할 수 있는가?

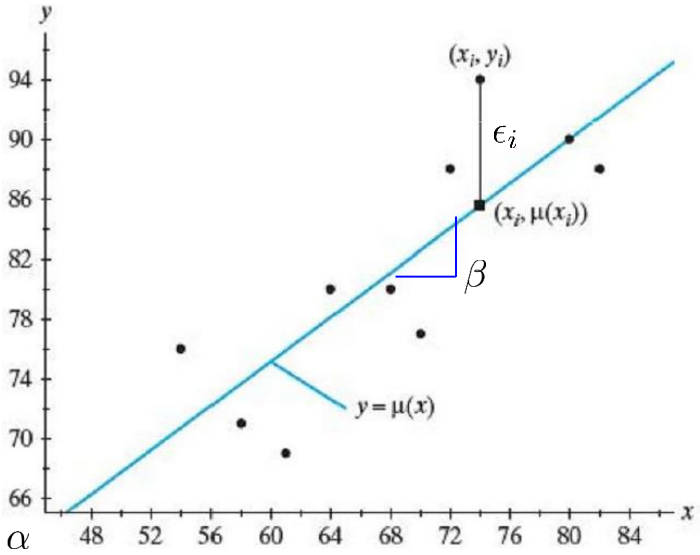
10주: Linear Regression

- **단순선형 회귀모형(Simple Linear Regression Model)**

  종속변수 Y를 설명하는 변수가 X 하나인 경우의 확률적 모형이다.

  예를 들어, 주택의 가격(Y)를 주택의 크기(X)로 설명하는 모형이다.

$$Y_i = \alpha + \beta x_i + \epsilon_i, \epsilon_i \sim N(0, \sigma^2),\ i = 1, \cdots, n$$



여기서, 오차제곱합 SSE가 최소화되도록 선을 그으면 그 선이 바로 단순선형 회귀모형 식이 된다.

$$SSE = \sum(\varepsilon_i{}^2) = \sum(Y_i - \alpha - \beta x_i)^2$$

10주: Linear Regression

- Least Square Estimation for α, β

$$\frac{\partial SSE}{\partial \alpha} = 2\sum(Y_i - \alpha - \beta x_i)(-1) = 0 \cdots\cdots\cdots ①$$

$$\frac{\partial SSE}{\partial \beta} = 2\sum(Y_i - \alpha - \beta x_i)(-x_i) = 0 \cdots\cdots\cdots ②$$

①× $\bar{x}$ 를 하면,

$$\alpha\sum x_i + \beta\bar{x}\sum x_i = n\bar{x}\bar{Y} \cdots\cdots\cdots\cdots ③$$

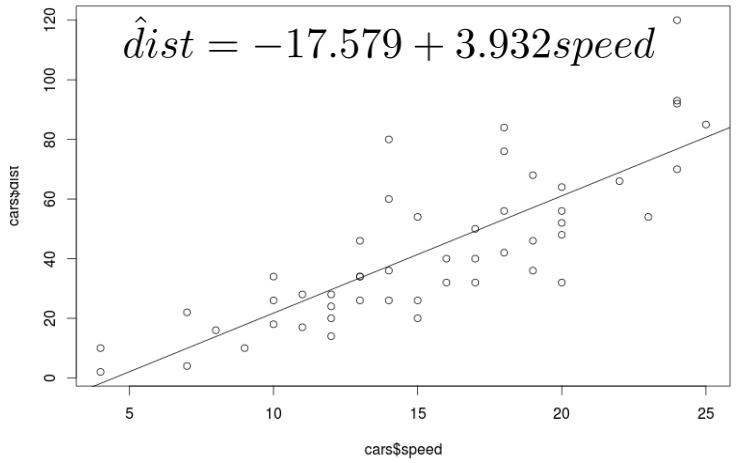
$$\alpha\sum x_i + \beta\sum {x_i}^2 = \sum x_i Y_i \cdots\cdots\cdots ④$$

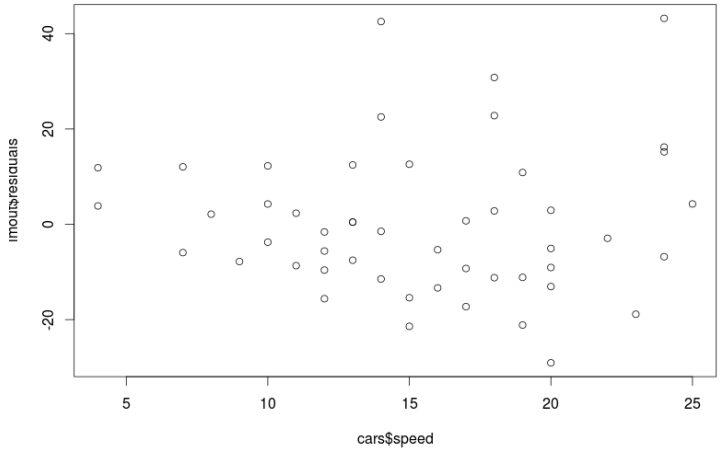
④-③을 하면,

$$\hat{\beta} = \frac{\sum x_i Y_i - n\bar{x}\bar{Y}}{\sum {x_i}^2 - n\bar{x}^2} = \frac{\sum(x_i - \bar{x})(Y_i - \bar{Y})}{\sum(x_i - \bar{x})^2}, \qquad \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}$$

10주: Linear Regression

```
data(cars)
head(cars)
plot(cars$speed,cars$dist)
y<-cars$dist
x<-cars$speed
lmout<-lm(y~x)
abline(lmout)
lmout
plot(cars$speed,lmout$residuals)
summary(lmout)
anova(lmout)
```





10주: Linear Regression

```
lm(formula = cars$dist ~ cars$speed)

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 *
cars$speed     3.9324     0.4155   9.464    1.49e-12 ***

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12


Analysis of Variance Table

Response: cars$dist
              Df  Sum Sq  Mean Sq  F value        Pr(>F)
cars$speed     1   21186  21185.5   89.567     1.49e-12 ***
Residuals     48   11354    236.5
```

10주: Linear Regression

$$\sum(Y_i - \bar{Y})^2 = \sum(Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2$$
$$= \sum(Y_i - \hat{Y}_i)^2 + \sum(\hat{Y}_i - \bar{Y})^2 + 2\sum(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y})$$
$$= \sum(Y_i - \hat{Y}_i)^2 + \sum(\hat{Y}_i - \bar{Y})^2$$

$$SST = SSE + SSR, R^2 = \frac{SSR}{SST}$$

$$SST = \sum(Y_i - \bar{Y})^2 \quad SSE = \sum(Y_i - \hat{Y}_i)^2 \quad SSR = \sum(\hat{Y}_i - \bar{Y})^2$$

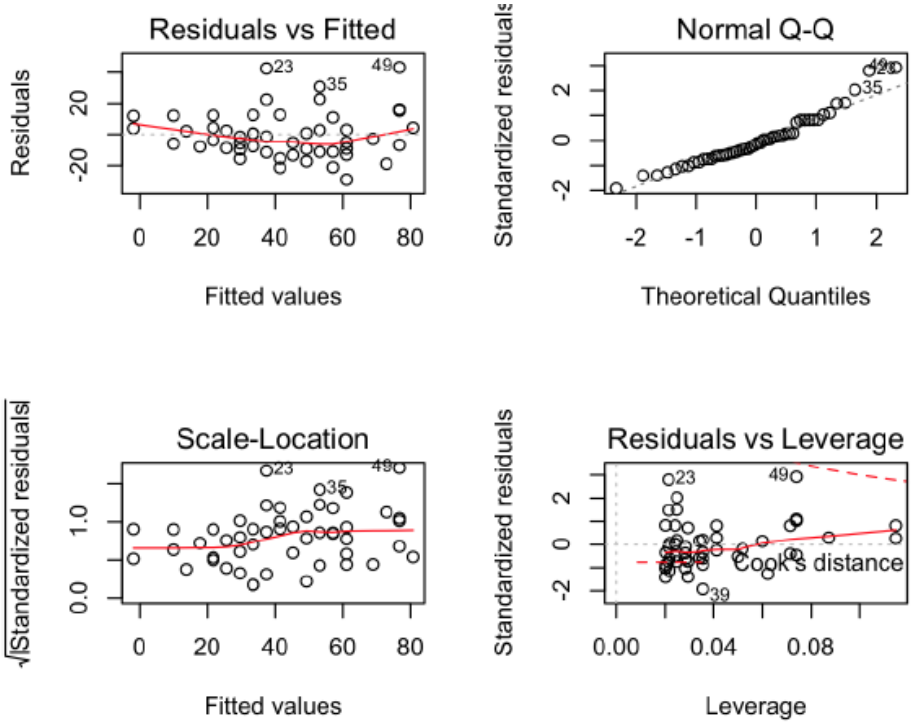
$$\sum(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y})$$
$$= \sum \hat{Y}_i e_i - \bar{Y}\sum e_i = \sum \hat{Y}_i e_i = \sum(\hat{\beta}_0 + \hat{\beta}_1 x_i)e_i = \hat{\beta}_0 \sum e_i + \hat{\beta}_1 \sum x_i e_i$$
$$= \hat{\beta}_1 \sum x_i(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = \hat{\beta}_1(\sum x_i Y_i - \hat{\beta}_0 \sum x_i - \hat{\beta}_1 \sum x_i^2) = 0$$

10주: Linear Regression

```
confint(lmout) # Confidence Interval for beta
deviance(lmout) # SSE
predict(lmout, data.frame(x=3)) # predict value at x=3
plot(lmout)
plot(lmout, which=c(4,6))
```

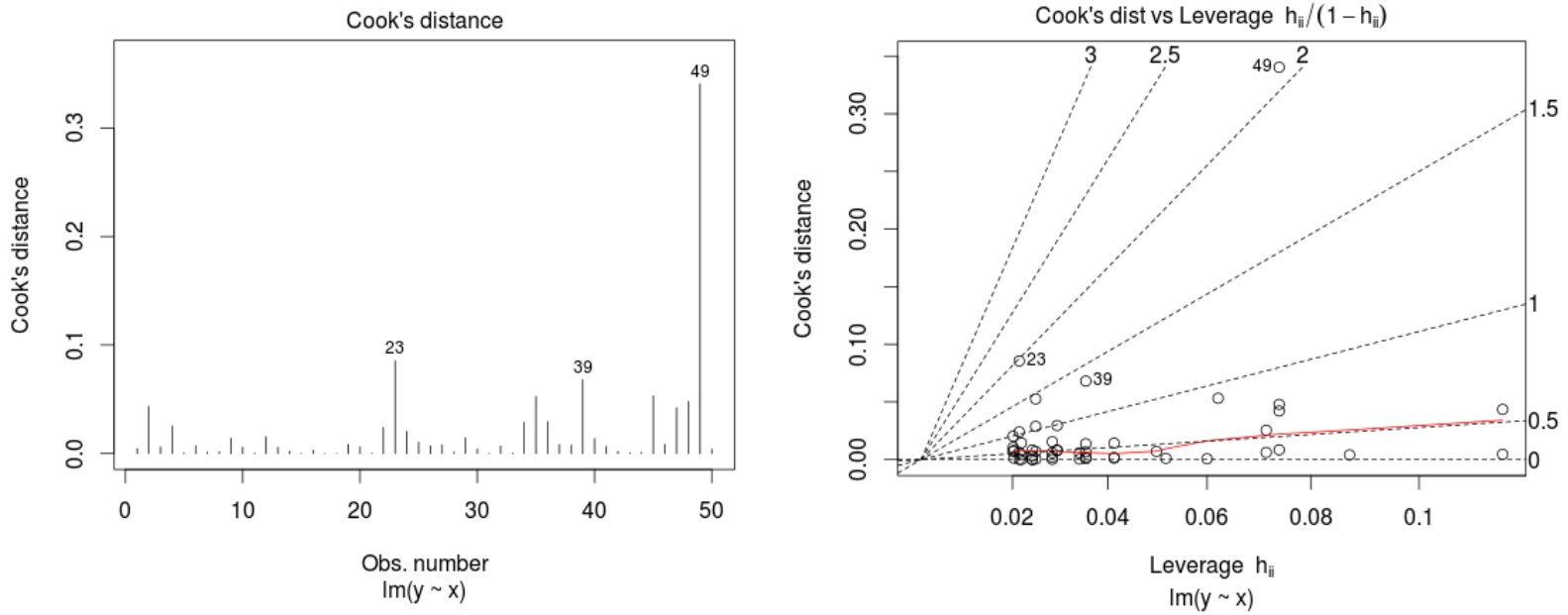
회귀분석에서 모형구축이 잘 이루어졌는가를 평가는 방법으로 R-Sqaure와 잔차가 규칙성을 가지는지를 판단하는 방법이 있다. R-Square는 데이터의 변동 중 모형이 설명하는 량으로 0~100% 사이의 값이고 이 값이 크면 클수록 모형이 데이터를 잘 설명한다고 볼 수 있다.

또한, 잔차는 Pure Random의 성질을 가져야 한다. 우측 그림은 일단, 잔차의 모양이 2차식으로 보이고 분산역시 커지는 모양으로 규칙성이 존재함을 알 수 있다.



10주: Linear Regression

Leverage는 지랫대라는 뜻으로 독립변수의 값들 중 극단값을 보여주고 이 값의 잔차를 보여주고 있다. 이 결과에서는 Leverage가 큰 값이 큰 잔차를 가지지 않는다. 그 뜻은 몇개의 극단값이 모형에 영향을 많이 주고 있음(Large Influenced Observation)을 의미한다. 그러므로, 이 관측값이 정확한 것인지를 파악할 필요가 있다.



Cook's distance는 Large Influenced Observation을 찾아주는 그림으로 23, 39, 49 번째 관측값이 모형을 영향을 많이 주는 값이라는 것을 보여주고 있고, 49번은 Leverage도 큰 값임을 알 수 있다.

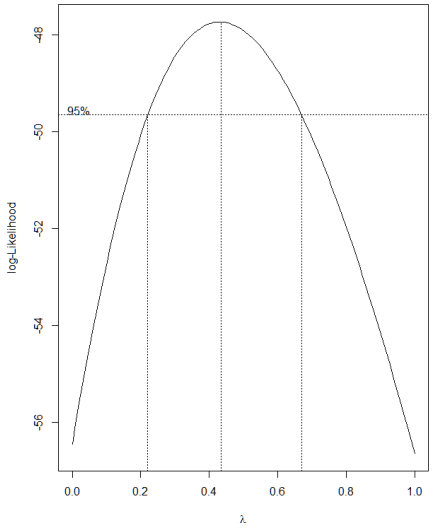
10주: Linear Regression

- 변수변환: 선형 회귀모형은 두개의 변수가 선형 관계에 있어야 한다.
  만약, 그렇지 않다면 적절한 변수변환을 할 수 있다. 예를 들어, 독립변수, 혹은 종속변수에
  log, sqrt, exp, sin, cos 등의 함수를 취해서 최대한 종속변수와 독립변수가 선형성을 가지도록 하는 것이다.
  이 때, 일반적으로 하는 방법으로 Box-Cox Transformation이 있다.

Box-Cox Transformation

$$y_i(\lambda) = \frac{y_i^{\lambda}-1}{\lambda} \text{ for } \lambda \neq 0,\ y_i(\lambda) = log(y_i) \text{ for } \lambda = 0$$

```
data(cars)
View(cars)
plot(cars$speed, cars$dist)
library(MASS)
boxcox(lm(dist~speed,data=cars),lambda=seq(0,1,by=.1))

cars$dist1<-(cars$dist^0.5 - 1)/0.5
plot(cars$speed, cars$dist1)

summary(lm(dist~speed, data=cars))
summary(lm(dist1~speed, data=cars))
```



11주: Multiple Linear Regression

종속변수 Y를 설명하는 독립변수가 k개 존재할 때의 회귀분석이다.

여기서, 종속변수와 독립변수는 반드시 양적 변수여야 한다.

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i, \epsilon_i \sim N(0, \sigma^2),\ i = 1, \cdots, n$$

$$\hat{\beta} = (X'X)^{-1}X'Y, \qquad X = \begin{pmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{pmatrix}, Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

$$SE(\hat{\beta}) = \sqrt{MSE \cdot C_{jj}}, C_{jj} \text{is the j'th diagonal element of the } (X'X)^{-1}$$

```
data(iris)
lmout<-lm(Sepal.Length~Sepal.Width+Petal.Length+Petal.Width, data=iris)
summary(lmout)
```

$$\hat{S.L} = 1.856 + 0.65S.W + 0.71P.L - 0.56P.W$$

11주: Multiple Linear Regression

독립변수 중에 범주형 변수가 포함되어 있다면, Dummy Variable을 사용하여 회귀분석을 진행할 수 있다.
iris data에서 Species는 "setosa", "versicolor","virginica"로 세가지 범주를 가지는 변수이다.
이 때에는 아래와 같이 Dummy Variable을 사용하여 모형을 세울 수 있다.

$$D_{1i} = 0, D_{2i} = 0 \text{ for setosa}$$
$$D_{1i} = 1, D_{2i} = 0 \text{ for versicolor}$$
$$D_{1i} = 0, D_{2i} = 1 \text{ for virginica}$$

$$S.L = \beta_0 + \beta_1 S.W + \beta_2 P.L + \beta_3 P.W + \beta_4 D_1 + \beta_5 D_2 + \epsilon$$

$$S.L = \beta_0 + \beta_1 S.W + \beta_2 P.L + \beta_3 P.W + \epsilon \text{ for setosa}$$
$$S.L = \beta_0 + \beta_1 S.W + \beta_2 P.L + \beta_3 P.W + \beta_4 D_1 + \epsilon \text{ for versicolor}$$
$$S.L = \beta_0 + \beta_1 S.W + \beta_2 P.L + \beta_3 P.W + \beta_4 D_2 + \epsilon \text{ for virginica}$$

```
lmout1<-lm(Sepal.Length~., data=iris)
summary(lmout1)
```

이 방법은 Species에 따라 절편만 다른 3개의 모형을 구하는 방법이다. 절편 이외에 기울기가 다른 모형을 구할 수 도 있다.

11주: Multiple Linear Regression

잔차의 검토

단순선형회귀모형에서는 산점도를 통해 쉽게 X, Y의 관계를 알 수 있고 회귀선이 데이터를 잘 설명하는지를 육안으로 확인 가능하지만, 독립변수가 여러 개이면 산점도를 통해 회귀분석이 잘되었는지를 알 수 없다.
회귀분석이 잘되었는지 확인하는 방법은 잔차가 오차의 성질을 만족하는지를 검사하는 것이다.
회귀분석의 기본 가정을 살펴보자.

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i, \epsilon_i \sim N(0, \sigma^2),\ i = 1, \cdots, n$$

(1) 오차의 평균은 "0"이다. (2) 오차들은 서로 독립이다.(독립성)
(3) 오차의 분산은 일정하다.(등분산성) (4)오차의 분포는 정규분포를 따른다.(정규성)

오차의 등분산성 검사

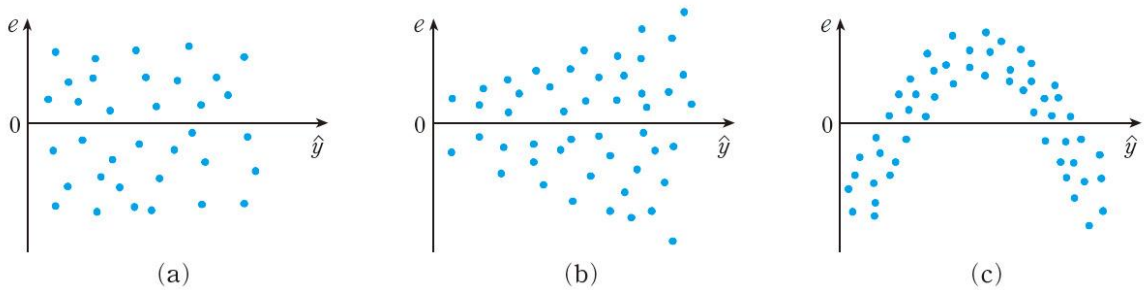


그림 9 잔차와 예측값의 산점도

(a)는 등분산성을 만족하는 것이고, (b), (C)는 등분산성을 만족하지 않는 그림이다. 이 경우, 가중회귀분석, 곡선회귀등의 방법을 써야 한다.

11주: Multiple Linear Regression

오차의 독립성 검사

오차가 독립인지를 검사하는 가장 간단한 방법은 (e_i, e_{i-1}) 의 산점도를 그리는 것이다.

다른 방법으로 Durbin-Watson 검정이 있다. 더빈 왓슨 검정은 1차 자기상관을 검사하는 방법으로 아래의 통계량을 구해 이 값이 속한 구간에 따라 자기상관 존재 여부를 결정하는 방법이다.

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

Q-Q Plot

어떤 자료가 특정한 분포를 따르는지를 알아보는 방법이 Q-Q Plot이다. 수많은 분포 중 특히, 정규분포를 따르는지 여부를 알 수 있는 방법으로 히스토그램을 그려 확인하는 방법과 Q-Q Plot을 그려보는 방법이 있다. 정규확률그림은 $(z_{k/(n+1)}, x_{(k)})$에 대한 산점도를 그리면 된다.

이 그림은 자료가 정규분포에 따른다면 직선을 보일 것이다.

```
hist(lmout1$residuals)
qqnorm(lmout1$residuals)
plot(lmout1$residuals,iris$Sepal.Width)
plot(lmout1$residuals,iris$Petal.Length)
plot(lmout1$residuals,iris$Petal.Width)
library(lmtest)
dwtest(Sepal.Length~Sepal.Width+Petal.Length+Petal.Width, data=iris)
```

11주: Multiple Linear Regression

• **상호작용(Interaction)**

모형에서 상호작용 혹은 교호작용이라는 것은 변수들간의 양의 시너지 혹은 음의 시너지를 말하는 것이다.

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$$

위의 식에서 각 독립변수가 추가될 때, 해당 독립변수의 효과가 더해지는 구조이다.

아래의 식은 두 변수의 조합된 상황에서 시너지가 나타나는 것을 모형으로 표현한 것이다.

아래의 모형에서 $H_0 : \beta_3 = 0, H_1 : \beta_3 \neq 0$ 의 가설을 검정해 귀무가설이 기각된다면

시너지가 존재한다고 결론내릴 수 있다.

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \epsilon_i$$

```
data(Orange)
head(Orange)
with ( Orange ,plot (Tree , circumference , xlab =" tree ", ylab =" circumference "))
with ( Orange , interaction.plot (age , Tree , circumference ))
lmout1<-lm(circumference~factor(Tree)+age, data=Orange)
anova(lmout1)
lmout2<-lm(circumference~factor(Tree)*age, data=Orange)
anova(lmout2)
```

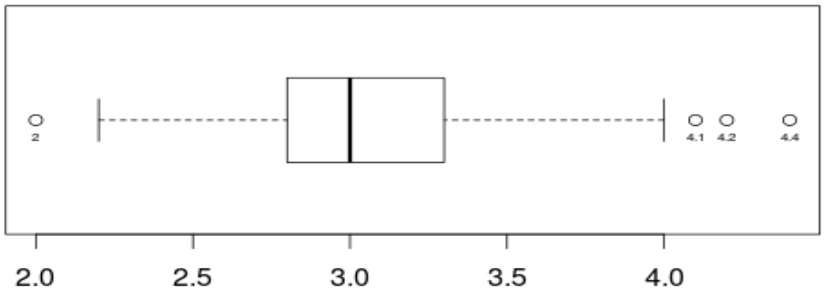
11주: Multiple Linear Regression

• **이상값(Outlier)**

Data set에 이상값이 존재하는 경우가 많다. 이상값은 분석에 아주 소중한 정보일수도 있고, 분석을 방해하는 자료일 수도 있다. 문제는 이상값이 소중한 정보로 Keep해야 하는지, 분석에 방해가 되는 것으로 버려야 하는지 알수는 없다. 다만, 통계적 방법을 통해 이상값을 찾고 이 이상값에 의해 분석이 어떤 영향을 받는지를 알 수는 있다. 한개의 변수에서는 Boxplot을 이용하여 이상값을 검출할 수 있고, 회귀분석에서는 Plot을 통해 이상값을 검출하거나 표준화 잔차의 절대값이 2를 초과하는 관측값을 이상값으로 정의할 수 있다.

```
lmout<-lm(Sepal.Length~Sepal.Width+Petal.Length+Petal.Width, data=iris)
plot(rstudent(lmout))
abline(a=2, b=0)
abline(a=-2, b=0)
identify (rstudent(lmout))
```

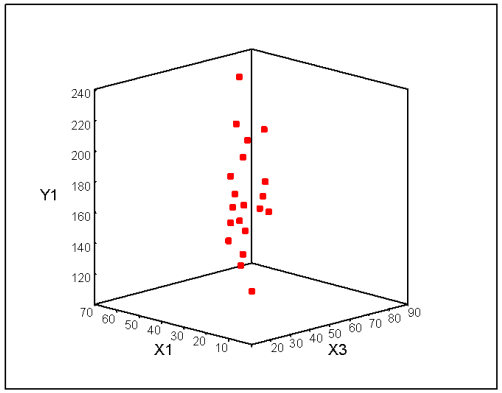
identify 함수로 85, 107, 135, 136, 142번째 관측값이 이상값으로 나타남을 알 수 있다.

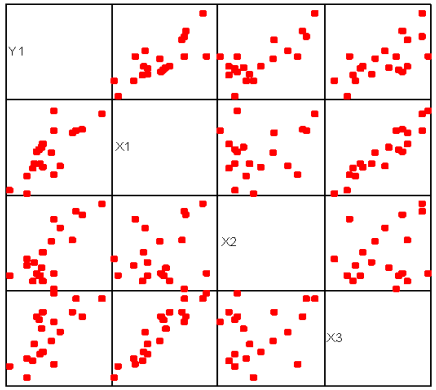


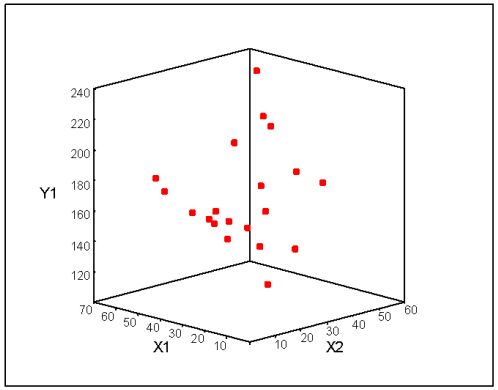
11주: Multiple Linear Regression

• **다중공선성(Multicollinearlity)**

다중공선성은 독립변수들간에 선형관계가 존재하는 상황의 이슈이다. 독립변수들끼리 상관관계가 존재하면 회귀계수의 추정에 대한 분산이 커지는 문제가 발생한다. 이러한 문제로 회귀계수가 참값과 동떨어진 값을 가질 확률이 커지는 문제가 발생할 수 있다. 심한 경우에는 회귀계수의 부호가 상식과 반대로 나오는 경우도 발생한다. 이를 막기 위해서는 선형관계가 존재하는 독립변수가 존재하는지 검사하고 불필요한 변수는 Forward Selection, Backward Elimination, Stepwise Regression 방법을 사용할 수 있다.







다중공선성을 판단하는 통계량으로 V.I.F.(Variance Inflation Factors)가 널리 사용된다.

V.I.F.는 j번째 독립변수를 종속변수로 하고 나머지를 독립변수로 하는 회귀분석을 수행한 모형의결정계수를 R_j^2으로 표현할 때, j번째 독립변수에 대한 VIF는 $VIF_j = \frac{1}{1-R_j^2}$ 과 같이 표현되고 이 식이 일반적으로 10보다 크면 다중공선성이 존재한다고 판단한다.

11주: Multiple Linear Regression

독립변수간에 상관관계가 크면 회귀계수 추정 오차가 커지는데 이는 아래의 식의 값이 커짐을 말한다.

$$SE(\hat{\beta}) = \sqrt{MSE \cdot C_{jj}}, C_{jj} \text{is the j'th diagonal element of the } (X'X)^{-1}$$

왜냐하면, C값은 (X'X)의 역행렬의 대각선 값을 의미하고 이 값은 아래의 예 처럼 (X'X)의 |X'X|를 구해야 하는데 분모인 |X'X| 값이 너무 작아져 "0"에 가까워지기 때문이다. 그러므로, C 값은 커져 회귀계수의 추정오차가 커지는 것이다.

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

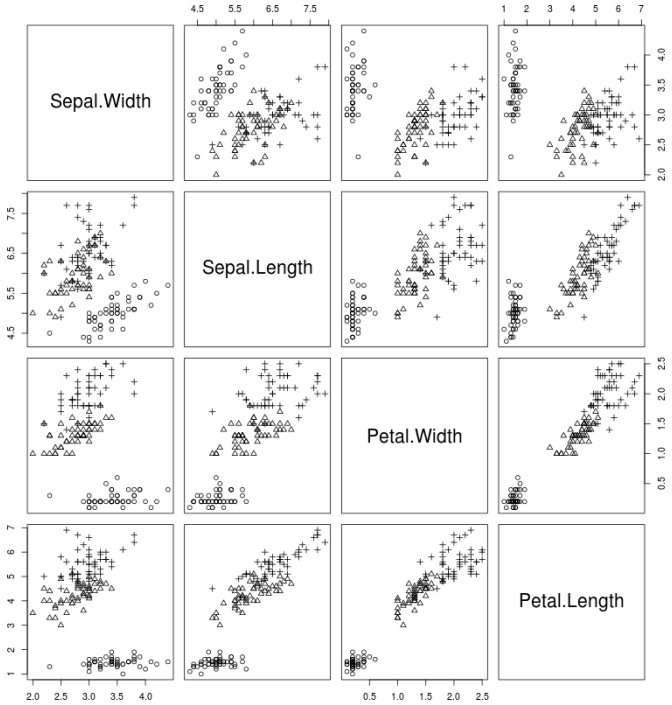
There exists an inverse matrix of A when
$\det A = a_{11}a_{22}a_{33} + a_{21}a_{32}a_{13} + a_{31}a_{12}a_{23} - a_{11}a_{32}a_{23} - a_{31}a_{22}a_{13} - a_{21}a_{12}a_{33} \neq 0$, and it is

$$\mathbf{A}^{-1} = \frac{1}{\det \mathbf{A}} \begin{pmatrix} a_{22}a_{33} - a_{23}a_{32} & a_{13}a_{32} - a_{12}a_{33} & a_{12}a_{23} - a_{13}a_{22} \\ a_{23}a_{31} - a_{21}a_{33} & a_{11}a_{33} - a_{13}a_{31} & a_{13}a_{21} - a_{11}a_{23} \\ a_{21}a_{32} - a_{22}a_{31} & a_{12}a_{31} - a_{11}a_{32} & a_{11}a_{22} - a_{12}a_{21} \end{pmatrix}$$

11주: Multiple Linear Regression

```
lmout<-lm(Sepal.Length~Sepal.Width+Petal.Length+Petal.Width, data=iris)
library(car)
vif(lmout)
Sepal.Width Petal.Length Petal.Width 1.270815 15.097572 14.234335
```

위의 결과로 볼 때, Petal.Length, Petal.Width는 다중공선성의 위험이 큰 변수임을 알 수 있다.



11주: Multiple Linear Regression

• 변수선택법(Variable Selection)

모형 개발 시, 여러 개의 독립변수 중에 최적으로 모형에 필요한 변수만을 골라야 하는 경우가 있다.
이를 위해서는 다양한 독립변수의 조합에 대해 회귀분석을 실시하고 이 중 최적의 모형을 선택한다.
여기서, 최적의 모형이란 (1) 해석의 용이성, (2) 높은 설명력, (3) 최소의 독립변수 수를 만족하는 모형이다.

1. 전진선택(Forward Selection)

이 방법은 여러 개의 독립변수 중 영향이 큰 변수들만으로 모형을 만드는 방법입니다. 독립변수 중 영향력이 가장 큰 변수를 선택하고 모형이 의미 있는 것인가를 확인한 후, 만약 현재 모형이 의미가 있다면 2번째로 영향이 큰 변수를 선택하여 2개의 변수로 모형을 다시 만들어 모형의 의미를 확인하는 방식으로 계속해서, 영향력이 큰 변수부터 하나씩 모형에 포함시켜 가면서 모형을 만들어가는 방법이다. 모형이 의미가 있다는 것은 추가적인 변수로 인해 설명력이 증가하고 그 변수 역시 의미 있는 변수로 확인됨을 의미한다.

2. 후방제거(Backward Elimination)

이것은 전진선택의 반대되는 방법으로 처음에 모든 변수들로 모형을 만든 다음 그 중에서 영향력이 가장 작은 변수를 모형에서 제거하여 모형을 만들고 제거하기 전, 후의 설명력의 차이를 검사하여 이 차이가 통계적으로 큰 차이가 없다고 판단하면 그 다음으로 영향력이 작은 변수를 제거는 방식으로 마지막에 남은 변수들은 모두 영향력이 큰 변수들만이 남게 되는 방법이다.

3. 단계별회귀(Stepwise Selection)

전진선택에서는 한번 선택된 변수는 모형의 끝까지 남게 되고, 후방제거는 한번 제거된 변수는 끝의 모형에서도 선택되지 않게 된다는 단점을 보완하는 방법으로 전진선택과 후방제거를 교차로 수행해 나가는 방식이다.

11주: Multiple Linear Regression

```
lmoutF <- step (lmout, direction ="forward")
lmoutB <- step (lmout, direction ="backward")
lmoutS <- step (lmout, direction ="both")
formula(lmoutF)
formula(lmoutB)
formula(lmoutS)
```

- **Akaike information criterion**

$$\mathrm{AIC} = 2k - 2\ln(L)$$

Suppose that we have a statistical model of some data. Let L be the maximum value of the likelihood function for the model; let k be the number of estimatedparameters in the model. Then the AIC value of the model is the following.[1][2]

Given a set of candidate models for the data, *the preferred model is the one with the minimum AIC value.* Hence AIC rewards goodness of fit (as assessed by the likelihood function), but it also includes a penalty that is an increasing function of the number of estimated parameters. The penalty discourages overfitting(increasing the number of parameters in the model almost always improves the goodness of the fit).

모형 선택시, AIC, BIC 값을 최소화시켜주는 모형을 선택하는 것도 하나의 기준이다.

위의 식에서 모형이 복잡해질수록 k가 증가하여 AIC 값은 커지고, 모형이 잘 적합될 수록 ln(L) 값이 커져 AIC 값이 작아진다. 그러므로, 최소 AIC는 복잡도가 낮으면서 동시에 모형의 적합성이 좋은 모형에서 나온다.

11주: Multiple Linear Regression

숙제 8: GSS2008 Data를 이용하여 소득(Income)에 대한 모형을 개발해 발표하세요

변수 설명

연령(Age), 교육년수(Educ), 주당 근로시간(Hrs), 배우자 주당 근로시간(SPHrs), 직업평판점수(Prestg80), 자녀의 수(Childs), 돈을 버는 가족의 수(Earnrs), 현직장 고용년수(CurEmpYr)

12주: Logistic Regression

회귀분석에서 종속변수는 반드시 양적 변수여야 한다.

하지만, 실제 문제에서 종속변수가 Binary 혹은 3개 이상의 범주를 가지는 경우가 많이 있다.

Binary Case: 기업 부도, 개인 신용 상태, 정상/비정상 접속 등 ...

로지스틱 회귀분석은 종속변수가 범주형 자료인 경우의 회귀분석법이다.

아래의 모형에서 우측은 $(-\infty \sim \infty)$ 의 범위인 반면 좌변 Y=0, 1 두가지 값을 가지므로 모형이 적절치 못하다.

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i$$

이 식에서 좌변과 우변의 범위를 맞추기 위해 아래와 같이 변형 해보자.

$$ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}$$

여기서, $p_i = P(Y_i = 1)$ 이라 하면 이 값은 0~1 사이에 있고, $\frac{p_i}{1-p_i}$ 은 $(0 \sim \infty)$ 사이에 있으므로 최종적으로 좌변은 $(-\infty \sim \infty)$ 가 성립한다.

$\frac{p_i}{1-p_i}$ 은 odds Ratio 라고 부른다. 위의 로지스틱 회귀식을 풀면 아래와 같이 표현 가능하다.

$$\hat{p}_i = \frac{exp\left[\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_k x_{ki}\right]}{1 + exp\left[\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_k x_{ki}\right]}$$

12주: Logistic Regression

```
data(iris)
iris2<-subset(iris, Species!="setosa")
iris2$Species<-factor(iris2$Species)
lout<-glm(Species~., family="binomial", data=iris2)
summary(lout)
```

```
Call:
glm(formula = Species ~ ., family = "binomial", data = iris2)
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.01105  -0.00541  -0.00001   0.00677   1.78065
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -42.638     25.707  -1.659   0.0972 .
Sepal.Length   -2.465      2.394  -1.030   0.3032
Sepal.Width    -6.681      4.480  -1.491   0.1359
Petal.Length    9.429      4.737   1.991   0.0465 *
Petal.Width    18.286      9.743   1.877   0.0605 .
Null deviance: 138.629  on 99  degrees of freedom
Residual deviance:  11.899  on 95  degrees of freedom
AIC: 21.899

Number of Fisher Scoring iterations: 10
```

$$\hat{p} = \frac{exp(-42.64-2.47SL-6.68SW+9.43PL+18.29PW)}{1+exp(-42.64-2.47SL-6.68SW+9.43PL+18.29PW)}$$

12주: Logistic Regression

```
lout1<-glm(Species~., family="binomial", data=iris2)
summary(lout1)
lout2<-glm(Species~Petal.Length+Petal.Width, family="binomial", data=iris2)
summary(lout2)
iris2$res1<-ifelse(lout1$fitted.values > 0.5, "virginica","versicolor")
iris2$res2<-ifelse(lout2$fitted.values > 0.5, "virginica","versicolor")
xtabs(~res1+Species,data=iris2)
xtabs(~res2+Species,data=iris2)
```

```
> xtabs(~res1+Species,data=iris2)
            Species
res1         versicolor virginica
  versicolor         49         1
  virginica           1        49

> xtabs(~res2+Species,data=iris2)
            Species
res2         versicolor virginica
  versicolor         47         3
  virginica           3        47
```

lout1 모형이 오분류 확률이 작음을 알 수 있다.

12주: Logistic Regression

로지스틱 회귀모형에서 범주가 3개 이상인 경우에는 모형을 아래와 같이 변환할 수 있다.

$$ln\frac{P(Y_i=1)}{P(Y_i=k)} = \beta_0 + \beta_1 X_i \qquad p_1 = p_k \cdot exp(\beta_0 + \beta_1 X_i)$$

$$ln\frac{P(Y_i=2)}{P(Y_i=k)} = \gamma_0 + \gamma_1 X_i \qquad p_2 = p_k \cdot exp(\gamma_0 + \gamma_1 X_i)$$

$$\vdots \qquad\qquad \vdots$$

$$ln\frac{P(Y_i=2)}{P(Y_i=k)} = \delta_0 + \delta_1 X_i \qquad p_{k-1} = p_k \cdot exp(\delta_0 + \delta_1 X_i)$$

위의 식에서 $\sum p_i = 1$ 이므로 $p_k = \frac{1}{1+exp(\beta_0+\beta_1 X_i)+exp(\gamma_0+\gamma_1 X_i)+\cdots+exp(\delta_0+\delta_1 X_i)}$ 이 된다.

최종적으로 $p_1 = \frac{exp(\beta_0+\beta_1 X_i)}{1+exp(\beta_0+\beta_1 X_i)+exp(\gamma_0+\gamma_1 X_i)+\cdots+exp(\delta_0+\delta_1 X_i)}$

$$p_1 = \frac{exp(\gamma_0+\gamma_1 X_i)}{1+exp(\beta_0+\beta_1 X_i)+exp(\gamma_0+\gamma_1 X_i)+\cdots+exp(\delta_0+\delta_1 X_i)}$$

$$\vdots$$

$$p_{k-1} = \frac{exp(\delta_0+\delta_1 X_i)}{1+exp(\beta_0+\beta_1 X_i)+exp(\gamma_0+\gamma_1 X_i)+\cdots+exp(\delta_0+\delta_1 X_i)}$$ 이 된다.

12주: Logistic Regression

```
library(nnet)
lout3<-multinom(Species~., data=iris)
summary(lout3)
head(lout3$fitted.values)
res<-as.data.frame(lout3$fitted.values)
res$Species<-iris$Species
res$dec<-””
res$dec<-ifelse(res$setosa>0.33, "setosa",res$dec)
res$dec<-ifelse(res$versicolor>0.33, "versicolor",res$dec)
res$dec<-ifelse(res$virginica>0.33, "virginica",res$dec)
xtabs(~Species+dec, data=res)
```

```
> xtabs(~Species+dec, data=res)
            dec
Species      setosa versicolor virginica
  setosa         50          0         0
  versicolor      0         48         2
  virginica       0          1        49
```

숙제 9: 위의 코드에서 아래와 같이 새로 발견된 꽃에 대해 종을 판별해 보시오.

S.L=5.7	S.W=1.6	P.L=2.4	P.W=2.2

13주: 재밌는 R 실습

- **WordCloud**

```
library(KoNLP)
library(RColorBrewer)
library(wordcloud)
useSejongDic()
txt<-readLines("d:/RNote/bigdata.txt")
txtNoun<-sapply(txt,extractNoun,USE.NAMES=F)
c <- unlist(txtNoun)
txtNoun <- Filter(function(x) {nchar(x) >= 2} ,c)
write.table(txtNoun,"d:/RNote/noun.csv", sep=",", quote=F,row.names=F, col.names=T)
noun<-read.table("d:/RNote/noun.csv", header=TRUE)
wordcount <- sort(table(noun),decreasing=T)
palete <- brewer.pal(9,"Set1")
png(filename="d:/RNote/bigdata.png", height=500, width=500)
wordcloud(names(wordcount),freq=wordcount,scale=c(5,1),rot.per=0.25,min.freq=2,
random.order=F,random.color=T,colors=palete)
dev.off()
```

숙제 10 : 관심 주제어를 정하고 WordCloud 그림을 작성해 제출하세요.

13주: 재밌는 R 실습


데이터
자동차
활용
과정
미래
시스템
실력
수집
자율
직업
예측
때문
개인정보보호
개념
러닝
주행
분야
하게
경로
유망
이것
컴퓨터
IT
Java
빅데이터를
빅데이터교육은
이미지를
빅데이터전문가는
빅데이터가
빅데이터전문가
전문가
특성
메신저
서버
업체
그림
3D
지역
전기
등등중요
프로그램
정리
추천
인공
자극
판단
기반
가지
영역
메시지
뉴런
전망
이해
기본적
소비
경쟁적
인식
방식
해서
결정
위배
생각
인력
통계
영향
'동물'이라는
구축
평소
개발
관련
다양
구매
스티커
하기
흥내
대상
학문
무엇
취향
인지
특정
경우
역할
동물
사용
전문
주변
링크
상상
강화
입장
사람
진행
채용
등장
방법
인간
시작
탑재
교육
취업
스마트폰
시대
계속
지리
Reality
App

13주: 재밌는 R 실습

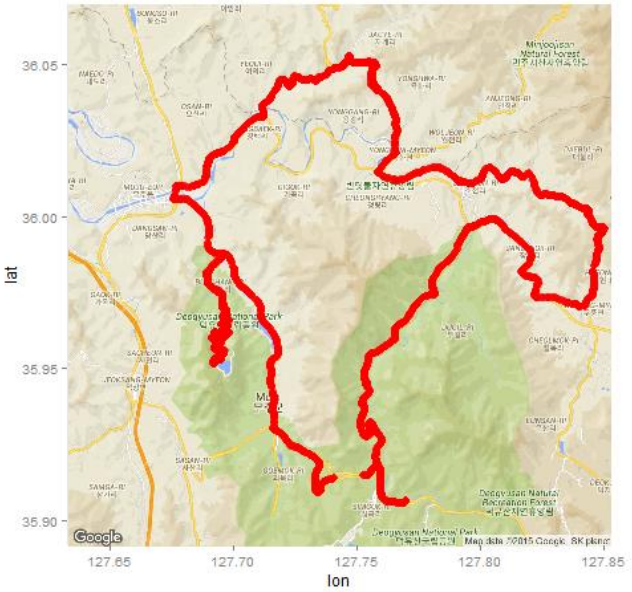
- **Google Map**

```
setwd("d:/rnote")
library(ggplot2)
library(ggmap)
loc <-read.csv("course.csv",header=T)
loc
kor <-get_map("Hallasan", zoom="auto", maptype = "roadmap")
kor.map<-ggmap(kor)+geom_point(data=loc, aes(x=LON, y=LAT),size=3,alpha=0.7,
                               col="red")
kor.map + geom_path(data=loc,aes(x=LON,y=LAT),size=1,linetype=1,col="green")+
geom_text(data=loc, aes(x = LON, y = LAT+0.005, label=장소),size=2)
```

13주: 재밌는 R 실습

- **GPX Data Plotting 응용**

```
library(plotKML)
gData<-readGPX("d:/rnote/pina.gpx")
pina<-gData$tracks[[1]][[1]]
muju <- get_map( location=c(lon = mean(pina$lon), lat = mean(pina$lat)), zoom=12,
maptype = "roadmap")
pina.map <- ggmap(muju)+geom_point(data=pina, aes(x=lon,
y=lat),size=3,alpha=0.7,col="red")
pina.map
```



Last Page

한 학기 동안 수고 많았습니다.

열심히 공부해줘서 감사합니다.