	단위별 학습내용 (Week5)
wk5-1	데이터시각화 : R그래픽
wk5-2	R 그래픽 : ggplot2의 활용
wk5-3	R 그래픽 : 3D와 히트맵
wk5-4	R 그래픽 : 공간지도분석

이혜선
POSTECH 산업경영공학과
email: hyelee@postech.ac.kr
web: http://www.postech.ac.kr/~hyelee/

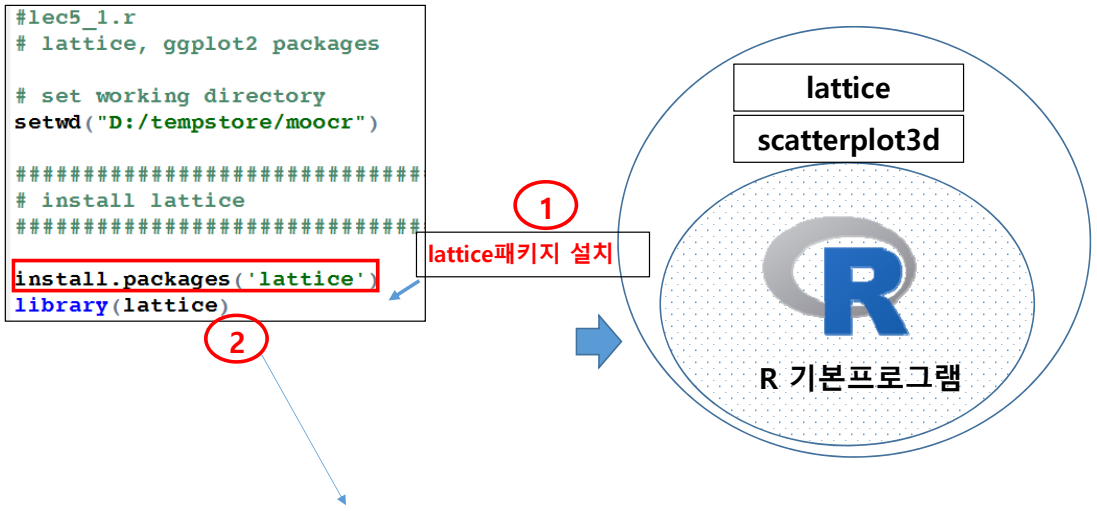
Wk5-1 : 데이터시각화

- R 그래픽 -

R 그래픽 - 데이터시각화

- 추가패키지 설치 (install.packages)

```
#lec5_1.r
# lattice, ggplot2 packages

# set working directory
setwd("D:/tempstore/moocr")

##############################
# install lattice
##############################

install.packages('lattice')
library(lattice)
```



* 해당패키지를 사용할때 library(lattice)는 매번 필요함 !!!

R그래픽 : 데이터시각화

1. R 기본 그래픽스 (Base에 포함되어 있음)
2. Lattice 그래픽스 : 직교형태의 멀티패널 툴
3. ggplot2 그래픽 시스템 : Hadley Wickham이 구현

(1) Grammar of Graphic라는 개념은 그래픽을 생성할때 각 요소를 구분하여 취급한다는 의미

(2) Incremental graphic : 기본 R그래픽스보다 인터액티브한 그래프를 그릴수 있음. 기초 그림을 생성한 후 그래픽스 요소를 필요에 따라 붙이거나 수정

R그래픽 : 데이터시각화

lattice : 직교형태의 그래픽툴

• lattice 함수 : xyplot, bwplot, contourplot, levelplot 등

(1)xyplot : 산점도

(2)bwplot : box whiskers plot, 상자그림

(3)dotplot

(4)levelplot

(5)stripplot : 점을 함께 표시한 상자그림

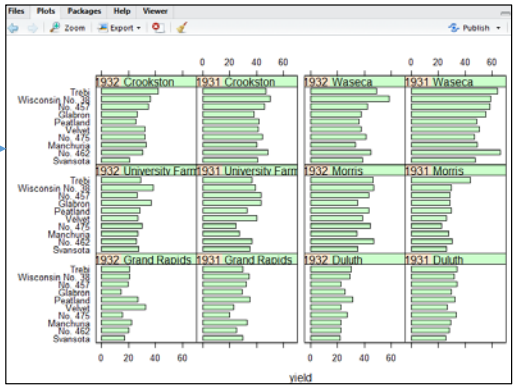
(6)splom : 산점도 매트릭스

(7)contourplot : 등고선그림

lattice 설치와 사용

• lattice : 직교형태의 그래픽툴

```
###################################
# install lattice
###################################

install.packages('lattice')
library(lattice)

# to see several plot using lattice pacakge
demo(lattice)
```



lattice 설치와 활용

- lattice 패키지의 데이터 사용

프로그램 편집창

```
# lec5_2.r
# Grapic using lattice

install.packages('lattice')
library(lattice)

# ethanol data in lattice
help("ethanol")
```

데이터 ethanol (lattice패키지에 들어있는 실습데이터)

A data frame with 88 observations on the following 3 variables.

NOx

Concentration of nitrogen oxides (NO and NO2) in micrograms/J.

C

Compression ratio of the engine.

E

Equivalence ratio–a measure of the richness of the air and ethanol fuel mixture.

lattice 활용 그래픽

- lattice 패키지의 데이터 ethanol

```
# ethanol data in lattice
help("ethanol")

head(ethanol)
dim(ethanol)
str(ethanol)
```

```
> head(ethanol)
    NOx  C     E
1 3.741 12 0.907
2 2.295 12 0.761
3 1.498 12 1.108
4 2.881 12 1.016
5 0.760 12 1.189
6 3.120  9 1.001
> dim(ethanol)
[1] 88  3
> str(ethanol)
'data.frame':   88 obs. of  3 variables:
 $ NOx: num  3.74 2.29 1.5 2.88 0.76 ...
 $ C  : num  12 12 12 12 12 9 9 9 12 12 ...
 $ E  : num  0.907 0.761 1.108 1.016 1.189 ...
```

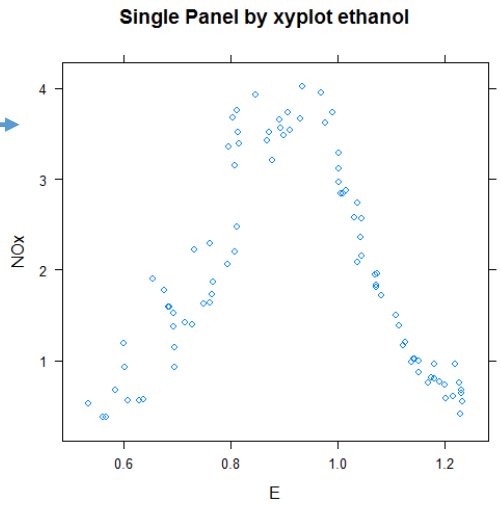
```
table(ethanol$C)
```

```
> table(ethanol$C)

7.5   9  12  15  18
 22  17  14  19  16
```

lattice 활용 그래픽

lattice함수 xyplot을 이용한 그래프 (기본 산점도와 동일)

```
# basic xyplot
xyplot(NOx ~ E , data = ethanol, main = "Single Panel by xyplot")
```



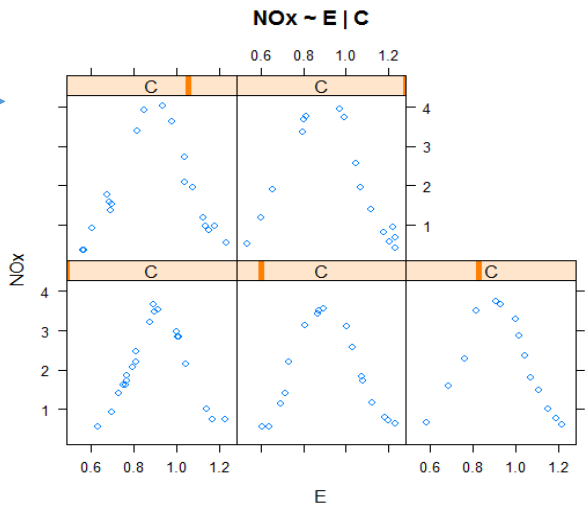
lattice 활용 그래픽

xyplot을 이용한 멀티패널 산점도 : xyplot(y변수~x변수 | 조건부변수, data=)

```
xyplot(NOx ~ E | C , data = ethanol, main = "NOx ~ E | C ")
```

엔진압축비(C)가 조건부변수 :
공기와 연료의 혼합비(E) 와 질소 산화도배출량(NOx))의 산점도

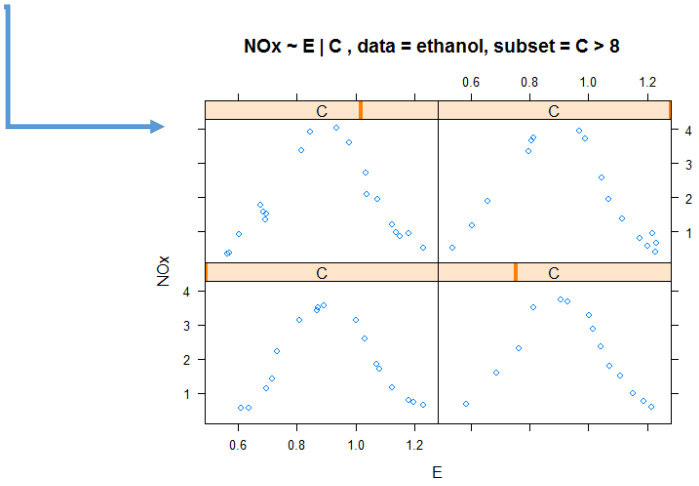
*엔진압축비 (7.5, 9, 12, 15, 18)



lattice 활용 그래픽

xyplot을 이용한 멀티패널 산점도 (subset 데이터)

```
# multi panel graph for subset
xyplot(NOx ~ E | C , data = ethanol, subset = C > 8,
       main = "NOx ~ E | C , data = ethanol, subset = C > 8")
```



R그래픽 : 데이터시각화

- http://www.r-graph-gallery.com/portfolio/ggplot2-package/
- http://www.cookbook-r.com/Graphs/
- http://gglots.org : ggplot2공식 사이트
- http://docs.ggplot2.org : 하위 옵션 사용법 및 예제
- http://groups.google.com/group/ggplot2
- R을 활용한 데이터시각화, 유충현, 홍성학

Wk5-2 : R 그래픽

- ggplot2 활용-

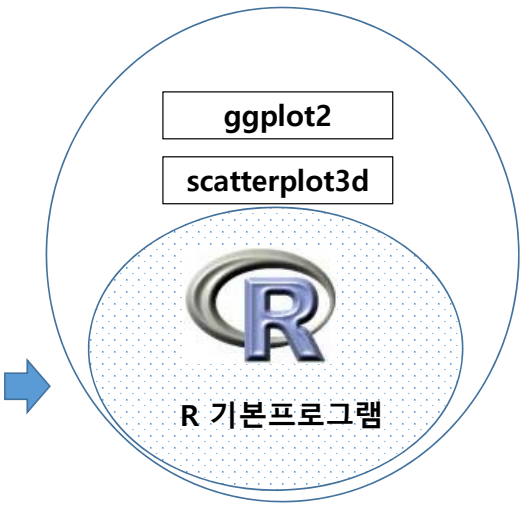
- 추가패키지 설치 (install.packages)

프로그램 편집창

```
# lec5_2.r
# Grapic using ggplot2

# set working directory
setwd("D:/tempstore/moocr")

# library
install.packages('ggplot2')
library(ggplot2)
```

1

2



R 그래픽 : ggplot2 패키지 구조

Grammar of graphics

(1) ggplot()이라는 기본 함수

+

(2-1) Layers : aes (Aesthetic) : 데이터를 어떻게 넣을건지

(2-2) Layer : geom (Geometric objects) : point(점), line(선) 등

(2-3) Layer : coor (coordinate system)

1. scale+coordinate system은 그림을 그릴때 캔버스로 생각하면 됨
2. 그 위에 data+mapping+geom을 추가
3. geom(기하학적 요소): geom_point, geom_smooth등을 이미 그려진 산점도에 추가하여(incremental) 그릴수 있음

R 그래픽 : ggplot2 그래프

- ggplot(데이터이름, aes(x=x축변수, y=y축변수, color=factor변수, shape=factor변수))+geom_point(size=3)

```
# lec5_2.r
# Grapic using ggplot2

# set working directory
setwd("D:/tempstore/moocr")

# library
# install.packages('ggplot2')
library(ggplot2)

# Read in R : autompg data
car<-read.csv("autompg.csv")
head(car)
str(car)

# subset of car : cyl (4,6,8)
car1<-subset(car, cyl==4 | cyl==6 | cyl==8)
attach(car1)

# 5-1 ggplot2 for scatterplot
# Color and shape display by factor (categorical variable)
# check the variable type(integer, numeric, factor) and define
str(car1)
car1$cyl<-as.factor(car1$cyl)
# Now, you can draw one of plot using ggplot
ggplot(car1, aes(x=wt, y=disp, color=cyl, shape=cyl)) +
  geom_point(size=3)
```

Step0 : 분석을 위한 설정(library, package, setwd)

Step1 : 데이터핸들링
(불러들이기, subset, 변수값정의)

Step2 : 데이터탐색
(그래픽)

5-1. ggplot2 : scatterplot by group (cyl) 변수 (lec5_2.R)

5-1. ggplot2 : ggplot객체들의 설명

```
# 5-1 ggplot2 for scatterplot
# Color and shape display by factor (categorical variable)
# check the variable type(integer, numeric, factor) and def
str(car1)
car1$cyl<-as.factor(car1$cyl)
# Now, you can draw one of plot using ggplot
ggplot(car1, aes(x=wt, y=disp, color=cyl, shape=cyl)) +
  geom_point(size=3)
```

```
ggplot(car1, aes(x=wt, y=disp, color=cyl, shape=cyl))
```

1. ggplot함수에 데이터는 car1을 이용하고, x축에는 wt(차의 무게)를, y축에는 disp(배기량)의 산점도를 그리고, 점의 색상은 cyl(실린더 수)로 표현한다

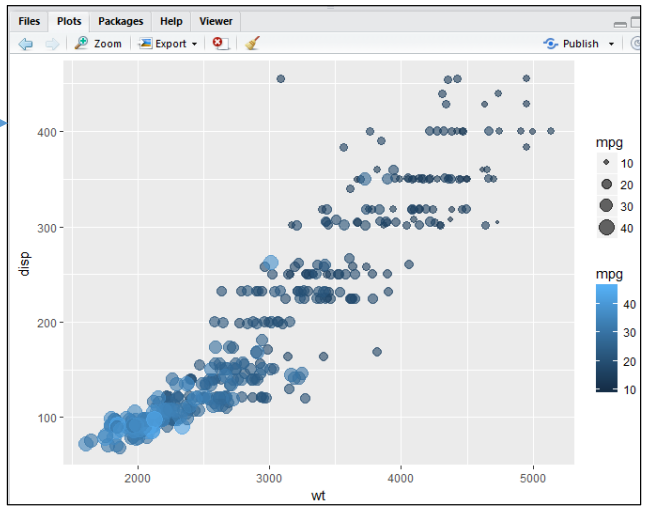
```
geom_point(size=3, alpha=0.6)
```

2. geom_point는 size=3(숫자 클수록 점 크기가 커짐)

R 그래픽 : ggplot2 그래프

5-1. mpg의 크기를 표시한 그래프

```
# mapping (continuous variable : mpg) on the scatterplot for wt and disp
ggplot(car1, aes(x=wt, y=disp, color=mpg, size=mpg)) +
  geom_point(alpha=0.6)
```



설명 : 차의 무게와 배기량의 산점도에 연비의 높고 낮음을 원의 사이즈와 색으로 표시한 그래프

R 그래픽 : ggplot2 구조

ggplot의 기본

ggplot : 새로운 ggplot을 생성
aes : aesthetic mapping을 구성 (데이터, 그래프구조)
qplot: 즉석 그림

geom 함수군

geom_abline, geom_hline, geom_vline
geom_bar
geom_point
geom_boxplot
geom_map
geom_smooth, stat_smooth

geom (geometric) 함수군

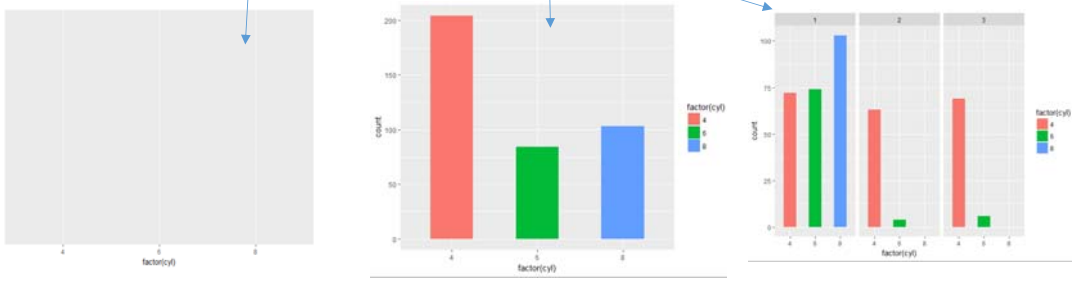
geom메뉴 | **help.search("geom")**

ggplot2::geom_abline	Reference lines: horizontal, vertical, and diagonal
ggplot2::geom_bar	Bars charts
ggplot2::geom_bin2d	Heatmap of 2d bin counts
ggplot2::geom_blank	Draw nothing
ggplot2::geom_boxplot	A box and whiskers plot (in the style of Tukey)
ggplot2::geom_contour	2d contours of a 3d surface
ggplot2::geom_count	Count overlapping points
ggplot2::geom_density	Smoothed density estimates
plot2::geom_density_2d	Contours of a 2d density estimate
ggplot2::geom_dotplot	Dot plot
ggplot2::geom_errorbarh	Horizontal error bars
ggplot2::geom_hex	Hexagonal heatmap of 2d bin counts
ggplot2::geom_freqpoly	Histograms and frequency polygons
ggplot2::geom_jitter	Jittered points
ggplot2::geom_crossbar	Vertical intervals: lines, crossbars & errorbars
ggplot2::geom_map	Polygons from a reference map
ggplot2::geom_path	Connect observations
ggplot2::geom_point	Points
ggplot2::geom_polygon	Polygons
ggplot2::geom_qq	A quantile-quantile plot

5-1. geom_bar 을 이용한 단계별 그래프 설명

```
# 5-2-1 : geom_bar : asethetic mapping (4,6,8 cyl)
p1<-ggplot(car1, aes(factor(cyl), fill=factor(cyl)))
p1
# barplot define
p1<-p1 + geom_bar(width=.5)
p1
# output by origin(1,2,3)
p1<-p1 + facet_grid(. ~ origin)
p1
```

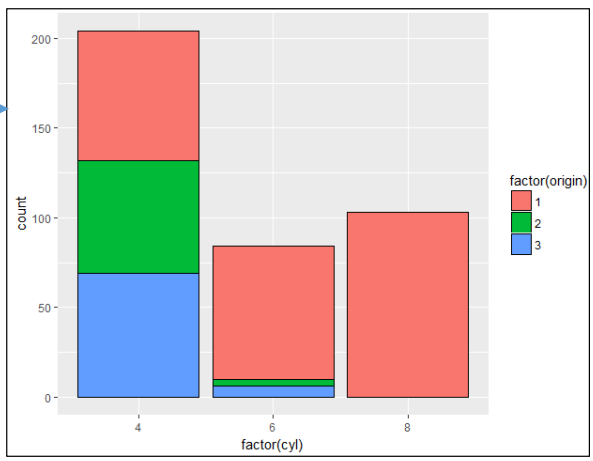
1. grid를 그림-cyl에 따라서
2. geom_bar을 이용한 cyl 빈도의 막대그래프
3. cyl의 bar chart를 변수'origin'에 따라 그림



```
# 5-2-1 : same plot with the above
ggplot(car1, aes(factor(cyl), fill=factor(cyl)))+ geom_bar(width=.5)+ facet_grid(. ~ origin)
```

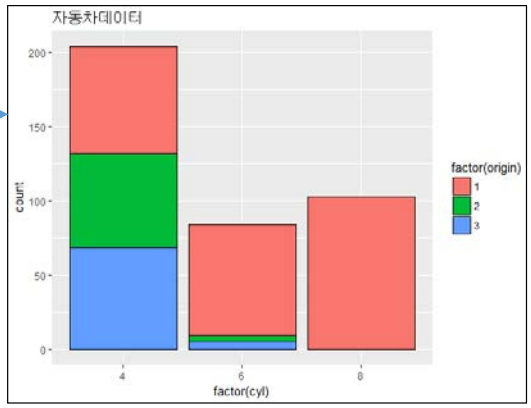

5-2. geom_bar 을 이용한 누적 막대그래프

```
# 5-2-2 : geom_bar : asethetic mapping (4,6,8 cyl)
p <- ggplot(data=car1, aes(factor(cyl)))
p + geom_bar(aes(fill=factor(origin)), colour="black")
```



5-2. geom_bar 을 이용한 누적 막대그래프- 한글제목 넣기

```
# for Korean title
# 5-2-2 : geom bar : asethetic mapping (4,6,8 cyl)
par(family="나눔고딕", cex=1.3)
p <- ggplot(data=car1, aes(factor(cyl)))
p<-p + geom_bar(aes(fill=factor(origin)), colour="black")
p<-p+ggtitle("자동차데이터")
p
```



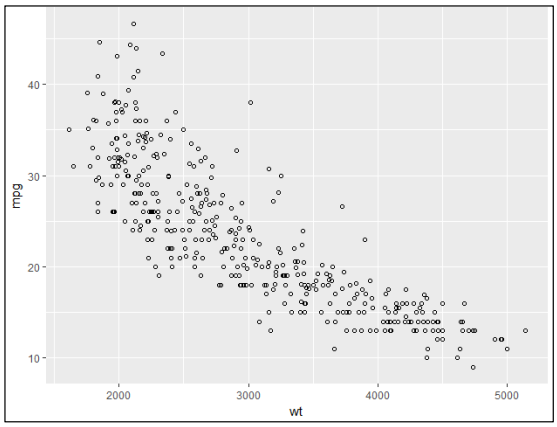
5-3. ggplot 산점도에 회귀선 넣기 : step1

autpmpg 데이터 사용 : 차량무게에 따른 연비 예측

```
# 5-3 ggplots for scatterplot with regression line
# step1 : # Use hollow circles
ggplot(car1, aes(x=wt, y=mpg))+geom_point(shape=1)
```

1. ggplot
2. geom_point : shape=1, size=

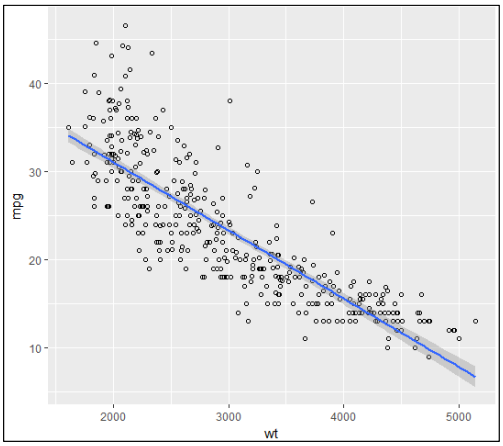
ggplot(데이터이름, aes(x=x축변수, y=y축변수))
geom_point(shape=1, size=)



5-3. ggplot 산점도에 회귀선 넣기 : step2

```
# step2 :  # Add linear regression line
# by default includes 95% confidence region
ggplot(car1, aes(x=wt, y=mpg)) + geom_point(shape=1) + geom_smooth(method=lm)
```

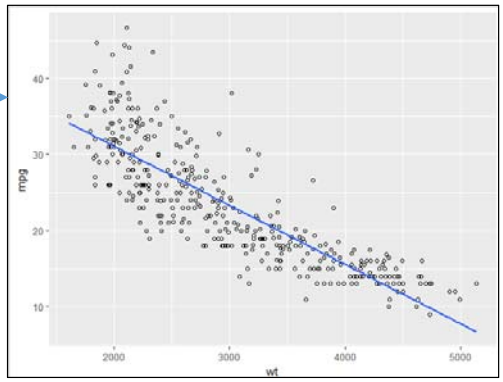
geom_smooth(method=lm) :
선형회귀식 추가, 선형식의 95% 신뢰구간이 default로 그려짐



5-3. ggplot 산점도에 회귀선 넣기 : step2

- 옵션 : 신뢰구간이 없는 회귀선 -

```
# excluding 95% confidence region
ggplot(car1, aes(x=wt, y=mpg)) +
  geom_point(shape=1) +  geom_smooth(method=lm, se=FALSE)
```

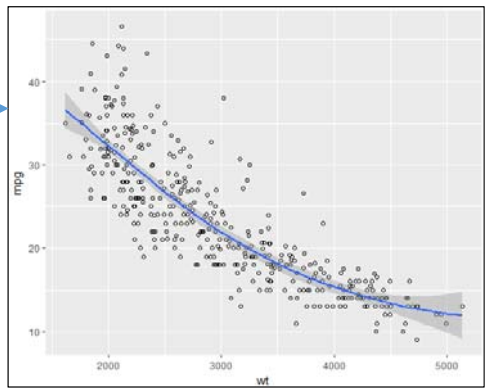


R 그래픽 : ggplot2 활용

5-3. ggplot 산점도에 비선형회귀식 적합 : step3

```
# step3: Add a loess smoothed fit curve with confidence region
# geom_smooth() use loess
ggplot(car1, aes(x=wt, y=mpg)) + geom_point(shape=1) + geom_smooth(method="loess")
```

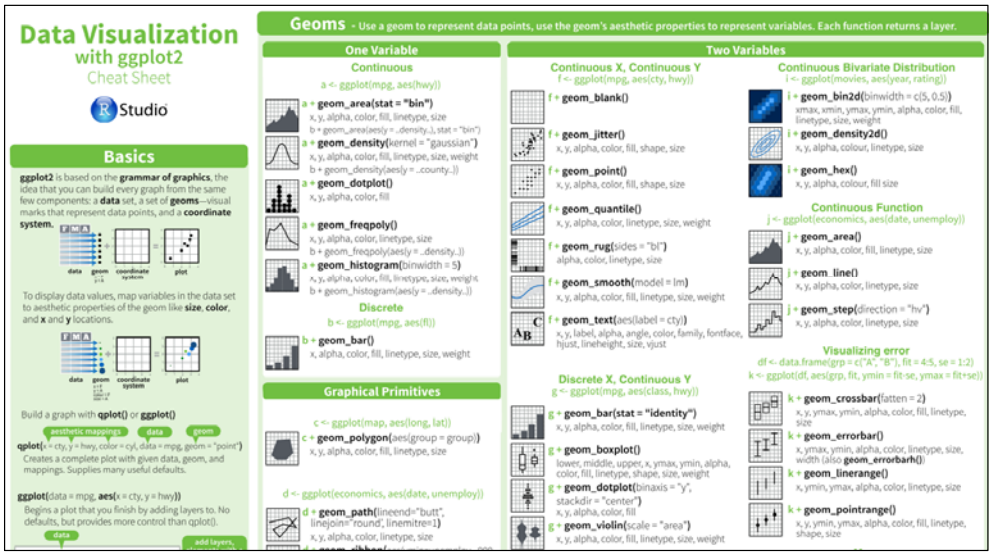
loess(local polynomial regression)



ggplot2 그래픽 사이트

- R-studio의 ggplot2활용 요약

https://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf



Wk5-3 : R 그래픽-3D, 히트맵

R 추가패키지

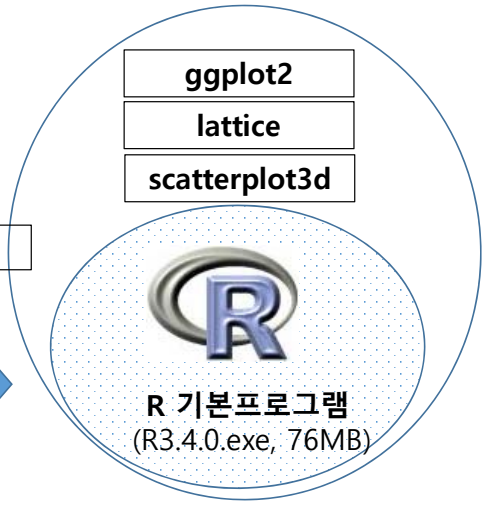
- 추가패키지 설치 (install.packages)

프로그램 편집창

```
# lec5_3.r
# 3D-plot, Heatmap

# set working directory
setwd("D:/tempstore/moocr")

# install 3d scatterplot package
install.packages('scatterplot3d')
library(scatterplot3d)
```

① scatterplot3d 패키지 설치

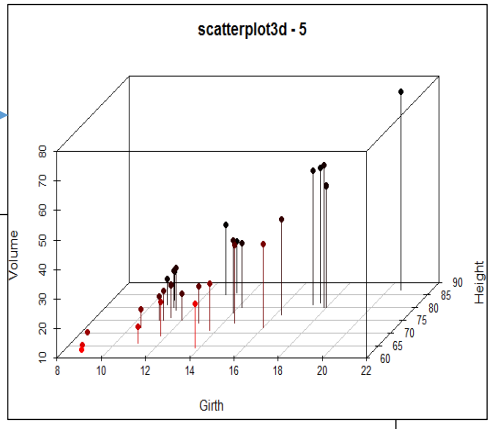
② library(scatterplot3d)



3-D plot

• 3D scatterplot : scatterplot(데이터$변수, pch= , ...)

```
# lec5_3.r
# 3D-plot, Heatmap

# set working directory
setwd("D:/tempstore/moocr")

# install 3d scatterplot package
install.packages('scatterplot3d')
library(scatterplot3d)

# 5-3 3D scatterplot with data trees
data(trees)
s3d <- scatterplot3d(trees, type="h", highlight.3d=TRUE,
                     angle=55, scale.y=0.7, pch=16, main="scatterplot3d - 5")
```



3-D plot

• 데이터 trees에 관하여 : help(trees), head(tress), write.csv(....)

```
# to know about data "trees"
help(trees)
head(trees)

# export to csv file
write.csv(trees,file="trees.csv", row.names = FALSE)
```

trees 데이터

Girth : 체리나무의 지름 (inch)

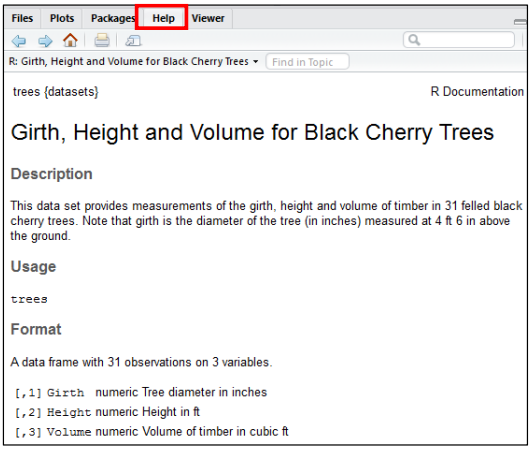
Height : 체리나무의 키 (feet)

Volume : 나무의 부피(ft^3)

⇒ 나무의 둘레(x1)와 키(x2)로 목재의 부피(y)를 예측

⇒ y=f(x1, x2)

도움말


Files Plots Packages Help Viewer

R: Girth, Height and Volume for Black Cherry Trees

Find in Topic

trees {datasets} R Documentation

Girth, Height and Volume for Black Cherry Trees

Description

This data set provides measurements of the girth, height and volume of timber in 31 felled black cherry trees. Note that girth is the diameter of the tree (in inches) measured at 4 ft 6 in above the ground.

Usage

```
trees
```

Format

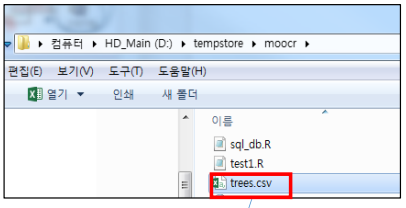
A data frame with 31 observations on 3 variables.

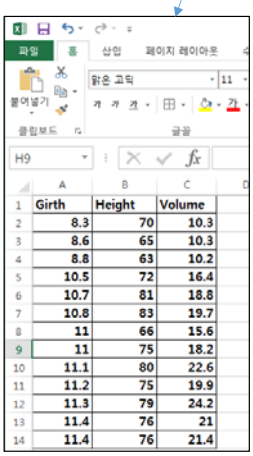
```
[,1] Girth   numeric Tree diameter in inches
[,2] Height numeric Height in ft
[,3] Volume numeric Volume of timber in cubic ft
```

3-D plot

• 데이터 내보내기 (실습)

현재 작업창 : working directory





프로그램 편집창

```
# to know about data "trees"
help(trees)
head(trees)

# export to csv file
write.csv(trees,file="trees.csv", row.names = FALSE)
```

컨솔창

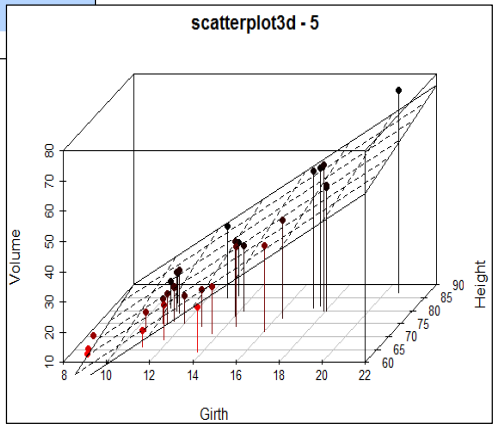
```
> head(trees)
  Girth Height Volume
1   8.3     70   10.3
2   8.6     65   10.3
3   8.8     63   10.2
4  10.5     72   16.4
5  10.7     81   18.8
6  10.8     83   19.7
> write.csv(trees,file="trees.csv", row.names = FALSE)
```


3-D plot

• 3D 산점도에 선형식 추가

```
# 5-3 3D scatterplot with data trees
data(trees)
s3d <- scatterplot3d(trees, type="h", highlight.3d=TRUE,
                     angle=55, scale.y=0.7, pch=16, main="

# Now adding some points to the "scatterplot3d"
#s3d$points3d(seq(10,20,2), seq(85,60,-5), seq(60,10,-10),

# Now adding a regression plane to the "scatterplot3d"
attach(trees)
my.lm <- lm(Volume ~ Girth + Height)
s3d$plane3d(my.lm, lty.box = "solid")
```

1. 산점도를 s3d(이름은 자유로 지정)로 저장하고,
2. s3d의 요소중 planed3d에 lm(Volume~Girth+Hieght)의 선형식을 추가



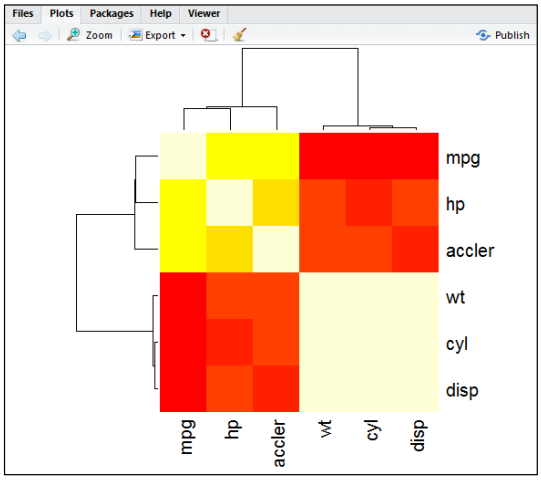
lm(linear model)
lm(y변수~x변수+x변수+..)
여기서는 y변수(나무의 부피), x변수(나무의 지름, 키)
=> 나무의 부피는 나무의 지름이 클수록, 나무의 키가 클수록 크다

히트맵 (heatmap)

• 히트맵 : 통계치를 구한후 크기에 비례하여 그라데이션 색상으로 표현한 시각화기법. 색상을 열의온도를 연상하게 해서 열(heatmap)이라고 함

Autompg 데이터의 상관계수를 이용한 히트맵

```
# heatmap 1 using autompg data
par(mfrow=c(1, 1))
cor.x<-cor(car[,1:6])
heatmap(cor.x, symm=TRUE)
```



상관행렬의 시각화 :
대각선에 위치한 변수들이 상관계수가 가장 큰것을 알수 있음

```
> round(cor.x,2)
        mpg   cyl  disp    hp    wt accler
mpg    1.00 -0.78 -0.80  0.42 -0.83    0.42
cyl   -0.78  1.00  0.95 -0.54  0.90   -0.51
disp  -0.80  0.95  1.00 -0.48  0.93   -0.54
hp     0.42 -0.54 -0.48  1.00 -0.48    0.26
wt    -0.83  0.90  0.93 -0.48  1.00   -0.42
accler 0.42 -0.51 -0.54  0.26 -0.42    1.00
```


히트맵 (heatmap)

• 데이터 : USArrest (미국범죄율 데이터, 1973년)

```
# Crime rate by US State (1973)
# Arrests per 100,000 residents for assault, murder, and rape
# in each of the 50 US states in 1973

help(USArrests)
head(USArrests)
```

```
> head(USArrests)
           Murder Assault UrbanPop Rape
Alabama      13.2     236       58 21.2
Alaska       10.0     263       48 44.5
Arizona       8.1     294       80 31.0
Arkansas      8.8     190       50 19.5
California    9.0     276       91 40.6
Colorado      7.9     204       78 38.7
```

상관계수(변수간 상관관계)

```
cor(USArrests)
round(cor(USArrests), 2)
```

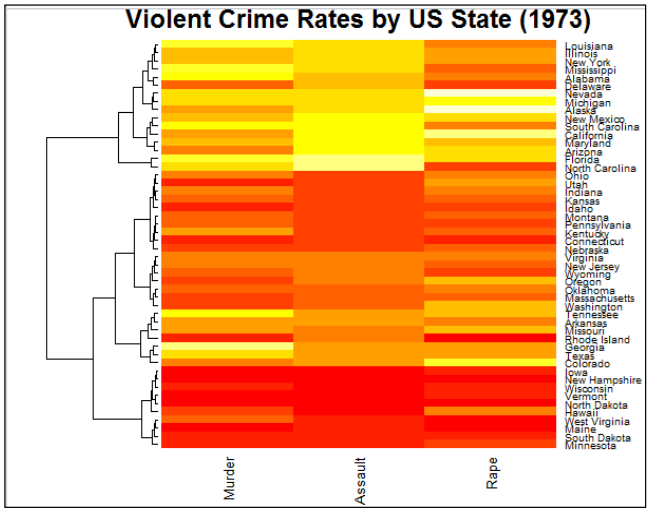
```
> round(cor(USArrests), 2)
         Murder Assault UrbanPop Rape
Murder     1.00    0.80     0.07 0.56
Assault    0.80    1.00     0.26 0.67
UrbanPop   0.07    0.26     1.00 0.41
Rape       0.56    0.67     0.41 1.00
```

히트맵 (heatmap)

• 미국 (1973년, State별 범죄)

4개변수들중 UrbanPop(도심인구비율) 변수는 제외한 행렬

```
# subset excluding 3th variable UrbanPop
x  <- as.matrix(USArrests[, -3])
result <- heatmap(x, scale="column", Colv=NA, cexCol=1,
                  main="Violent Crime Rates by US State (1973)")
```



인구 십만명당 폭행, 살인, 성범죄 비율에 따른 히트맵 : 범죄비율이 높은 주부터 낮은 주를 표시 (빨강색이 범죄율 낮은 주)

히트맵 (heatmap)

• 데이터 (미국 주별 범죄발생, 1973년)

```
# subset excluding 3th variable UrbanPop
x  <- as.matrix(USArrests[, -3])
result <- heatmap(x, scale="column", Colv=NA, cexCol=1,
                  main="Violent Crime Rates by US State (1973)")

row.names(USArrests)[result$rowInd[1:10]]
row.names(USArrests)[result$rowInd[35:50]]
```

범죄발생 낮은주(10개주 1:10)

```
> row.names(USArrests)[result$rowInd[1:10]]
 [1] "Minnesota"      "South Dakota"   "Maine"          "West Virginia"
 [5] "Hawaii"         "North Dakota"   "Vermont"        "Wisconsin"
 [9] "New Hampshire"  "Iowa"
```

범죄발생 높은주 (16개주 35:50)

```
> row.names(USArrests)[result$rowInd[35:50]]
 [1] "North Carolina" "Florida"        "Arizona"        "Maryland"
 [5] "California"     "South Carolina" "New Mexico"     "Alaska"
 [9] "Michigan"       "Nevada"         "Delaware"       "Alabama"
[13] "Mississippi"    "New York"       "Illinois"       "Louisiana"
```

Wk5-4 : R 그래픽-공간지도분석

R 추가패키지 설치

- 추가패키지 설치 (install.packages)

```
# lec5_4.R

#install.packages("ggplot2")
library(ggplot2)

# maps : world map
install.packages("maps")
library(maps)

# mapdata : more world map
install.packages("mapdata")
library(mapdata)

# mapdata : latitude and longitude
install.packages("mapproj")
library(mapproj)
```

1. maps – 세계의 지도 데이터베이스
2. mapdata – maps보다 정교한 지도
3. mapproj – 위도와 경도
4. ggplot2 – R 그래픽 패키지

R그래픽 : 공간지도

- 한국지도 : maps와 mapdata를 이용한 지도

```
# using map package
par(mfrow = c(1, 2),mar=c(2,2,2,2))
map(database = 'world', region = c('South Korea','North Korea'), col='green', fill = TRUE)
title("Korea")
# using mapdata package
map(database = 'worldHires', region = c('South Korea','North Korea'), col='green', fill = TRUE)
title("Korea")
```

R그래픽 : 공간지도

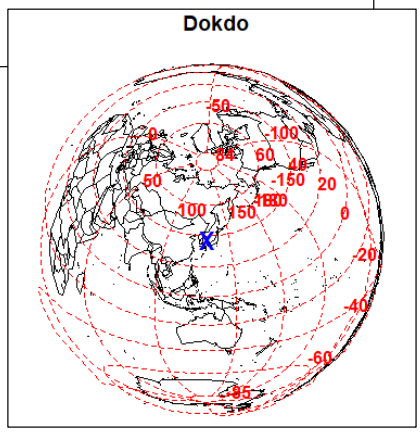
- 지도 : 이탈리아

```
# 2.Italy
par(mfrow = c(1, 1),mar=c(2,2,2,2))
map(database = 'world', region = c('Italy'), col='coral', fill = TRUE)
title("Italy")
```

R그래픽 : 공간지도

- **mapproj 패키지 - 위도, 경도 활용 (독도를 표시)**

```
# Dokdo map using mapproj package
library(mapproj)
map('world', proj = 'azequalarea', orient = c(37.24223, 131.8643, 0))
map.grid(col = 2)
points(mapproject(list(y = 37.24223, x = 131.8643)), col = "blue", pch = "x", cex = 2)
title("Dokdo")
# for reading Korean : encoding to UTF-8
# file menu: Tools_global options_code_saving
```



공간지도분석 예제 1 : 국내공항 및 항공노선

- **국내 공항 및 노선 현황**

```
# 4. Airport & route data (source : https://www.data.go.kr/)
airport<-read.csv("airport.csv")
route = read.csv("route.csv")
head(airport)
head(route)

head(route[order(route$id),])
```

- **airport.csv : 국내 공항 위치 정보**

```
head(airport)
  airport iata     lon     lat
1    강릉  KAG 128.944 37.7536
2    광주  KWJ 126.809 35.1264
3    군산  KUV 126.616 35.9038
4    김포  GMP 126.791 37.5583
5    대구  TAE 128.659 35.8941
6    목포  MPK 126.380 34.7589
```

- **route.csv : 국내선 노선 정보**

```
  id airport     lon     lat
1  1     CJJ 127.499 36.7166
2  7     CJJ 127.499 36.7166
3 45     CJJ 127.499 36.7166
4 77     CJJ 127.499 36.7166
5  2     CJJ 127.499 36.7166
6  8     CJJ 127.499 36.7166
```

id기준 정렬

```
   id airport     lon     lat
1   1     CJJ 127.499 36.7166
83  1     CJU 126.493 33.5113
2   2     CJU 126.493 33.5113
84  2     CJJ 127.499 36.7166
3   3     CJU 126.493 33.5113
85  3     GMP 126.791 37.5583
```

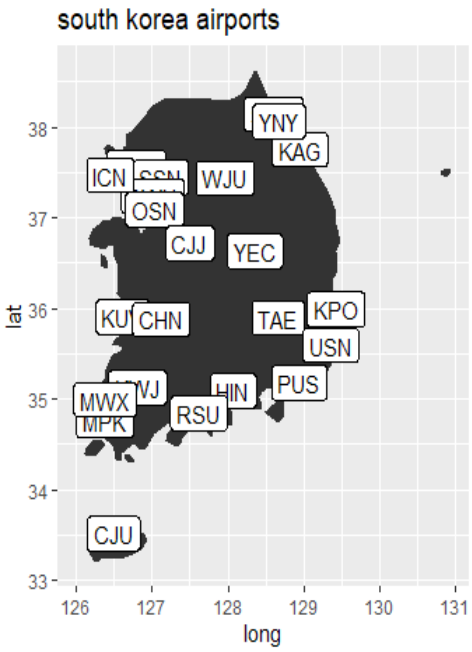
항공노선

```
head(route[order(route$id),])
```

공간지도분석 예제 : 국내공항 및 항공노선

- 국내 공항위치

```
# Korea map (kr.map)
world.map <- map_data("world")
kr.map <- world.map %>% filter(region == "South Korea")

# ------------------------------------ #
# Korea map using ggplot
# ------------------------------------ #
# 5. Domestic airport location
ggplot() +
  geom_polygon(data=kr.map, aes(x=long, y=lat, group=group)) +
  geom_label(data=airport, aes(x = lon, y = lat, label=iata)) +
  labs(title = "south korea airports")
```



공간지도분석과 ggplot

- ggplot 은 레이어를 추가하는 방식으로 그래픽을 구현

1

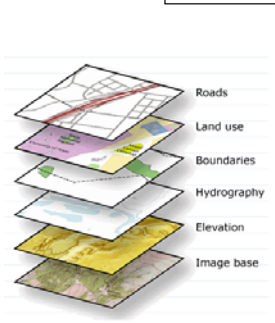
```
ggplot(data=kr.map, aes(x=long, y=lat, group=group))
```

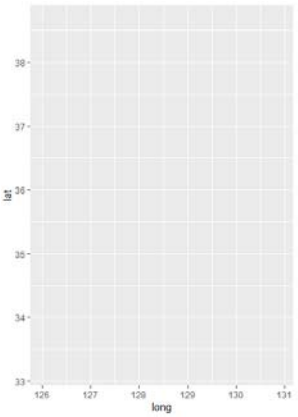
2

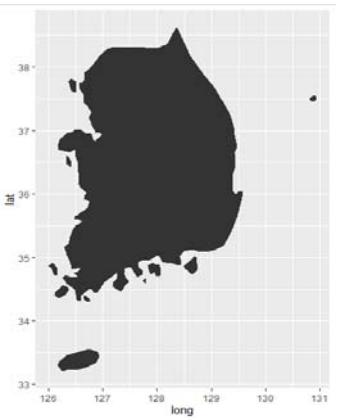
```
ggplot(data=kr.map, aes(x=long, y=lat, group=group)) +
  geom_polygon(fill="white", colour="black")
```

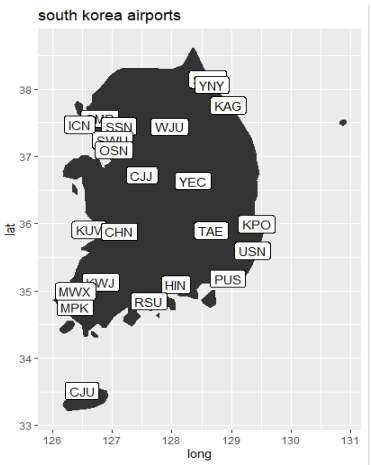
3

```
ggplot() +
  geom_polygon(data=kr.map, aes(x=long, y=lat, group=group)) +
  geom_label(data=airport, aes(x = lon, y = lat, label=iata)) +
  labs(title = "south korea airports")
```





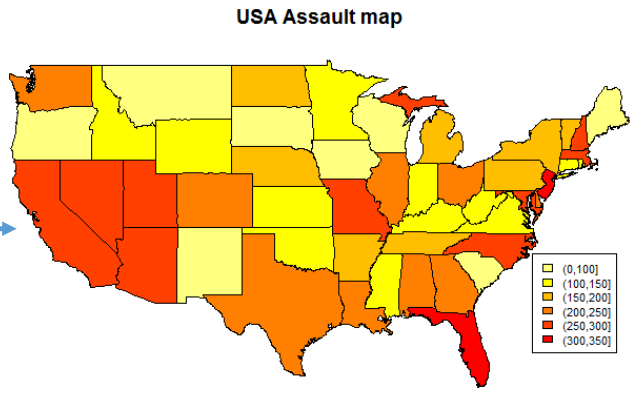




공간지도분석 예제 : 미국 행정데이터

- 지도 데이터베이스와 행정자료 결합 : 미국 (1973) 범죄수 지도

```
# Assault in US (1973)
par(mfrow = c(1, 1))
library(maps)
# excluding Alaska, Hawaii
sub.usa <- subset(USArrests,!rownames(USArrests) %in% c("Alaska", "Hawaii"))
# data with State name, Assult count
usa.data <- data.frame(states = rownames(sub.usa), Assault = sub.usa$Assault)
# legend
col.level <- cut(sub.usa[, 2], c(0, 100, 150, 200, 250, 300, 350))
legends <- levels(col.level)
# displaying color for the size
levels(col.level) <- sort(heat.colors(6), decreasing = TRUE)
usa.data <- data.frame(usa.data, col.level = col
# Map
map('state', region = usa.data$states, fill = TR
title("USA Assault map")
legend(-76, 35, legends, fill = sort(heat.colors
```



공간지도분석 예제 : 미국 행정데이터

- 행정자료 (USArrests 는 기본프로그램에 포함된 데이터)

```
help(USArrests)
head(USArrests)
```

Description

This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas.

Usage

```
USArrests
```

Format

A data frame with 50 observations on 4 variables.

[,1] Murder	numeric Murder arrests (per 100,000)
[,2] Assault	numeric Assault arrests (per 100,000)
[,3] UrbanPop	numeric Percent urban population
[,4] Rape	numeric Rape arrests (per 100,000)

```
> head(USArrests)
           Murder Assault UrbanPop Rape
Alabama      13.2     236       58 21.2
Alaska       10.0     263       48 44.5
Arizona       8.1     294       80 31.0
Arkansas      8.8     190       50 19.5
California    9.0     276       91 40.6
Colorado      7.9     204       78 38.7
```

	데이터탐색 (Week6)
wk6-1	데이터 다루기
wk6-2	데이터의 기술통계치요약
wk6-3	그래프를 이용한 데이터탐색
wk6-4	데이터의 정규성검정과 신뢰구간

6-1. 데이터 다루기 (데이터결합, 분할, 정렬)

Wk6-1 : 데이터 다루기

- 데이터결합, 분할, 정렬 -

데이터 다루기

- 데이터 결합 : merge(data1, data2, by="ID")

data1과 data2는 아래와 같이 식별변수 ID를 기준으로 결합
data1 : 게임장르, 나이, 성별 data2 : 주당게임시간, 음주경험, 흡연경험

data1.csv

A	B	C	D
ID	age	gender	game
111	16	F	RTS
112	17	F	FPS
113	15	M	Sport
114	18	M	MMORPG
115	14	F	MMORPG
116	15	F	FPS
117	13	M	Sport
118	19	F	FPS
119	17	M	Sport
120	18	F	RTS

data2.csv

A	B	C	D
ID	hourwk	alcohol	smoke
111	10	yes	yes
112	8	no	no
113	4	no	no
114	10	no	no
115	2	no	yes
116	10	yes	yes
117	12	yes	yes
118	8	no	no
119	6	no	no
120	4	no	no

Filter

	ID	age	gender	game	hourwk	alcohol	smoke
1	111	16	F	RTS	10	yes	yes
2	112	17	F	FPS	8	no	no
3	113	15	M	Sport	4	no	no
4	114	18	M	MMORPG	10	no	no
5	115	14	F	MMORPG	2	no	yes
6	116	15	F	FPS	10	yes	yes
7	117	13	M	Sport	12	yes	yes
8	118	19	F	FPS	8	no	no
9	119	17	M	Sport	6	no	no
10	120	18	F	RTS	4	no	no

데이터 다루기

- 데이터 결합 : merge(data1, data2, by="ID")

dat1과 dat2를 ID를 기준으로 결합 (관측치수는 동일함, 변수들의 정보가 추가됨)

```
# lec6_1.r
# Data exploration

# set working directory
setwd("D:/tempstore/moocr")

# practice data
dat1<-read.csv(file="data1.csv")
dat2<-read.csv(file="data2.csv")

# data merging
dat12<-merge(dat1, dat2, by="ID")
```

Filter

	ID	age	gender	game	hourwk	alcohol	smoke
1	111	16	F	RTS	10	yes	yes
2	112	17	F	FPS	8	no	no
3	113	15	M	Sport	4	no	no
4	114	18	M	MMORPG	10	no	no
5	115	14	F	MMORPG	2	no	yes
6	116	15	F	FPS	10	yes	yes
7	117	13	M	Sport	12	yes	yes
8	118	19	F	FPS	8	no	no
9	119	17	M	Sport	6	no	no
10	120	18	F	RTS	4	no	no

데이터 다루기

- 데이터 결합 : rbind(data3, data4)

data3과 data4가 동일한 변수들을 갖고 있을때 두개 데이터를 행(row)으로 결합

dat12

	ID	age	gender	game	hourwk	alcohol	smoke
1	111	16	F	RTS	10	yes	yes
2	112	17	F	FPS	8	no	no
3	113	15	M	Sport	4	no	no
4	114	18	M	MMORPG	10	no	no
5	115	14	F	MMORPG	2	no	yes
6	116	15	F	FPS	10	yes	yes
7	117	13	M	Sport	12	yes	yes
8	118	19	F	FPS	8	no	no
9	119	17	M	Sport	6	no	no
10	120	18	F	RTS	4	no	no

data3

ID	age	gender	game	hourwk	alcohol	smoke
121	20	F	RTS	10	yes	yes
122	21	F	FPS	8	no	no
123	20	M	Sport	12	no	no

```
# add more data (combine in a row)
dat3<-read.csv(file="data3.csv")
dat123<-rbind(dat12, dat3)
dat123
```

```
> dat3<-read.csv(file="data3.csv")
> dat123<-rbind(dat12, dat3)
> dat123
    ID age gender   game hourwk alcohol smoke
1  111  16      F    RTS     10     yes   yes
2  112  17      F    FPS      8      no    no
3  113  15      M  Sport      4      no    no
4  114  18      M MMORPG     10      no    no
5  115  14      F MMORPG      2      no   yes
6  116  15      F    FPS     10     yes   yes
7  117  13      M  Sport     12     yes   yes
8  118  19      F    FPS      8      no    no
9  119  17      M  Sport      6      no    no
10 120  18      F    RTS      4      no    no
11 121  20      F    RTS     10     yes   yes
12 122  21      F    FPS      8      no    no
13 123  20      M  Sport     12      no    no
```


데이터 다루기

- 데이터 정렬 : 데이터이름[order(변수1, 변수2),]

변수1로 먼저 정렬을 하고 그다음 변수2로 정렬

```
# data sorting
dats1<-dat12[order(dat12$age),]
dats1
dats2<-dat12[order(dat12$gender, dat12$age), ]
dats2
```

연령별(age)로 정렬

```
> dats1<-dat12[order(dat12$age),]
> dats1
    ID age gender   game hourwk alcohol smoke
7  117  13      M  Sport     12     yes   yes
5  115  14      F MMORPG      2      no   yes
3  113  15      M  Sport      4      no    no
6  116  15      F    FPS     10     yes   yes
1  111  16      F    RTS     10     yes   yes
2  112  17      F    FPS      8      no    no
9  119  17      M  Sport      6      no    no
4  114  18      M MMORPG     10      no    no
10 120  18      F    RTS      4      no    no
8  118  19      F    FPS      8      no    no
```

데이터 다루기

- 데이터 정렬 : 데이터이름[order(변수1, 변수2),]

```
# data sorting
dats1<-dat12[order(dat12$age),]
dats1
dats2<-dat12[order(dat12$gender, dat12$age), ]
dats2
```

성별(gender)로 정렬한다음 그 다음은 연령별(age)로 정렬

```
> dats2<-dat12[order(dat12$gender, dat12$age), ]
> dats2
    ID age gender   game hourwk alcohol smoke
5  115  14      F MMORPG      2      no   yes
6  116  15      F    FPS     10     yes   yes
1  111  16      F    RTS     10     yes   yes
2  112  17      F    FPS      8      no    no
10 120  18      F    RTS      4      no    no
8  118  19      F    FPS      8      no    no
7  117  13      M  Sport     12     yes   yes
3  113  15      M  Sport      4      no    no
9  119  17      M  Sport      6      no    no
4  114  18      M MMORPG     10      no    no
```


데이터 다루기

- 데이터 추출 – subset(데이터이름, 조건1 & 조건2)

dat12에서 gender=F이고 age>15이상인 데이터를 newdat라는 이름의 데이터로 저장

```
# data subset (selecting data)
#newdat<-dat12[which(dat12$gender=="F" & dat12$age>15),]
newdat<-subset(dat12, dat12$gender=="F" & dat12$age>15)
newdat
```

```
> newdat<-subset(dat12, dat12$gender=="F" & dat12$age>15)
> newdat
   ID age gender game hourwk alcohol smoke
1 111  16      F  RTS     10     yes   yes
2 112  17      F  FPS      8      no    no
8 118  19      F  FPS      8      no    no
10 120 18      F  RTS      4      no    no
```

데이터 다루기

- 데이터에서 일부변수 제거하기 – 데이터이름[!names(데이터) %in% c("변수1","변수2")]

dat12에서 age와 gender를 제외하고 exdat라는 이름의 데이터로 저장 (!는 not을 의미)

```
# excluding variables
exdat<-dat12[!names(dat12) %in% c("age","gender")]
exdat
```

```
> exdat<-dat12[!names(dat12) %in% c("age","gender")]
> exdat
    ID    game hourwk alcohol smoke
1  111     RTS     10     yes   yes
2  112     FPS      8      no    no
3  113   Sport      4      no    no
4  114 MMORPG     10      no    no
5  115 MMORPG      2      no   yes
6  116     FPS     10     yes   yes
7  117   Sport     12     yes   yes
8  118     FPS      8      no    no
9  119   Sport      6      no    no
10 120     RTS      4      no    no
```


데이터분석 : 데이터 사이언티스트

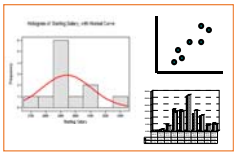
- 데이터 핸들링 -> 데이터 탐색 -> 통계적 모델링 (통계모형, 기계학습, 인공지능)

탐색적 데이터분석

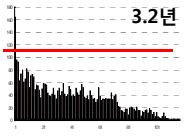
기술통계량 (평균, 빈도)
고객의 연령, 성별, 주거형태,
직업, 거주지

```
> describe(stud[vars])
   vars   n  mean   sd median
G1    1 395 10.91 3.32     11
G2    2 395 10.71 3.76     11
G3    3 395 10.42 4.58     11
```

히스토그램, 산점도, 파레토 그래프
연령대별, 제품가격대별, 구매수단별,
서비스, RFM



구매주기 - 제품교체주기 파악
2회이상 구매자들의 재구매시점을 계산
히스토그램 및 평균으로 분석



where we are!!

통계적 분석기법

상관분석
(X,Y 모두 continuous variable)

일반적으로 0.7이상이면 높다고
보지만 절대적 기준은 없다.

카이제곱분석 –범주형 변수간 상관관계
(X,Y 모두 범주형 변수)
유의수준 0.1, 0.05에서 판단

```
> #post-hoc analysis Tukey
> a1 <- aov(value ~ drug + age)
> posthoc <- TukeyHSD(x=a1, 'drug', conf.level=0.95)
> posthoc
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = value ~ drug + age)

$drug
                 diff        lwr        upr     p adj
5mg-10mg    -3.500000 -5.843288 -1.156712 0.0041976
placebo-10mg -7.333333 -9.676621 -4.990045 0.0000029
placebo-5mg  -3.833333 -6.176621 -1.490045 0.0020536
```

분산분석(ANOVA)
매장평수별 판매금액, 횟수의 차이
그룹간 유의한 차이는 0.05, 0.1에서 결정

구매 중요 요인 도출 (마케팅)
불량 요인 도출 (제조업)
위험요인 도출 (금융업)

Wk6-2 : 데이터의 기술통계치요약

데이터 기술통계치요약

- 데이터 : 학생들의 학업성취도* (포루투칼의 고등학생 수학점수)
- http://archive.ics.uci.edu/ml/datasets/Student+Performance

stud_math.csv

	A	B	C	D	E	F	G	H	I	J	K	L	
1	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	trav
2	GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother	
3	GP	F	17	U	GT3	T	1	1	at_home	other	course	father	
4	GP	F	15	U	LE3	T	1	1	at_home	other	other	mother	
5	GP	F	15	U	GT3	T	4	2	health	services	home	mother	
6	GP	F	16	U	GT3	T	3	3	other	other	home	father	
7	GP	M	16	U	LE3	T	4	3	services	other	reputation	mother	
8	GP	M	16	U	LE3	T	2	2	other	other	home	mother	
9	GP	F	17	U	GT3	A	4	4	other	teacher	home	mother	
10	GP	M	15	U	LE3	A	3	2	services	other	home	mother	
11	GP	M	15	U	GT3	T	3	4	other	other	home	mother	
12	GP	F	15	U	GT3	T	4	4	teacher	health	reputation	mother	
13	GP	F	15	U	GT3	T	2	1	services	other	reputation	father	
14	GP	M	15	U	LE3	T	4	4	health	services	course	father	
15	GP	M	15	U	GT3	T	4	3	teacher	other	course	mother	

* P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

데이터 기술통계치요약

- 데이터설명 (stud_math_desc.doc참고)

school : 학교이름 (GP, MS)
sex : 성별 (F, M)
age : 나이 (15-22)
address : 주소 (Urban:도심, Rural:외곽)
famsize : 가족수 (LE3 :≤3, GT3: >3)
Medu : 엄마교육수준
Fedu: 아빠교육수준
traveltime: 통학시간: 1(15분이하),2,3,4(1시간이상)
studytime: 주중공부시간: 1(<2시간), 2(2-5시간), 3(5-10시간), 4(>10시간)
activties : 방과후활동(yes, no)
nursery:유치원다녔는지여부(yes,no)
internet : 집에서 인터넷사용(yes,no)
romantic : 이성교제여부(yes, no)
soout : 친구들과 외출 (1-5)
Dalc : 음주(1-5)
health : 건강상태 (1(매우나쁨)-5(매우좋음))
absences : 학교결석 (0-93)
타겟변수 : G3(최종성적, 0-20), G2(2학년), G1(1학년)

Attribute Information:

Attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course) datasets:
1 school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2 sex - student's sex (binary: 'F' - female or 'M' - male)
3 age - student's age (numeric: from 15 to 22)
4 address - student's home address type (binary: 'U' - urban or 'R' - rural)
5 famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6 Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7 Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€" 5th to 9th grade, 3 â€" secondary edu
8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€" 5th to 9th grade, 3 â€" secondary educa
9 Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
10 Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
11 reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12 guardian - student's guardian (nominal: 'mother', 'father' or 'other')
13 traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14 studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15 failures - number of past class failures (numeric: n if 1<=n<3, else 4)
16 schoolsup - extra educational support (binary: yes or no)
17 famsup - family educational support (binary: yes or no)
18 paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19 activities - extra-curricular activities (binary: yes or no)
20 nursery - attended nursery school (binary: yes or no)
21 higher - wants to take higher education (binary: yes or no)
22 internet - Internet access at home (binary: yes or no)
23 romantic - with a romantic relationship (binary: yes or no)
24 famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25 freetime - free time after school (numeric: from 1 - very low to 5 - very high)
26 goout - going out with friends (numeric: from 1 - very low to 5 - very high)
27 Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28 Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29 health - current health status (numeric: from 1 - very bad to 5 - very good)
30 absences - number of school absences (numeric: from 0 to 93)

these grades are related with the course subject, Math or Portuguese:
31 G1 - first period grade (numeric: from 0 to 20)
31 G2 - second period grade (numeric: from 0 to 20)
32 G3 - final grade (numeric: from 0 to 20, output target)

6-2. 데이터의 기술통계치요약

데이터 기술통계치요약

- 데이터 : 학생들의 학업성취도* (포루투칼의 고등학생 수학성적)

```
# lec6_2.r
# Data exploration : Numerical summary statistics

# set working directory
setwd("D:/tempstore/moocr")

### student math grade data ####

stud<-read.csv("stud_math.csv")

head(stud)
dim(stud)
str(stud)

attach(stud)
```

stud 데이터는 n=395관측치와 33개의 변수

```
> dim(stud)
[1] 395  33
> str(stud)
'data.frame':   395 obs. of  33 variables:
 $ school    : Factor w/ 2 levels "GP","MS": 1
1 ...
 $ sex       : Factor w/ 2 levels "F","M": 1 1
```

데이터 기술통계치요약

- summary(데이터이름) : 각 변수별로 요약통계량을 제공.

문자변수에 대해서는 빈도를 주고, 숫자변수에 대해서는 (최소값, 25%, 중위값, 평균, 75%, 최대값)을 제공

```
# 1-1 Numeriacl analytics
summary(stud)
```

```
> summary(stud)
 school   sex      age       address famsize   Pstatus
 GP:349   F:208   Min.   :15.0   R: 88   GT3:281   A: 41
 MS: 46   M:187   1st Qu.:16.0   U:307   LE3:114   T:354
                  Median :17.0
                  Mean   :16.7
                  3rd Qu.:18.0
                  Max.   :22.0

      Medu            Fedu            Mjob
 Min.   :0.000   Min.   :0.000   at_home : 59
 1st Qu.:2.000   1st Qu.:2.000   health  : 34
 Median :3.000   Median :2.000   other   :141
 Mean   :2.749   Mean   :2.522   services:103
 3rd Qu.:4.000   3rd Qu.:3.000   teacher : 58
 Max.   :4.000   Max.   :4.000
      Fjob            reason      guardian
 at_home : 20   course     :145   father: 90
 health  : 18   home       :109   mother:273
 other   :217   other      : 36   other : 32
 services:111   reputation:105

       G2              G3
 Min.   : 0.00   Min.   : 0.00
 1st Qu.: 9.00   1st Qu.: 8.00
 Median :11.00   Median :11.00
 Mean   :10.71   Mean   :10.42
 3rd Qu.:13.00   3rd Qu.:14.00
 Max.   :19.00   Max.   :20.00
```


데이터 기술통계치요약

- mean(변수) : 평균
- sd(변수) : 표준편차 (분산의 제곱근)
- var(변수) : 분산

```
mean(G3)
sd(G3)
sqrt(var(G3))
```

```
> mean(G3)
[1] 10.41519
> sd(G3)
[1] 4.581443
> sqrt(var(G3))
[1] 4.581443
```

통계함수	설명
mean(x)	평균
median(x)	중앙값
sd(x)	표준편차
mad(x)	Median absolute deviation
var(x)	분산

데이터 기술통계치요약

- 특정변수들에 대한 요약통계량 : vars<-c("변수1", "변수2", "변수3")

```
# creating interested variable list
vars<-c("G1", "G2", "G3")
head(stud[vars])
summary(stud[vars])
```

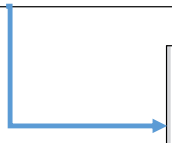
```
> vars<-c("G1", "G2", "G3")
> head(stud[vars])
  G1 G2 G3
1  5  6  6
2  5  5  6
3  7  8 10
4 15 14 15
5  6 10 10
6 15 15 15
> summary(stud[vars])
       G1              G2              G3
 Min.   : 3.00   Min.   : 0.00   Min.   : 0.00
 1st Qu.: 8.00   1st Qu.: 9.00   1st Qu.: 8.00
 Median :11.00   Median :11.00   Median :11.00
 Mean   :10.91   Mean   :10.71   Mean   :10.42
 3rd Qu.:13.00   3rd Qu.:13.00   3rd Qu.:14.00
 Max.   :19.00   Max.   :19.00   Max.   :20.00
```

stud데이터는 33개의 변수를 가짐!!
=> 특정변수들에 대해 탐색하고자 할때

데이터 기술통계치요약

- 함수 describe를 사용한 데이터요약통계량 (psych 패키지 필요)

```
# descriptive statistics using "psych" package
install.packages("psych")
library(psych)
# require "psych" for "describe" function
describe(stud[vars])
```



```
> describe(stud[vars])
   vars   n  mean   sd median trimmed  mad min max range  skew
G1    1 395 10.91 3.32     11   10.80 4.45   3  19    16  0.24
G2    2 395 10.71 3.76     11   10.84 2.97   0  19    19 -0.43
G3    3 395 10.42 4.58     11   10.84 4.45   0  20    20 -0.73
   kurtosis   se
G1    -0.71 0.17
G2     0.59 0.19
G3     0.37 0.23
```

- sapply함수를 사용

```
# sapply function
sapply(stud[vars], mean)
```

```
> sapply(stud[vars], mean)
      G1       G2       G3
10.90886 10.71392 10.41519
```

데이터 기술통계치요약

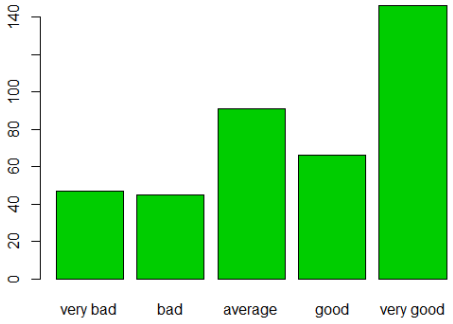
- 범주형 변수의 요약 : table(변수이름)

```
# categorical data
table(health)
```

```
> table(health)
health
  1   2   3   4   5
 47  45  91  66 146
```

막대그림 (이름주기)

```
health_freq<-table(health)
names(health_freq) <- c ("very bad", "bad", "average", "good",
                         "very good")
barplot(health_freq, col=3)
```



데이터 기술통계치요약

- 범주형 변수의 요약 : table(변수1, 변수2)

2*2 분할표

```
# 2*2 contingency table
table(health,studytime)
```

```
> table(health,studytime)
      studytime
health  1  2  3  4
     1 11 29  3  4
     2  9 27  6  3
     3 20 43 18 10
     4 15 30 17  4
     5 50 69 21  6
```

Wk6-3 : 그래프를 이용한 데이터탐색

데이터 설명

- **stud_math 데이터 : 포루투칼의 고등학생 수학성적 (stud_math_desc.doc참고)**

school : 학교이름 (GP, MS)
sex : 성별 (F, M)
age : 나이 (15-22)
address : 주소 (Urban:도심, Rural:외곽)
famsize : 가족수 (LE3 :≤3, GT3: >3)
Medu : 엄마교육수준
Fedu: 아빠교육수준
traveltime: 통학시간: 1(15분이하),2,3,4(1시간이상)
studytime: 주중공부시간: 1(<2시간), 2(2-5시간), 3(5-10시간), 4(>10시간)
activties : 방과후활동(yes, no), freetime : 자유시간(1-5)
nursery:유치원다녔는지여부(yes,no)
internet : 집에서 인터넷사용(yes,no)
romantic : 이성교제여부(yes, no)
soout : 친구들과 외출 (1-5)
Dalc : 음주(1-5)
health : 건강상태 (1(매우나쁨)-5(매우좋음))
absences : 학교결석 (0-93)
타겟변수 : G3(최종성적, 0-20), G2(2학년), G1(1학년)

Attribute Information:

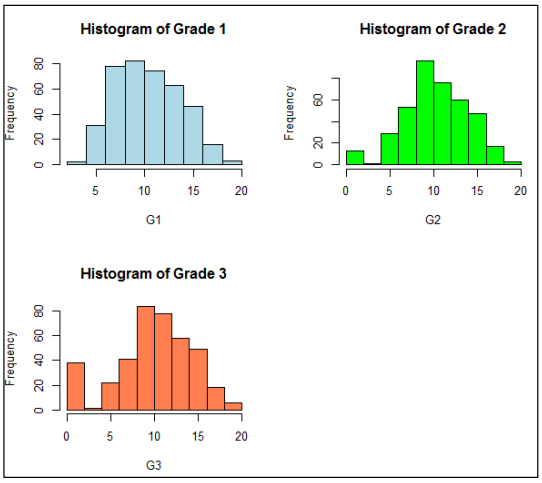
Attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course) datasets:
1 school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2 sex - student's sex (binary: 'F' - female or 'M' - male)
3 age - student's age (numeric: from 15 to 22)
4 address - student's home address type (binary: 'U' - urban or 'R' - rural)
5 famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6 Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7 Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary edu
8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary educa
9 Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
10 Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
11 reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12 guardian - student's guardian (nominal: 'mother', 'father' or 'other')
13 traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14 studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15 failures - number of past class failures (numeric: n if 1<=n<3, else 4)
16 schoolsup - extra educational support (binary: yes or no)
17 famsup - family educational support (binary: yes or no)
18 paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19 activities - extra-curricular activities (binary: yes or no)
20 nursery - attended nursery school (binary: yes or no)
21 higher - wants to take higher education (binary: yes or no)
22 internet - Internet access at home (binary: yes or no)
23 romantic - with a romantic relationship (binary: yes or no)
24 famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25 freetime - free time after school (numeric: from 1 - very low to 5 - very high)
26 goout - going out with friends (numeric: from 1 - very low to 5 - very high)
27 Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28 Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29 health - current health status (numeric: from 1 - very bad to 5 - very good)
30 absences - number of school absences (numeric: from 0 to 93)

these grades are related with the course subject, Math or Portuguese:
31 G1 - first period grade (numeric: from 0 to 20)
31 G2 - second period grade (numeric: from 0 to 20)
32 G3 - final grade (numeric: from 0 to 20, output target)

그래프를 이용한 데이터탐색

- 히스토그램 (1학년, 2학년, 3학년 성적의 분포)

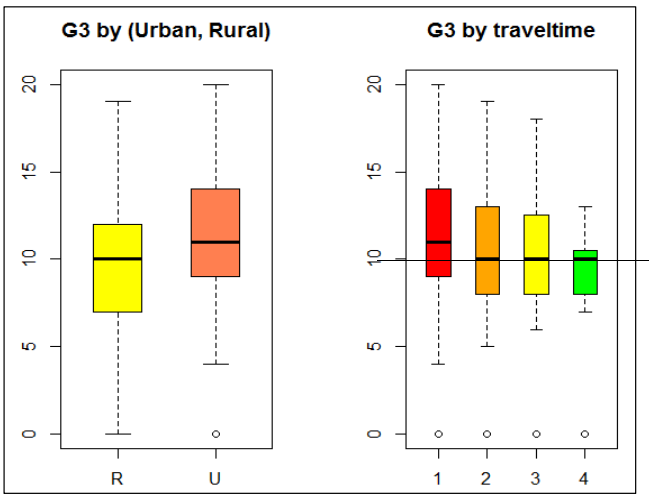
```
# lec6_3.r
# Normality test, confidence interval

# set working directory
setwd("D:/tempstore/moocr")

### student math grade data ####

stud<-read.csv("stud_math.csv")

head(stud)
dim(stud)
str(stud)

attach(stud)

# 1. histogram with color and title, legend
par(mfrow=c(2,2))
hist(G1, breaks = 10, col = "lightblue", main="Histogram of Grade 1" )
hist(G2, breaks = 10, col = "green", main="Histogram of Grade 2" )
hist(G3, breaks = 10, col = "coral", main="Histogram of Grade 3" )
```



그래프를 이용한 데이터탐색

- 상자그림 (거주지역에 따른 G3, 통학시간에 따른 G3)

```
# 2. boxplot
par(mfrow=c(1,2))
boxplot(G3~address, boxwex = 0.5, col = c("yellow", "coral"), main="G3 by
boxplot(G3~traveltime, boxwex = 0.5, col = c("red","orange","yellow","gree
```



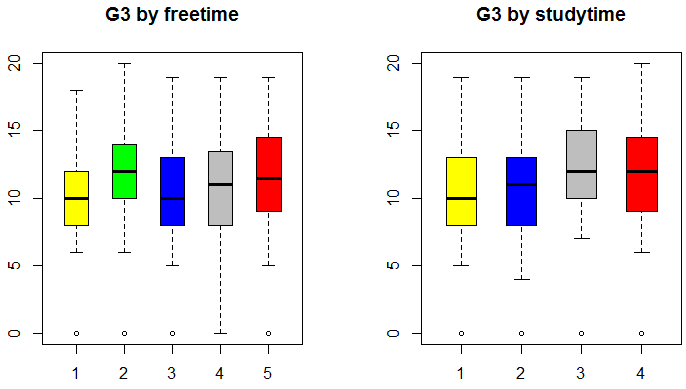
(1) 도심지역 학생들 성적이 외곽지역 학생들보다 높다

(2) 통학시간이 짧은(15분이내)의 학생들의 성적이 더 높다

그래프를 이용한 데이터탐색

• 상자그림 (자유시간에 따른 G3, 공부시간에 따른 G3)

```
# boxplot
par(mfrow=c(1,2))
# academic achievement by freetime
# 1 - very low to 5 - very high
boxplot(G3~freetime, boxwex = 0.5, col = c("yellow", "green", "blue
# academic achievement by studytime
# 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hou
boxplot(G3~studytime, boxwex = 0.5, col = c("yellow", "blue","grey"
```



(1) 방과후 자유시간에 따른 G3의 차이 : 자유시간이 적은편(low)이라고 응답한 학생들의 성적이 다소 높은데..특별히 해석하기는 어려움

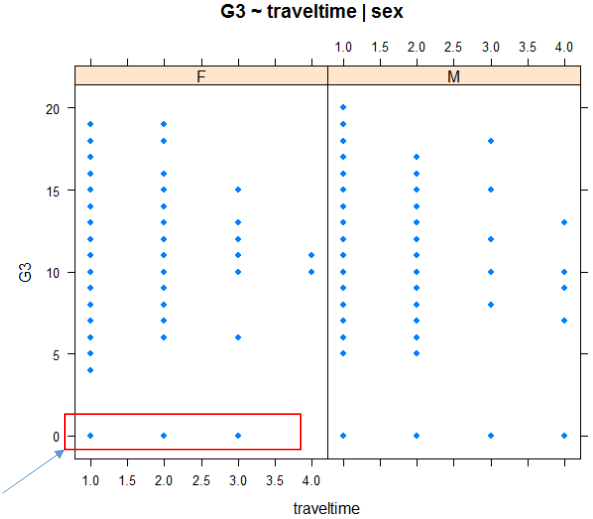
(2) 주중공부시간이 5시간이상 (3: 5-10시간, 4: 10시간이상)인 학생들의 성적이 높은편임

6-3. 그래프를 이용한 데이터탐색

그래프를 이용한 데이터탐색

• 통학시간과 최종성적(G3)의 멀티패널 그림 (성별) : lattice 패키지 사용

```
# lattice package
library(lattice)
xyplot(G3 ~ traveltime | sex , data = stud, pch=16, main = "G3 ~ travelt
```



(1) 학생들 대부분은 30분이내의 통학거리에 있으며,

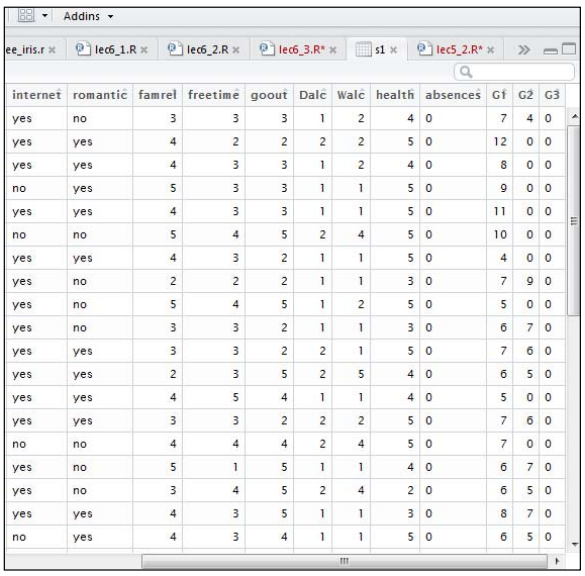
(2) 통학거리가 짧은 학생들의 성적평균이 다소 높게 나타남

(3) 통학거리가 1시간 이상은 표본이 상대적으로 적음

0점인 데이터 확인, 점검 필요

그래프를 이용한 데이터탐색

- G3=0인 데이터 (n=38명)



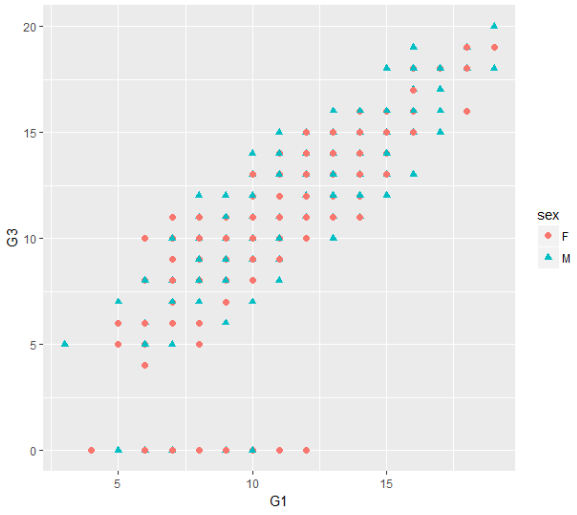
internet	romantic	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
yes	no	3	3	3	1	2	4	0	7	4	0
yes	yes	4	2	2	2	2	5	0	12	0	0
yes	yes	4	3	3	1	2	4	0	8	0	0
no	yes	5	3	3	1	1	5	0	9	0	0
yes	yes	4	3	3	1	1	5	0	11	0	0
no	no	5	4	5	2	4	5	0	10	0	0
yes	yes	4	3	2	1	1	5	0	4	0	0
yes	no	2	2	2	1	1	3	0	7	9	0
yes	no	5	4	5	1	2	5	0	5	0	0
yes	no	3	3	2	1	1	3	0	6	7	0
yes	yes	3	3	2	2	1	5	0	7	6	0
yes	yes	2	3	5	2	5	4	0	6	5	0
yes	yes	4	5	4	1	1	4	0	5	0	0
yes	yes	3	3	2	2	2	5	0	7	6	0
no	no	4	4	4	2	4	5	0	7	0	0
yes	no	5	1	5	1	1	4	0	6	7	0
yes	no	3	4	5	2	4	2	0	6	5	0
yes	yes	4	3	5	1	1	3	0	8	7	0
no	yes	4	3	4	1	1	5	0	6	5	0

```
# data (G3=0)
s1<-subset(stud, G3==0)
#ggplot(data=s1, aes(factor(s1$add
#ggplot(data=s1, aes(factor(s1$int
```


그래프를 이용한 데이터탐색

- 산점도 (ggplot2 패키지의 ggplot이용)

```
# ggplot2 package
library(ggplot2)
# scatterplot for G1 and G3 by sex
ggplot(stud, aes(x=G1, y=G3, color=sex, shape=sex)) +
  geom_point(size=2)
```

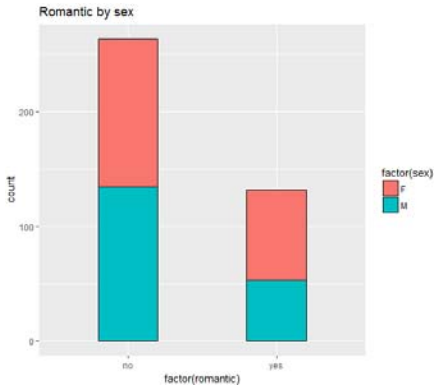


성별에 따른 차이는 없음

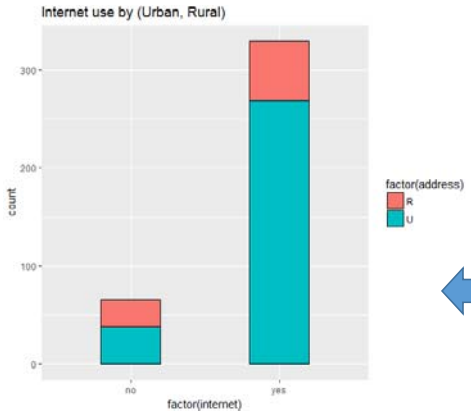
그래프를 이용한 데이터탐색

• 막대그림 (ggplot2 패키지의 ggplot이용)

```
# bar chart for romantic by sex
ggplot(data=stud, aes(factor(romantic)))+geom_bar(aes(fill=factor(sex)),
```



연애경험 있는경우 여학생 비율이 높음



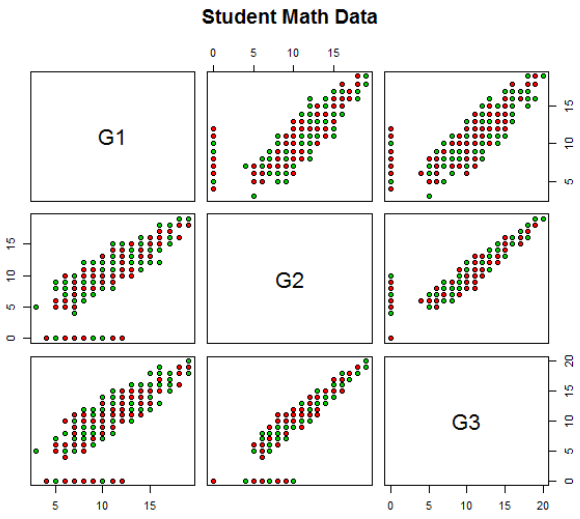

```
# bar chart for internet use by (Urban, Rural)
ggplot(data=stud, aes(factor(internet)))+geom_bar(aes(fill=factor(address)
```

인터넷사용자 중에는 도심지역에 사는 경우가 훨씬 높음

그래프를 이용한 데이터탐색

• pairwise scatterplot : pairs(변수리스트)

```
# new variable lists
vars1<-c("G1", "G2", "G3")
# pariwise plot
pairs (stud[vars1], main = "Student Math Data",
       pch = 21,bg = c ("red","green3")[ unclass (stud$sex)])
```



(1) G1, G2, G3간의 상관성은 매우높다

(2) 성별간 차이는 없다

Wk6-4 : 데이터의 정규성검정과 신뢰구간

데이터 설명

- **stud_math 데이터 : 포루투칼의 고등학생 수학성적 (stud_math_desc.doc참고)**

school : 학교이름 (GP, MS)
sex : 성별 (F, M)
age : 나이 (15-22)
address : 주소 (Urban:도심, Rural:외곽)
famsize : 가족수 (LE3 :≤3, GT3: >3)
Medu : 엄마교육수준
Fedu: 아빠교육수준
traveltime: 통학시간: 1(15분이하),2,3,4(1시간이상)
studytime: 주중공부시간: 1(<2시간), 2(2-5시간), 3(5-10시간), 4(>10시간)
activties : 방과후활동(yes, no), freetime : 자유시간(1-5)
nursery:유치원다녔는지여부(yes,no)
internet : 집에서 인터넷사용(yes,no)
romantic : 이성교제여부(yes, no)
soout : 친구들과 외출 (1-5)
Dalc : 음주(1-5)
health : 건강상태 (1(매우나쁨)-5(매우좋음))
absences : 학교결석 (0-93)
타겟변수 : G3(최종성적, 0-20), G2(2학년), G1(1학년)

Attribute Information:

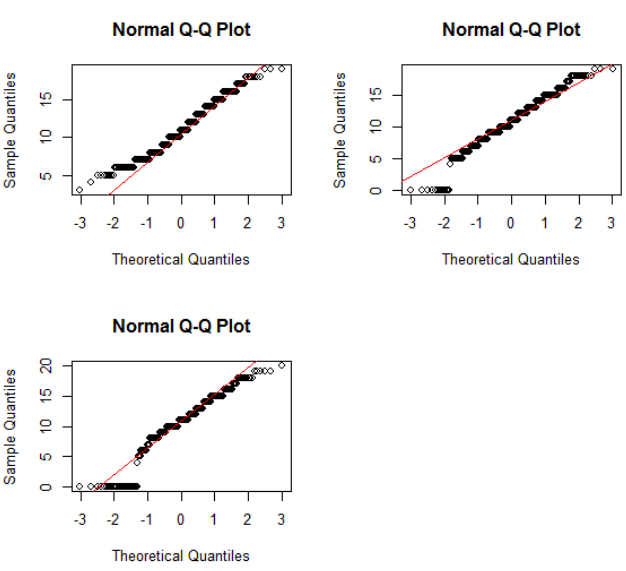
Attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course) datasets:
1 school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2 sex - student's sex (binary: 'F' - female or 'M' - male)
3 age - student's age (numeric: from 15 to 22)
4 address - student's home address type (binary: 'U' - urban or 'R' - rural)
5 famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6 Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7 Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary edu
8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary educa
9 Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
10 Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
11 reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12 guardian - student's guardian (nominal: 'mother', 'father' or 'other')
13 traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14 studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15 failures - number of past class failures (numeric: n if 1<=n<3, else 4)
16 schoolsup - extra educational support (binary: yes or no)
17 famsup - family educational support (binary: yes or no)
18 paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19 activities - extra-curricular activities (binary: yes or no)
20 nursery - attended nursery school (binary: yes or no)
21 higher - wants to take higher education (binary: yes or no)
22 internet - Internet access at home (binary: yes or no)
23 romantic - with a romantic relationship (binary: yes or no)
24 famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25 freetime - free time after school (numeric: from 1 - very low to 5 - very high)
26 goout - going out with friends (numeric: from 1 - very low to 5 - very high)
27 Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28 Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29 health - current health status (numeric: from 1 - very bad to 5 - very good)
30 absences - number of school absences (numeric: from 0 to 93)

these grades are related with the course subject, Math or Portuguese:
31 G1 - first period grade (numeric: from 0 to 20)
31 G2 - second period grade (numeric: from 0 to 20)
32 G3 - final grade (numeric: from 0 to 20, output target)

데이터의 정규성검정

• 정규확률도 (Normal Q-Q plot) : 데이터가 정규분포하는가?

```
# 1-2 Testing for normality
# multiple plot (2 by 2)
par(mfrow=c(2,2))
#Quantile plot
qqnorm(G1)
qqline(G1, col = 2, cex=7)

qqnorm(G2)
qqline(G2, col = 2, cex=7)

qqnorm(G3)
qqline(G3, col = 2, cex=7)
```



qqline의 디폴트는 정규분포의 1사분위, 3사분위를 직선

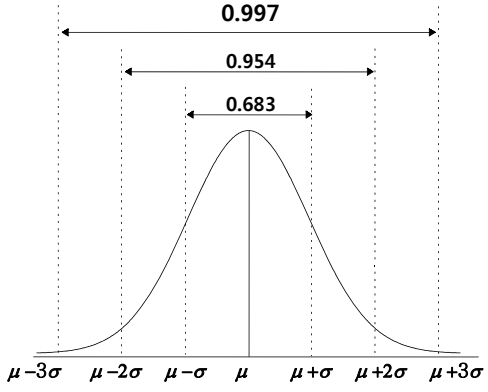
qqline(y, distribution = qqnorm, probs = c(0.25, 0.75))

데이터의 정규성검정

▪ 정규분포(Normal distribution)

- 확률변수 X 의 확률밀도함수가 다음과 같이 주어질 때 X 는 정규분포 N(μ, σ2) 을 따른다고 함

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}},\quad -\infty<x<\infty$$



• $\mu\pm\sigma$에 대한 설명

(예) 중1학년 평균신장 150cm, 편차가 5cm라고하면

1시그마범위 : 150±5 => (145cm-155cm)

2시그마범위 : 150±10 => (140cm-160cm)

3시그마범위 : 150±15 => (135cm-165cm)

어떤 학생의 신장이 167cm이면 그학생의 신장은 $\pm3\sigma$를 벗어나 있다라고 말할수 있음

데이터의 정규성검정

- 정규분포 적합성검정 : 데이터가 정규분포 하는지에 대한 검정

(1) Shapiro-Wilks검정

```
#Shapiro-Wilks test
shapiro.test(G3)
```

```
> shapiro.test(G3)

        Shapiro-Wilk normality test

data:  G3
W = 0.92873, p-value = 8.836e-13
```

G3는 정규분포한다고 볼수없다 (p-value~0)

(2) Anderson-Darling검정 (추가패키지 필요)

```
#Anderson-Darling test require installing package "nortest"
install.packages('nortest')
library(nortest)
ad.test(G3)
```

```
> ad.test(G3)

        Anderson-Darling normality test

data:  G3
A = 8.3032, p-value < 2.2e-16
```

G3는 정규분포한다고 볼수없다 (p-value~0)

데이터 시뮬레이션

- 확률분포함수로부터 데이터생성

분포함수	설명
binom(x)	이항분포 rbinom(5, size=100,prob=.2)
exp(x)	지수분포
gamma(x)	감마분포 rgamma(5, shape=3, rate=2)
norm(x)	정규분포 rnorm(50, mean=10, sd=5)
pois(x)	포아송분포 rpois(n, lambda)
unif(x)	균일분포 runif(30)

p : 누적함수
d : 확률밀도함수
q : quantile 함수
r : 랜덤넘버의 생성

데이터 시뮬레이션

- 확률분포함수로부터 데이터생성

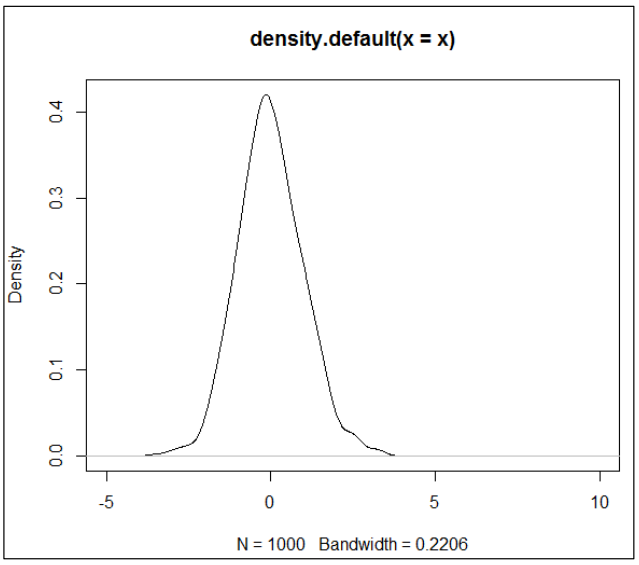
```
# data simulation
# Simulation examples
runif(5,min=1,max=5)
rnorm(5,mean=5,sd=1)
rgamma(5,shape=3,rate=2)
rbinom(5,size=100,prob=.2)
```

```
> runif(5,min=1,max=5)
[1] 3.001213 4.377249 3.358988 3.343567 2.458019
> rnorm(5,mean=5,sd=1)
[1] 4.454306 5.583966 6.178074 5.799947 5.294765
> rgamma(5,shape=3,rate=2)
[1] 1.496415 1.206023 1.987550 1.623255 1.513995
> rbinom(5,size=100,prob=.2)
[1] 25 18 23 19 18
```


데이터 시뮬레이션

- 정규분포로부터 데이터생성, 밀도함수 그래프

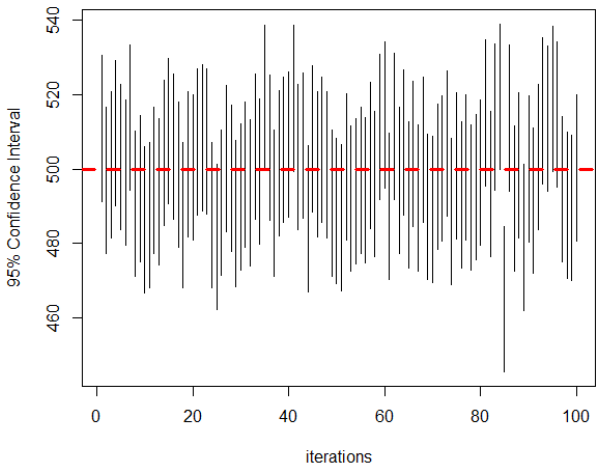
```
# from normal distribution
x<-rnorm(1000)
plot(density(x),xlim=c(-5,10))
```



데이터 시뮬레이션

• 데이터 생성 (정규분포(평균=500, 편차=100)에서 100개 데이터 생성)

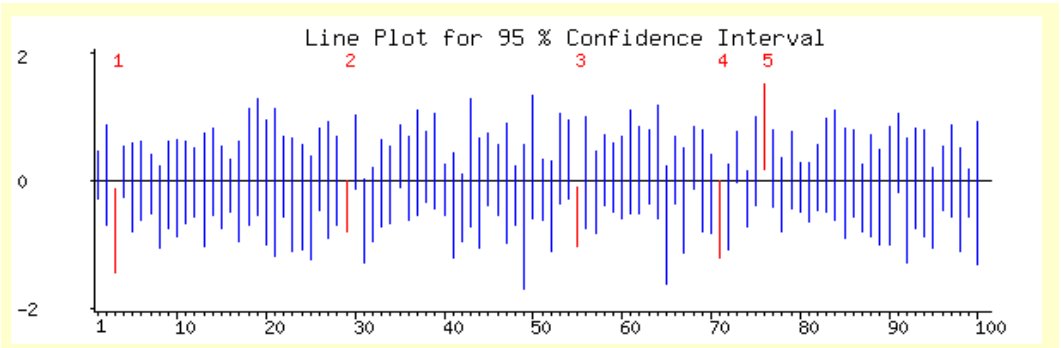
```
# confidence interval of normal distribution
nreps <- 100
ll <- numeric(nreps)
ul <- numeric(nreps)
n <- 100
mu <- 500
sigma <- 100
for(i in 1:nreps) {
  set.seed(i)
  x <- rnorm(n, mu, sigma)
  ll[i] <- mean(x) - qnorm(0.975)*sqrt(sigma^2/n)
  ul[i] <- mean(x) + qnorm(0.975)*sqrt(sigma^2/n)
}

# Draw 95% confidence interval
par(mfrow=c(1,1))
plot(1:100, ylim=c(min(ll), max(ul)),
     ylab="95% Confidence Interval", xlab="iterations")

for(i in 1:100) lines(c(i, i), c(ll[i], ul[i]))
abline(h=mu, col="red", lty=2, lwd=3)
```



신뢰구간

▪ 신뢰구간

- 신뢰구간은 실제 모수(parameter=모평균, 모분산 등)를 추정하는데 몇 퍼센트의 확률로 그 신뢰구간이 실제 모수를 포함하게 될 것이냐 하는 것이다. 예를 들어 모평균(μ)의 추정을 위해 100번의 sampling을 통해 표본평균과 표본분산을 구하여 100개의 신뢰구간을 얻었을 때, 그 100개의 신뢰구간 중 95개에 모평균(μ)이 포함되게 설정된 신뢰구간을 95% 신뢰구간이라고 한다.



신뢰구간의 의미

- 신뢰수준, 표본오차

" 전국의 유권자 1,500명을 조사한 결과에 의하면 A 후보 지지율은 45%이며, 95% 신뢰수준에서 오차한계는 3.5%이다."

==> 지지도에 대한 95% 신뢰구간 :

표본 지지율±오차한계 ⇔ 45%±3.5% ⇔ (41.5%, 48.5%)

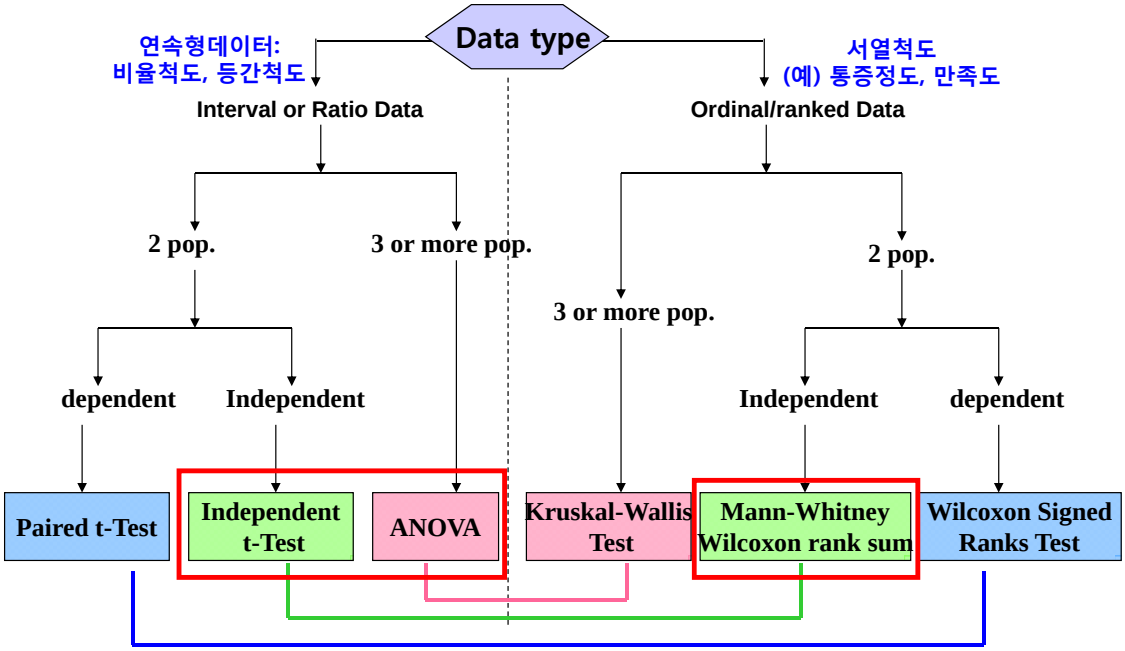
(예) A후보자 득표율 24% (허용오차 ±2%, 95% 신뢰수준)이라고 하면,
그 의미는 n명을 표본으로 하여 random 하게 반복 조사하였을 때
100번 중 95번은 22%-26% 득표율을 가질 것이라는 의미

	R을 이용한 통계분석 (Week7)
wk7-1	두 그룹간 평균비교(t-test)
wk7-2	짝을 이룬 그룹간 비교(paired t-test)
wk7-3	분산분석(ANOVA)
wk7-4	이원분산분석(two-way ANOVA)

7-1. 두그룹간 평균비교(t-test)

Wk7-1 : 두그룹간 평균비교 (t-test)

모집단간 차이에 대한 검정 (모수/비모수 검정)



1. 단일표본의 평균검정

- 단일표본의 평균검정 : t.test(변수, mu=검정하고자 하는 평균값)

가설 1 : G3(최종성적)의 평균은 10인가? H_0 (null Hypothesis: 귀무가설) : μ=10

```
# lec7_1.r
# t-test for two sample means

# set working directory
setwd("D:/tempstore/moocr")

### student math grade data ###
stud<-read.csv("stud_math.csv")

head(stud)
dim(stud)
str(stud)

attach(stud)
```

```
# single t-test : to test whether or not mean of G3 is 10
t.test(G3, mu=10)
```

```
> t.test(G3, mu=10)

        One Sample t-test

data:  G3
t = 1.8011, df = 394, p-value = 0.07245
alternative hypothesis: true mean is not equal to 10
95 percent confidence interval:
  9.961992 10.868388
sample estimates:
mean of x
 10.41519
```

결론: α=0.05에서는 G3의 평균이 10이라고 할 수 있는 근거가있다

t검정통계량, 자유도, p-value

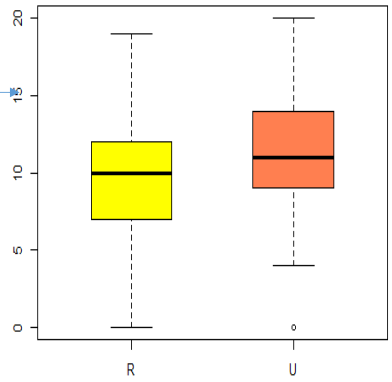
Ha 대립가설 : 모평균은 10이 아니다
95% 신뢰구간 : (9.96, 10.86)
표본평균값 : 10.415

2. 두 집단의 평균검정

• 두집단 표본평균 비교검정 : t.test(타겟변수~범주형변수, data=)

가설 2 : 거주지역(R, U)에 따른 G3(최종성적) 평균에 차이가 있는가? (양측검정)

```
# two sample t-test :
# to test whether or not mean of G3 is same between Urban and Rural
t.test(G3~address, data=stud)
boxplot(G3~address, boxwex = 0.5, col = c("yellow", "coral"))
```




```
> t.test(G3~address, data=stud)

        Welch Two Sample t-test

data:  G3 by address
t = -2.1101, df = 140.91, p-value = 0.03661
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.25240320 -0.07340373
sample estimates:
mean in group R mean in group U
       9.511364       10.674267
```

양측검정

p-value=0.03으로 유의수준 0.05 (α=0.05)에서 거주지역에 따라 G3는 유의한 차이가 있다고 할 수 있다.

2. 두 집단의 평균검정

• 두집단 표본평균 비교검정 : t.test(연속변수~범주형변수, data=)

단측검정 : 기각역이 한쪽에만 있는 경우, alternative=c("less") 혹은 alternative=c("greater")

```
# alternative H : mu(Rural)< mu(Urban)
t.test(G3~address, data=stud, alternative = c("less"))

help(t.test)
```

$H_0 : \mu_R = \mu_U \ (\mu_R - \mu_U = 0)$
$H_1 : \mu_R < \mu_U \ (\mu_R - \mu_U < 0)$

```
> t.test(G3~address, data=stud, alternative = c("less"))

        Welch Two Sample t-test

data:  G3 by address
t = -2.1101, df = 140.91, p-value = 0.01831
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.2504199
sample estimates:
mean in group R mean in group U
       9.511364       10.674267
```

p-value=0.018로 유의수준을 0.05로 했을때 성적(Rural)<성적(Urban)이라고 할수 있다

2. 두 집단의 평균검정

- 두집단 표본평균 비교 도움말 보기 : help(t.test)

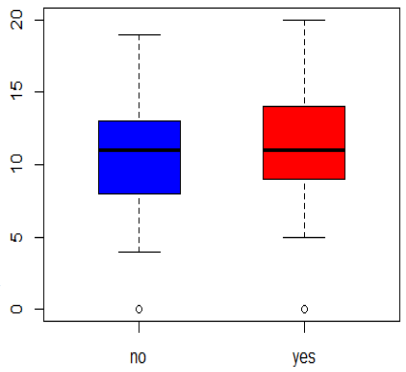
```
help(t.test)
```

2. 두 집단의 평균검정

- 두집단 표본평균 비교검정 : t.test(타겟변수~범주형변수, data=)

가설 3 : 방과후 활동여부(yes, no)에 따른 G3(최종성적) 평균에 차이가 있는가?

```
## example 2
# to test whether or not mean of G3 is equal for activities
t.test(G3~activities, data=stud)
boxplot(G3~activities, boxwex = 0.5, col = c("blue", "red"))
```



상자그림(Boxplot)에서 보면 방과후 활동여부는 G3(성적)에 뚜렷한 차이를 볼수 없음

2. 두 집단의 평균검정

• 두집단 표본평균 비교검정 : t.test(타겟변수~범주형변수, data=)

가설 3 : 방과후 활동여부(yes, no)에 따른 G3(최종성적) 평균에 차이가 있는가?

t-test 검정통계량에 의한 검정결과

```
> t.test(G3~activities, data=stud)

        Welch Two Sample t-test

data:  G3 by activities
t = -0.31944, df = 392.98, p-value = 0.7496
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.0542623  0.7595503
sample estimates:
 mean in group no mean in group yes
         10.34021          10.48756
```

p-value

양측검정

p-value=0.75는 유의수준 0.05보다 큽니다.
즉 검정통계량의 값이 기각역에 있지 않다.
=> 귀무가설(평균이같다)를 기각할수 없음
=> 방과후활동여부는 G3에 유의한 영향이 없다!

평균(G3(방과후활동없음)-G3(방과후활동)) 차이에 대한 신뢰구간 = (-1.05, 0.79)

0

신뢰구간 사이에 0값이 있다는것은 차이가 없음을 의미!!

3. 두 집단의 비모수적 비교검정

• 두 모집단의 비모수적 방법 (Wilcoxon rank sum Test) : wilcox.test(x,y)

wilcox.test는 타겟변수가 등간척도(통증정도, 만족도, ..)일때 사용할 수 있다

```
# Wilcoxon signed-rank test
# wilcox.test(G3, mu=10)
wilcox.test(G3~address)
```

wilcox.test(타겟변수~범주형변수)

```
> wilcox.test(G3~address)

        Wilcoxon rank sum test with continuity correction

data:  G3 by address
W = 11278, p-value = 0.01776
alternative hypothesis: true location shift is not equal to 0
```

```
help(wilcox.test)
```

Wilcoxon Rank Sum and Signed Rank Tests

Description

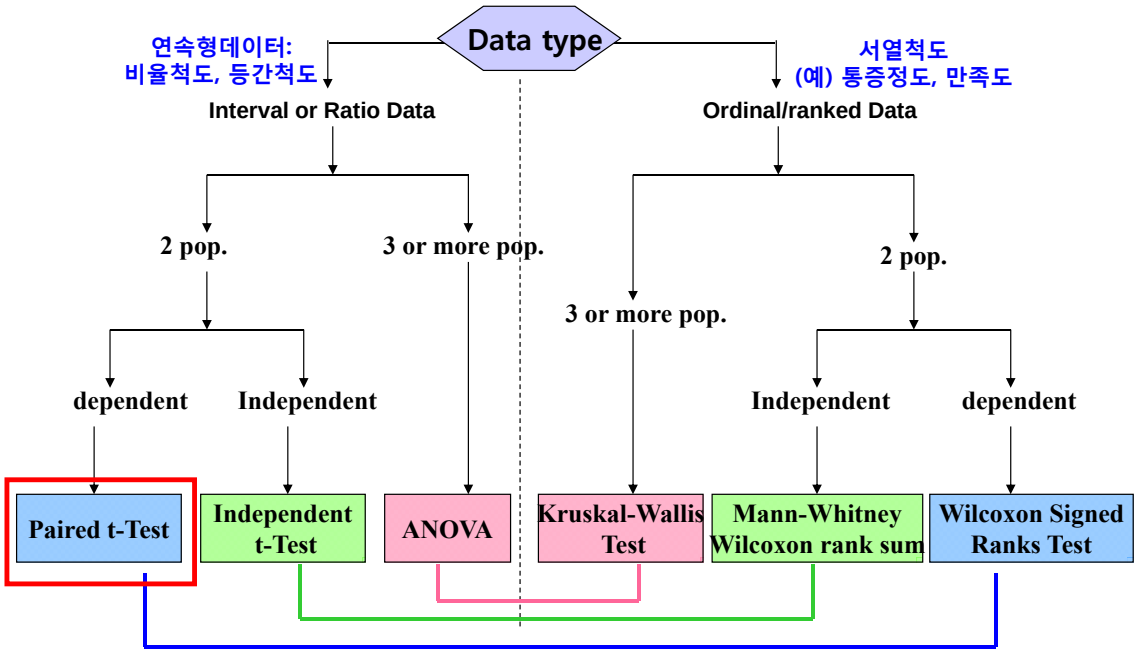
Performs one- and two-sample Wilcoxon tests on vectors of data; the latter is also known as 'Mann-Whitney' test.

Usage

```
wilcox.test(x, ...)

## Default S3 method:
wilcox.test(x, y = NULL,
            alternative = c("two.sided", "less", "greater"),
            mu = 0, paired = FALSE, exact = NULL, correct = TRUE,
            conf.int = FALSE, conf.level = 0.95, ...)
```

Wk7-2 : 짝을 이룬 그룹간 비교 (paired t-test)

모집단간 차이에 대한 검정 (모수/비모수 검정)



짝을 이룬 그룹간 비교 (paired t-test)

- 특정 처리(treatment)의 효과를 비교분석할 때 사용
- 동일한 실험표본 : before & after 측정

예제 :

(1)혈압강하제의 투약효과

(2)방과후프로그램의 성과 (학업흥미도)

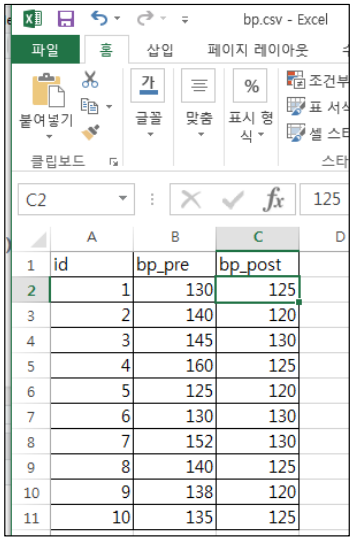
(3)다이어트 프로그램의 효과

(4)직무교육후의 생산성 향상의 효과

평균, 편차 계산=> 검정통계량

짝을 이룬 그룹간 비교 (paired t-test)

- 예제 1 : 고혈압 환자 10명에게 혈압강하제를 12주동안 투여한 후 복용전의 혈압과 복용후의 혈압을 비교하였다. 새로운 혈압강하제가 효과가 있다고 할수 있는가?



	A	B	C	D
1	id	bp_pre	bp_post	
2	1	130	125	
3	2	140	120	
4	3	145	130	
5	4	160	125	
6	5	125	120	
7	6	130	130	
8	7	152	130	
9	8	140	125	
10	9	138	120	
11	10	135	125	

짝을 이룬 그룹간 비교 (paired t-test)

짝을 이룬 그룹간 비교 (paired t-test)

- paired t-test : t.test(before, after, mu=0, paired=T))

```
# lec7_2.r

# paired t-test for two sample means

# set working directory
setwd("D:/tempstore/moocr")

## example 1: blood pressure data
bp<-read.csv("bp.csv")
attach(bp)

# paired t-test (two-sided)
t.test(bp_pre, bp_post, mu=0,  paired=T)
```

양측검정 : $H_0 : \mu_{(dif)} = 0$, $H_1 : \mu_{(dif)} \neq 0$
유의한 차이가 있는지 없는지에 대한 검정

```
> t.test(bp_pre, bp_post, mu=0,  paired=T)

        Paired t-test

data:  bp_pre and bp_post
t = 4.5095, df = 9, p-value = 0.001469
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  7.226228 21.773772
sample estimates:
mean of the differences
                  14.5
```

t검정통계량, 자유도, p-value

p-value=0.0015 (매우 작음)
유의수준 0.05 (α=0.05)보다 작으므로
H_0 를 기각=> 따라서 투약전과 투약후의
혈압에 유의한 차이가 있다고 볼 수 있음

짝을 이룬 그룹간 비교 (paired t-test)

- paired t-test의 검정통계량

```
> t.test(bp_pre, bp_post, mu=0,  paired=T)

        Paired t-test

data:  bp_pre and bp_post
t = 4.5095, df = 9, p-value = 0.001469
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  7.226228 21.773772
sample estimates:
mean of the differences
                  14.5
```

t검정통계량, 자유도, p-value

$$t=4.5095 = \frac{(\text{평균(Dif)}) - 0}{\text{편차}(Dif)/\sqrt{n}} = \frac{14.5 - 0}{10.168/\sqrt{10}}$$

짝을 이룬 그룹간 비교 (paired t-test)

- paired t-test : t.test(before, after, mu=0, alternative="greater", paired=T))

```
# paired t-test (one-sided)
t.test(bp_pre, bp_post, mu=0, alternative="greater", paired=T)
```

단측검정 : $H_0 : \mu_{(dif)} = 0$, $H_1 : \mu_{(dif)} > 0$
혈압(투약전-투약후)의 차이가 0보다 큰가?

```
> # paired t-test (one-sided)
> t.test(bp_pre, bp_post, mu=0, alternative="greater", paired=T)

        Paired t-test

data:  bp_pre and bp_post
t = 4.5095, df = 9, p-value = 0.0007344
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 8.605783      Inf
sample estimates:
mean of the differences
                  14.5
```

p-value=0.0007 (매우 작음)
유의수준 0.05 (α=0.05)보다 작으므로 H_0 를 기각=> 따라서 투약효과가 매우 유의하다고 볼 수 있다 (즉, 투약전보다 투약후의 혈압이 유의하게 낮아진다는 것이 검정됨)

1. 짝을 이룬 그룹간 비교 (paired t-test)

- 예제 2* : 비만 대상자들(성인)에게 12주동안 극저 칼로리 식이요법(very low-calorie diet: VLCD)을 실시한 후 그 효과를 비교. 이 프로그램이 체중감소에 효과가 있다고 할 수 있는가?

weight.csv

	A	B	C
	id	wt_pre	wt_post
1	1	117.3	83.3
2	2	111.4	85.9
3	3	98.6	75.8
4	4	104.3	82.9
5	5	105.4	82.3
6	6	100.4	77.7
7	7	81.7	62.7
8	8	89.5	69
9	9	78.2	63.9

very low-calorie**	<=800 calories/day
low-calorie	1,000–1,200 calories/day for a woman 1,200–1,600 calories/day for a man
standard-calorie	2000 calories/day

* 박미라, 이재원, 의학데이터의 통계분석, 자유아카데미
* https://www.niddk.nih.gov/health-information/weight-management/very-low-calorie-diets

짝을 이룬 그룹간 비교 (paired t-test)

• 예제 2 : 극저 칼로리 식이요법(very low-calorie diet: VLCD)의 효과

```
## example 2: Very Low-calroie diet
diet<-read.csv("weight.csv")
attach(diet)

# paired t-test (two-sided)
t.test(wt_pre, wt_post, mu=0,  paired=T)
```

양측검정 : $H_0 : \mu_{(dif)} = 0$, $H_1 : \mu_{(dif)} \neq 0$
극저칼로리 식이요법이 체중감량에 유의한 효과가 있는지 없는지에 대한 검정

```
> t.test(wt_pre, wt_post, mu=0,  paired=T)

        Paired t-test

data:  wt_pre and wt_post
t = 12.74, df = 8, p-value = 1.357e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 18.50003 26.67775
sample estimates:
mean of the differences
                22.58889
```

p-value=0.000001357

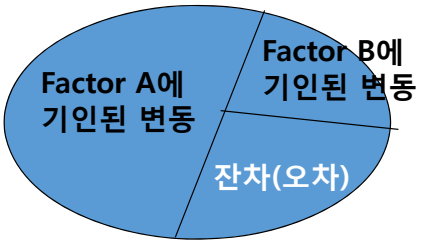
참고 : 0.001을 1e-3으로도 표기

Wk7-3 : 분산분석 (ANOVA)
- Analysis of Variance -

1. 분산분석의 개념

• **ANOVA (Analysis of Variance)** : 전체분산(variance)을 분할(분석, analysis)하여 어떤 요인(factor)의 영향이 유의한지(significant)한지 검정하는 방법.

(예) Drug effect (5mg, 10mg, placebo)
Age effect(young, old)



전체변동(전체분산)= $\sum_{i=1}^{a}\sum_{j=1}^{b}(Y_{ij}-\bar{Y}_{..})^2$

factor가 1개일때의 분산분석모형

$$y_{ij}=\mu+\tau_i+\varepsilon_{ij},\begin{cases} i=1,2,...,a \\ j=1,2,...,n \end{cases}$$

$\mu=$ an overall mean, $\tau_i = ith$ treatment effect,
$\varepsilon_{ij}=$ experimental error, $N(0,\sigma^2)$

2. 분산분석 : factor가 한개일때

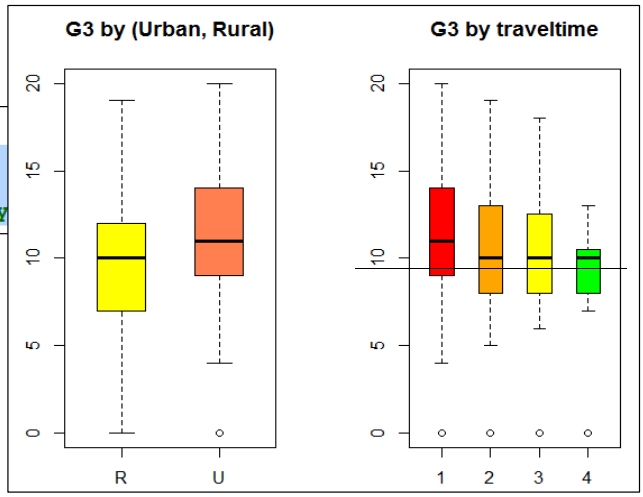
• 분산분석모형 적용

(1) 거주지역에 따른 학업성취도 : 거주지역(factor: R/U), 학업성적(1-20)

(2) 통학시간에 따른 학업성취도 : 통학시간(factor: 1-4), 학업성적(1-20)

Week6_3에서의 그래프를 이용한 데이터탐색

```
# 2. boxplot
par(mfrow=c(1,2))
boxplot(G3~address, boxwex = 0.5, col = c("yellow", "coral"),
boxplot(G3~traveltime, boxwex = 0.5, col = c("red","orange","y
```



(1) 도심지역 학생들 성적이 외곽지역 학생들보다 높다

(2) 통학시간이 짧은(15분이내)의 학생들의 성적이 더 높다

2. 분산분석 : factor가 한개일때

(1) 거주지역에 따른 학업성취도 : 거주지역(factor: R/U), 학업성적(1-20)

가설1 : 거주지역(R/U)에 따라 G3에 유의한 영향이 있나?

aov(타겟변수~factor)

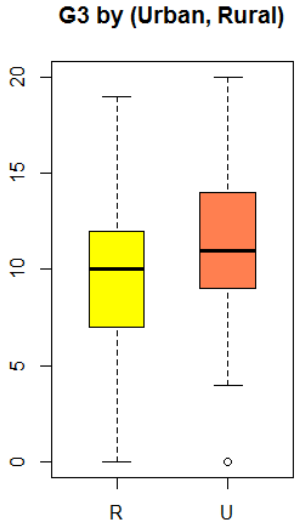
```
# ANOVA
a1 <- aov(G3~address)
summary(a1)
```

```
> a1 <- aov(G3~address)
> summary(a1)
             Df Sum Sq Mean Sq F value Pr(>F)
address       1     92   92.49   4.445 0.0356 *
Residuals   393   8177   20.81
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

p-value=0.035. 유의수준을 0.05로 잡을때 0.05보다 작으므로=>
거주지역에 따른 학업성적에는 유의한 차이가 있다고 할수 있음

2. 분산분석 : factor가 한개일때

(1) 거주지역에 따른 학업성취도 : 거주지역(factor: R/U), 학업성적(1-20)



분산분석 결과는 상자그림으로 본 거주지역에 따른 G3의 차이가 통계적으로 유의하다는 것을 보여줌!!

```
# tapply - give FUN value by address
round(tapply(G3, address, mean),2)
```

```
> round(tapply(G3, address, mean),2)
    R     U
 9.51 10.67
```

G3(Rural)=9.51, G3(Urban)=10.67

2. 분산분석 : factor가 한개일때

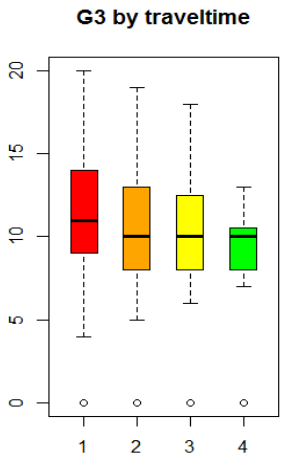
(2) 통학시간에 따른 학업성취도 : 통학시간(factor: 1-4), 학업성적(1-20)

가설2 : 통학시간에 따라 G3에는 유의한 차이가 있나?

aov(타겟변수~factor)

```
# 2. ANOVA by traveltime
traveltime<-as.factor(traveltime)
a2 <- aov(G3~traveltime)
summary(a2)
```

```
> # 2. ANOVA by traveltime
> a2 <- aov(G3~traveltime)
> summary(a2)
              Df Sum Sq Mean Sq F value Pr(>F)
traveltime     3    115   38.37    1.84  0.139
Residuals    391   8155   20.86
```



- p-value=0.139, 유의수준을 0.05로 잡을때 0.05보다 크므로=> 유의수준 0.05에서는 통학시간에 따른 학업성적에는 유의한 차이 없다고 할수 있다
- 그러나 p-value가 0.139이므로 어느정도 차이가 존재함을 알수 있다.

3. 사후검정(post-hoc analysis)

• 사후검정 : ANOVA에서 어떤 factor의 유의성이 검정되면, 그 다음단계에 하는 검정

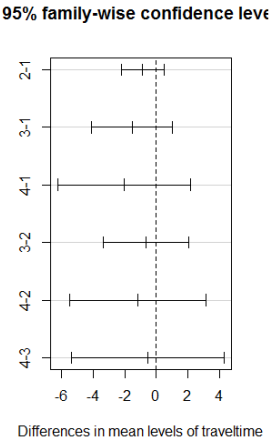
Tukey's Honest Significant Difference Test

```
# should be foctor for Tukey's Honest Sig
TukeyHSD(a2, "traveltime", ordered=TRUE)
plot(TukeyHSD(a2, "traveltime"))
```

```
> TukeyHSD(a2, "traveltime", ordered=TRUE)
  Tukey multiple comparisons of means
    95% family-wise confidence level
    factor levels have been ordered

Fit: aov(formula = G3 ~ traveltime)

$traveltime
         diff        lwr      upr     p adj
3-4 0.5108696 -4.3256048 5.347344 0.9929165
2-4 1.1565421 -3.1623166 5.475401 0.9005404
1-4 2.0321012 -2.1981712 6.262374 0.6021367
2-3 0.6456725 -2.0624782 3.353823 0.9272302
1-3 1.5212316 -1.0432848 4.085748 0.4202138
1-2 0.8755591 -0.4800959 2.231214 0.3429618
```

95% 신뢰구간의 lower bound, upper bound

• μ_1-μ_2 =0이라는 의미는 1그룹과 2그룹간 차이가 없다는 의미
• 즉 (μ_1-μ_2)의 신뢰구간에 0이 있다는것은 차이가 없다고 할 수있다

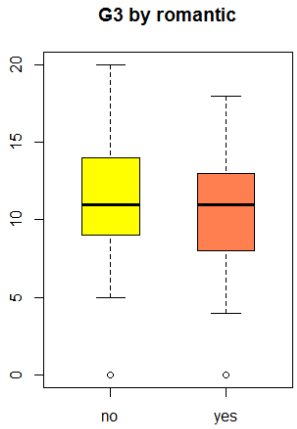


모든 pairwise 신뢰구간에 0이 포함됨
=> 유의한 차이가 없음

4. 추가예제: 분산분석

(4) 연애경험여부에 따른 학업성취도 : 연애경험(yes, no), 학업성적(1-20)

```
# 4. ANOVA by romantic
a4 <- aov(G3~romantic)
summary(a4)
# tapply - give FUN value by address
round(tapply(G3,romantic, mean),2)
```




```
> summary(a4)
             Df Sum Sq Mean Sq F value  Pr(>F)
romantic      1    140  139.70   6.753 0.00971 **
Residuals   393   8130   20.69
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
> # tapply - give FUN value by address
> round(tapply(G3,romantic, mean),2)
   no   yes
10.84  9.58
```

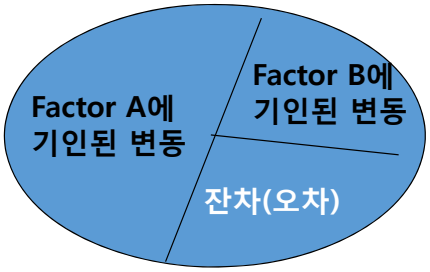
• 연애경험이 있는 경우 학업성적이 유의하게 낮음 (p-value=0.0097)
• median은 비슷해보이지만, 평균은 10.84-9.58=1.26 차이있음

Wk7-4 : 이원분산분석 (two-way ANOVA)

이원분산분석 (two-way ANOVA)

- **ANOVA (Analysis of Variance)** : 전체분산(variance)을 분할(분석, analysis)하여 어떤 요인(factor)의 영향이 유의한지(significant)한지 검정하는 방법.

(예) Drug effect (5mg, 10mg, placebo)
Age effect(young, old)



전체변동(전체분산)= $\sum_{i=1}^{a}\sum_{j=1}^{b}(Y_{ij}-\bar{Y}_{..})^2$

이원분산분석 (two-way ANOVA)

- 데이터 : High-Density Lipoprotein (HDL) 콜레스테롤

chol_ex.csv

1. ID
2. drug : 5mg, 10mg, placebo
3. age : young(18-39), old (≥40세)
4. value : HDL(투약전)-HDL(투약후)

id	drug	age	value
1	placebo	young	4
2	placebo	young	3
3	placebo	young	-1
4	placebo	old	3
5	placebo	old	1
6	placebo	old	2
7	5mg	young	7
8	5mg	young	5
9	5mg	young	6
10	5mg	old	5
11	5mg	old	7

*HDL(고밀도 리포 단백질)은 높을수록 좋은것으로 알려진 콜레스테롤. 40mg/dl이상이 정상범위

이원분산분석 (two-way ANOVA)

- 이원분산분석 (two-way ANOVA) : factor가 두개인 경우

(1) 투약효과가 있는가? (5mg, 10mg, 위약)

(2) 연령그룹(young/ old)에 따른 영향이 있는가?

가설 1 : 신약의 투약효과가 있는가? HDL을 상승시키는 효과가 있나?

가설 2 : 연령그룹에 따라 투약효과(HDL변화)에 차이가 있나?

가설 3 : 신약의 투약과 연령그룹간 상호작용 효과가 있는가?

이원분산분석 (two-way ANOVA)

- 이원분산분석 : aov(타겟변수~factor1+ factor2)

lec7_3.R

```
dat<-read.csv(file="chol_ex.csv")
head(dat)
dim(dat)
str(dat)
attach(dat)

# two-way ANOVA,
a6 <- aov(value ~ drug + age)
summary(a6)
```

```
> a6 <- aov(value ~ drug + age)
> summary(a6)
             Df Sum Sq Mean Sq F value   Pr(>F)
drug          2 161.44   80.72  33.568 4.55e-06 ***
age           1   4.50    4.50   1.871    0.193
Residuals    14  33.67    2.40
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
```

(1) drug effect : p-value~0이므로 HDL값에 통계적으로 유의한 차이가 있음

(2) age : p-value=0.19로 유의수준 0.05에서 유의한 차이는 없음

이원분산분석 (two-way ANOVA)

- 이원분산분석 : aov(타겟변수~factor1+ factor2+ 상호작용)

두개의 factor간 상호작용의 유의성을 검정하기 위한 분석

```
# two-way ANOVA with interaction
a7 <- aov(value ~ drug + age+ drug*age)
summary(a7)
```

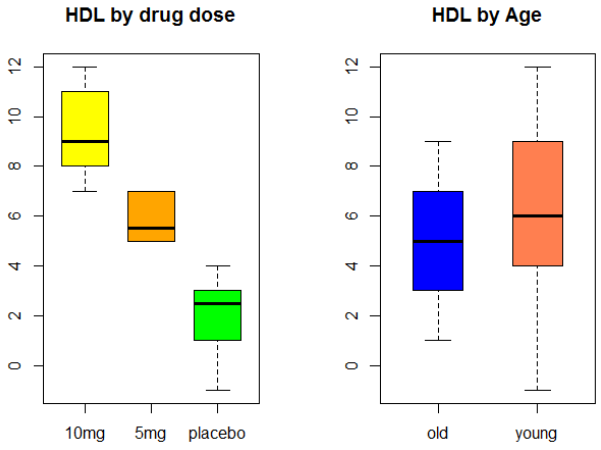
```
> a7 <- aov(value ~ drug + age+ drug*age)
> summary(a7)
             Df Sum Sq Mean Sq F value   Pr(>F)
drug          2 161.44   80.72  35.439 9.21e-06 ***
age           1   4.50    4.50   1.976    0.185
drug:age      2   6.33    3.17   1.390    0.286
Residuals    12  27.33    2.28
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.
```

(3) drug와 age그룹간 상호작용 : p-value=0.286으로 유의수준 0.05에서 유의한 차이는 없음

이원분산분석 (two-way ANOVA) : 두 요인의 상자그림

• 투약용량과 연령그룹에 따른 상자그림

```
# two-way ANOVA
par(mfrow=c(1,2))
boxplot(value ~ drug, boxwex = 0.7, main="HDL by drug dose", col = c("yellow","
boxplot(value ~ age, boxwex = 0.5, main="HDL by Age", col = c("blue", "coral"))
```



(1) drug effect : 10mg인 경우 HDL 상승효과가 가장 높음

```
# tapply - give FUN value by drug
round(tapply(value, drug, mean),2)
```

```
> round(tapply(value, drug, mean),2)
   10mg     5mg placebo
   9.33    5.83    2.00
```

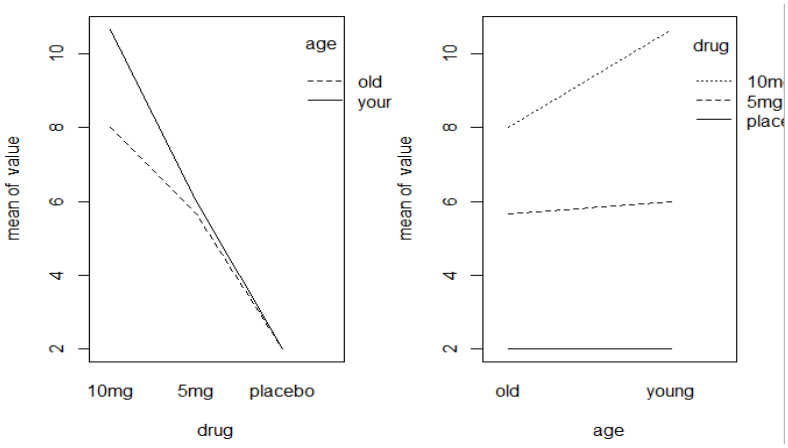
(2) age : young그룹(18-40)의 HDL 상승효과가 더 높음

```
> round(tapply(value, age, mean),2)
  old young
 5.22  6.22
```


이원분산분석 (two-way ANOVA) : 상호작용 그래프

• 상호작용 그래프

```
# interaction plot
par(mfrow=c(1,2))
interaction.plot(drug, age, value)
interaction.plot(age, drug, value)
```



(1) 투약용량을 기준(x축)으로 그릴때 (2) 연령그룹을 기준(x축)으로 그릴때

- 투약용량 10mg에서 young그룹의 상승효과가 old그룹보다 훨씬 높음
- 5mg에서와 placebo에서는 연령그룹의 차이가 거의없음

	단위별 학습내용 (Week8)
wk8-1	상관분석
wk8-2	회귀분석(선형모형)
wk8-3	텍스트마이닝 I
wk8-4	텍스트마이닝 II

Wk8-1 : 상관분석

1. 상관분석 : 상관계수

• 상관계수 : cor(변수1, 변수2)

```
# lec7_4.r
# Correlation coefficent

# set working directory
setwd("D:/tempstore/moocr")

# autompg data
car<-read.csv("autompg.csv")
#head(car)
#dim(car)

# subset of car : cyl (4,6,8)
car1<-subset(car, cyl==4 | cyl==6 | cyl==8)
attach(car1)
```

wt와 mpg는 음의 상관관계

```
#correlation
cor(wt, mpg)
cor(disp, mpg)
cor(accler, mpg)
```

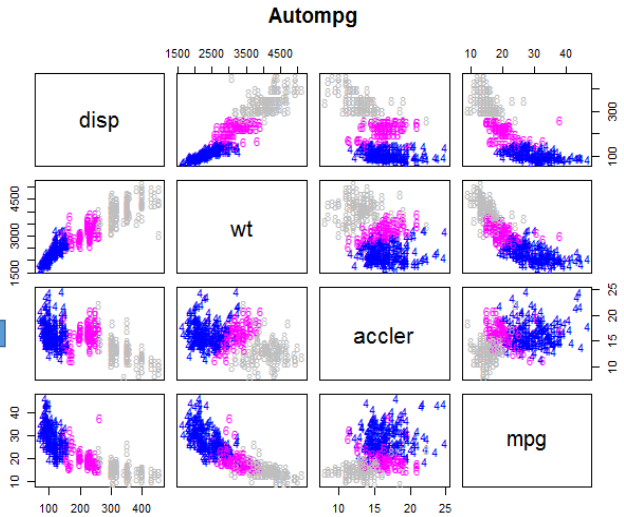
```
> cor(wt, mpg)
[1] -0.8420347
> cor(disp, mpg)
[1] -0.8168117
> cor(accler, mpg)
[1] 0.4163202
```

cor의 디폴트는 pearson의 상관계수
kendall의 상관계수 혹은 spearman의 상관계수를 구할때는
cor(변수1, 변수2, method=c("spearman"))

1. 상관분석 : 상관계수

• 상관계수와 산점도

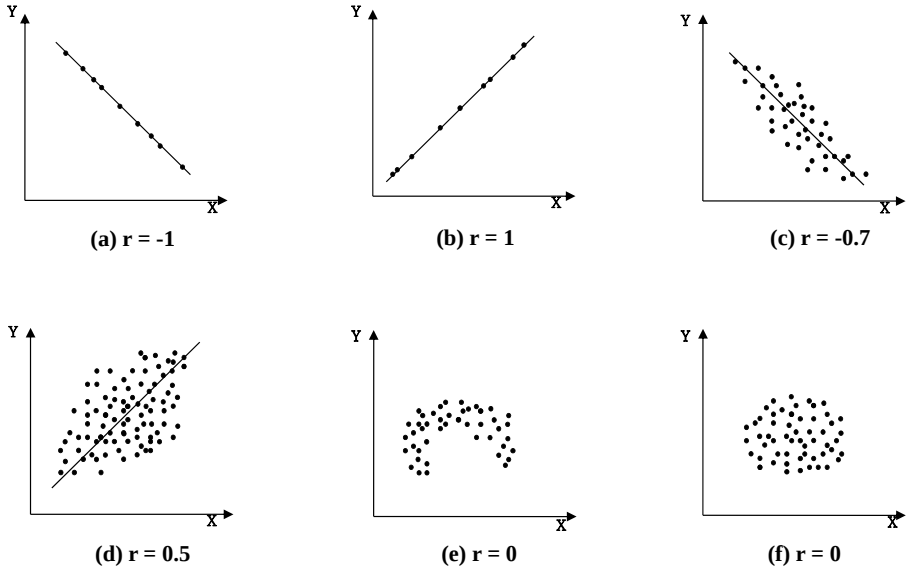
```
# 6. pariwise plot
# new variable lists
vars1<-c("disp", "wt", "accler", "mpg")
# pariwise plot
pairs(car1[vars1], main ="Autompg",cex=1, col=as.integer(car1$c
```



(1) 차량무게와 배기량과는 정비례관계 (양의 상관계수)
(2) MPG(연비)와 (wt, disp)는 상관성이 높다 (반비례 음의 상관계수)
(3) cylinder별로 색으로 표시 (파란색:4, 진한핑크: 6, 회색 : 8)

2. 상관분석 – 상관계수와 산점도

- 상관계수(r)은 절대값이 0-1사이 값을 갖는다



- 절대값이 0에 가까울수록 상관관계가 없다
- 절대값이 1에 가까운수록 강한 상관성이 있다

3. 통계치와 그래프 : 주의!!

- 통계치와 그래프 - Monkey 데이터 + King Kong 한마리

상관계수 : 0.53

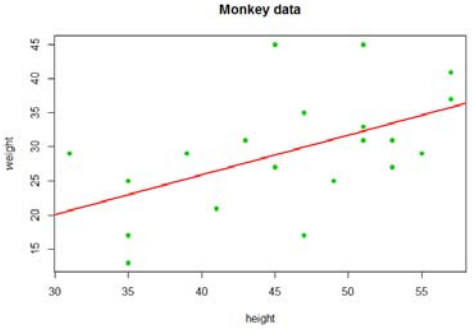
```
# correlation coefficients
cor(height, weight)
```

```
> cor(height, weight)
[1] 0.5267801
```

monkey.csv

ID	height	weight
1	55	29
2	45	27
3	35	17
4	39	29
5	53	31
6	41	21
7	51	31
8	35	13
9	57	37
10	57	41
11	45	45
12	47	35
13	35	25
14	49	25
15	43	31
16	51	33
17	31	29
18	53	27
19	47	17
20	51	45

```
# scatterplot for weight and height
par(mfrow=c(1, 1))
plot(height, weight, pch=16, col=3,main="Monkey data")

# add the best fit linear line (lec4_3.R)
abline(lm(weight~height), col="red", lwd=2, lty=1)
```



- weight와 height간 상관계수는 0.53으로 별로 높지 않다

3. 통계치와 그래프 : 주의!!

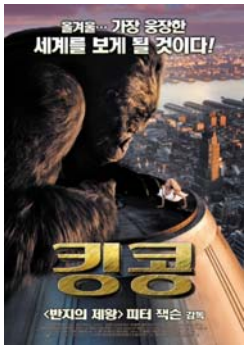
- 통계치와 그래프 - Monkey 데이터 + King Kong 한마리

```
## Monkey data + Kingkong
monkey1<-read.csv("monkey_k.csv")
head(monkey1)
dim(monkey1)
attach(monkey1)
```

```
# correlation coefficients
cor(height, weight)
```

```
> cor(height, weight)
[1] 0.940375
```

상관계수 : 0.94



ID	height	weight
1	55	29
2	45	27
3	35	17
4	39	29
5	53	31
6	41	21
7	51	31
8	35	13
9	57	37
10	57	41
11	45	45
12	47	35
13	35	25
14	49	25
15	43	31
16	51	33
17	31	29
18	53	27
19	47	17
20	51	45
21	130	150

3. 통계치와 그래프 : 주의!!

- 선형회귀식 - Monkey 데이터

```
# linear model and summary of linear model
m1<-lm(weight~height)
summary(m1)
```

```
> m1<-lm(weight~height)
> summary(m1)

Call:
lm(formula = weight ~ height)

Residuals:
     Min       1Q   Median       3Q      Max
-12.9797  -5.7186  -0.2983   3.9983  16.1797

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.7356    10.2815   0.266    0.793
height        0.5797     0.2205   2.629    0.017 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.573 on 18 degrees of freedom
Multiple R-squared:  0.2775,    Adjusted R-squared:  0.2374
F-statistic: 6.913 on 1 and 18 DF,  p-value: 0.01702
```

선형회귀식
Y(weight)=2.74+0.58X(height)

선형회귀식의 결정계수
R^2=0.27

3. 통계치와 그래프 : 주의!!

- 선형회귀식 - Monkey 데이터 + King Kong 한마리

```
# linear model and summary of linear model
m2<-lm(weight~height)
summary(m2)
```

```
> m2<-lm(weight~height)
> summary(m2)

Call:
lm(formula = weight ~ height)

Residuals:
    Min      1Q  Median      3Q     Max
-14.219  -7.298  -2.372   8.243  18.706

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -30.2495     5.8203  -5.197 5.13e-05 ***
height        1.3078     0.1085  12.051 2.41e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.646 on 19 degrees of freedom
Multiple R-squared:  0.8843,    Adjusted R-squared:  0.8782
F-statistic: 145.2 on 1 and 19 DF,  p-value: 2.412e-10
```

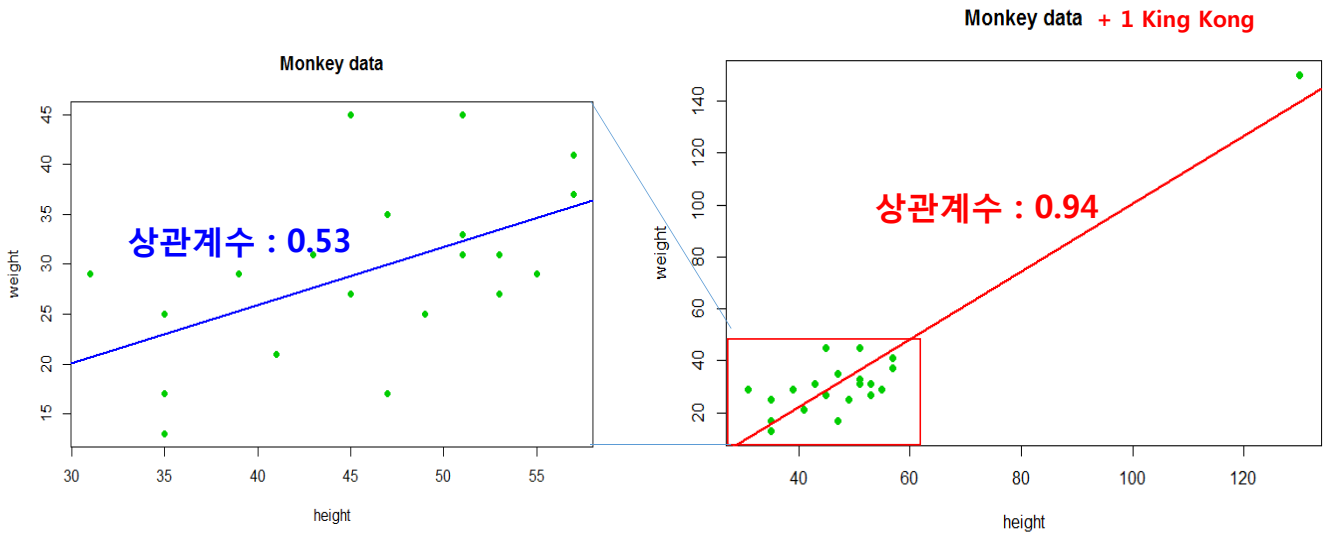
선형회귀식
Y(weight)=-30.25+1.31X(height)

선형회귀식의 결정계수
R^2=0.88

3. 통계치와 그래프 : 주의!!

- 통계치와 그래프 - Monkey 데이터 + King Kong 한마리

한마리의 킹콩 데이터가 몸무게와 신장의 상관관계에 대한 해석을 완전히 바꿔놓을수 있다 !!



Wk8-2 : 회귀분석 (선형모형)

1. 회귀분석 - 데이터

- **autompg 데이터**

1. mpg: continuous (연비 : 연속형변수)
2. cylinders: multi-valued discrete (실린더 : 정수값)
3. displacement: continuous (배기량 : 연속형변수)
4. horsepower: continuous (마력 : 연속형변수)
5. weight: continuous (무게 : 연속형변수)
6. acceleration: continuous (가속 : 연속형변수)
7. year: multi-valued discrete (모델연도 : 정수값)
8. origin: multi-valued discrete (정수값)
9. car name: string (unique for each instance) (차종류 이름)

```
# lec8_2.r : Linear model
# Regression

# set working directory
setwd("D:/tempstore/moocr")

# autompg data
car<-read.csv("autompg.csv")
head(car)
str(car)

# subset with cyl=4,6,8
car1<-subset(car, cyl==4 | cyl==6 | cyl==8)
attach(car1)
```

```
> car<-read.csv("autompg.csv")
> head(car)
  mpg cyl disp hp   wt accler year origin                  carname
1  18   8  307 17 3504   12.0   70      1 chevrolet chevelle malibu
2  15   8  350 35 3693   11.5   70      1         buick skylark 320
3  18   8  318 29 3436   11.0   70      1        plymouth satellite
4  16   8  304 29 3433   12.0   70      1             amc rebel sst
5  17   8  302 24 3449   10.5   70      1               ford torino
6  15   8  429 42 4341   10.0   70      1          ford galaxie 500
```

2. 회귀분석 - 단순회귀모형

• 단순회귀모형 : lm(y변수~x변수, data=)

```
# 1. simple Regression(independent variable : wt)
r1<-lm(mpg~wt, data=car1)
summary(r1)
anova(r1)
```

1. 단순회귀모형
종속변수 : mpg(연비), 독립변수: wt(차량무게)

```
> r1<-lm(mpg~wt, data=car1)
> summary(r1)

Call:
lm(formula = mpg ~ wt, data = car1)

Residuals:
    Min      1Q  Median      3Q     Max
-9.6770 -2.7567 -0.3636  2.1120 16.3712

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 46.600189   0.779849   59.76   <2e-16 ***
wt          -0.007759   0.000252  -30.79   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.239 on 389 degrees of freedom
Multiple R-squared:  0.709,     Adjusted R-squared:  0.7083
F-statistic: 947.9 on 1 and 389 DF,  p-value: < 2.2e-16
```

선형회귀식
y(mpg)=46.60-0.0077(wt)

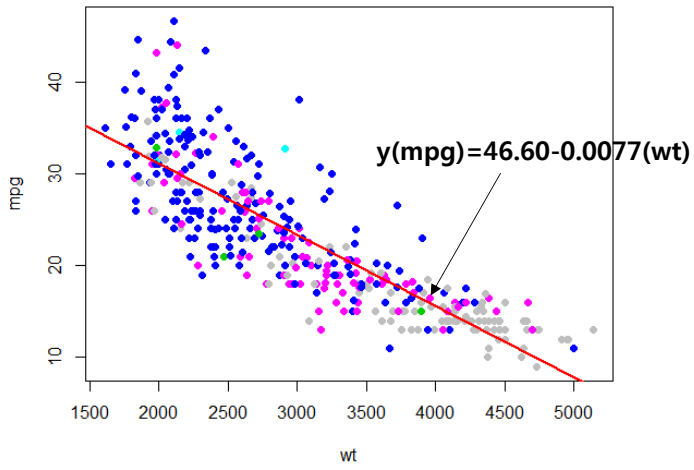
선형회귀식의 결정계수
R^2=0.71

2. 회귀분석 - 단순회귀모형

• 산점도에 회귀선 그리기

```
# (lec4_3.R) scatterplot with best fit lines
par(mfrow=c(1,1))
plot(wt, mpg,  col=as.integer(car1$cyl), pch=19)
# best fit linear line
abline(lm(mpg~wt), col="red", lwd=2, lty=1)
```

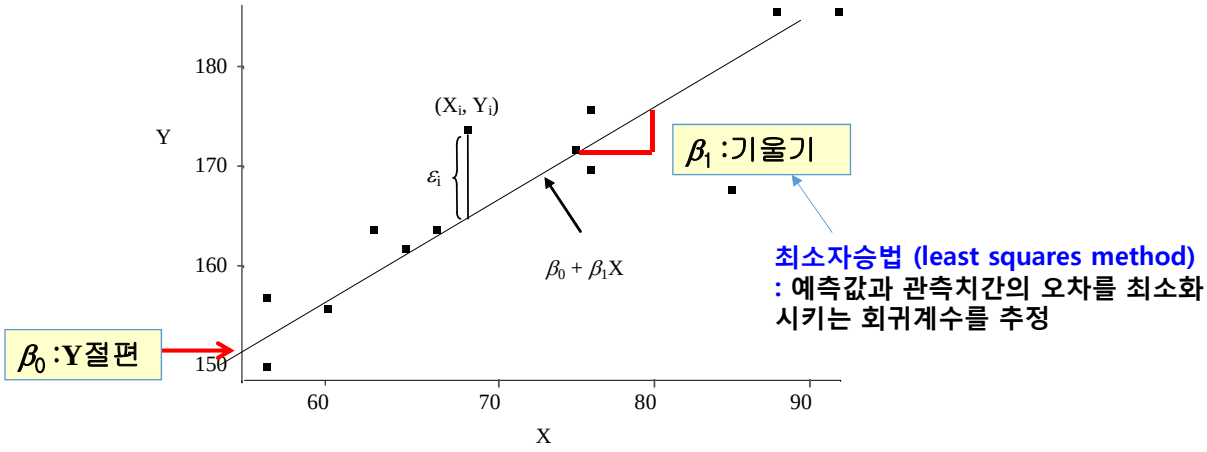
plot(x축변수, y축변수)
abline : add line (선을 추가하는 함수)
lm(y변수~x변수) : lm은 linear model(선형모형)의 약자



3. 회귀분석의 목적

- 회귀분석의 목적 : 예측(prediction)과 추정(estimation)
  - 선형모형 : 독립변수와 종속변수간의 관계가 선형식으로 적합

모형 : $Y_i=\beta_0 + \beta_1 X_i + \varepsilon_i, \quad i=1, 2, \ldots, n$



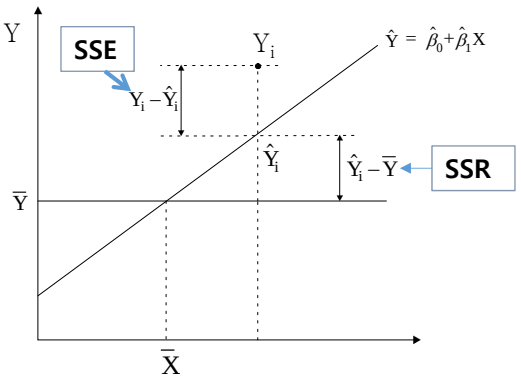
최소자승법 (least squares method)
: 예측값과 관측치간의 오차를 최소화 시키는 회귀계수를 추정

4. 회귀분석 - 모형의 적합도

- 모형의 적합도와 결정계수 (R^2) : $0 \le R^2 \le 1$

: 전체제곱합(SST)에 대한 회귀제곱합(SSR)의 비율, 즉 모형으로 설명할수 있는 부분의 비율

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$



전체제곱합의 분할: SST=SSR+SSE

전체제곱합(SST): $SST = \sum(Y_i - \bar{Y})^2$

회귀제곱합(SSR): $SSR = \sum(\hat{Y}_i - \bar{Y})^2$

잔차제곱합(SSE): $SSE = \sum(Y_i - \hat{Y}_i)^2$

4. 회귀분석 - 모형의 적합도

- 회귀식에 의해 설명되는 부분(SSR)과 설명되지 않는부분(SSE)

```
> anova(r1)
Analysis of Variance Table

Response: mpg
           Df  Sum Sq Mean Sq F value    Pr(>F)
wt          1 17029.1   17029  947.87 < 2.2e-16 ***
Residuals 389  6988.6      18
```

$$R^2 = \frac{SSR}{SST} = \frac{17029}{24017} = 0.709$$

- R^2는 1에 가까울수록 회귀식에 의해 적합되는 부분이 높음
- R^2는 0에 가까우면 주어진 독립변수들에 의해 설명(예측 혹은 적합)되는 부분이 없다고 할 수 있다

SST=total sum of squares
SSR=regression sum of squares
SSE=error(residual) sum of squares

4. 회귀분석

```
# 2. simple Regression(independent variable : disp)
r2<-lm(mpg~disp, data=car1)
summary(r2)
anova(r2)
```

#2. 단순회귀모형
종속변수 : mpg(연비), 독립변수: disp(배기량)

```
> summary(r2)

Call:
lm(formula = mpg ~ disp, data = car1)

Residuals:
     Min       1Q   Median       3Q      Max
-10.0627  -3.0037  -0.6113   2.3110  18.5978

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 35.49479    0.48640   72.97   <2e-16 ***
disp        -0.06142    0.00220  -27.93   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

Residual standard error: 4.533 on 389 degrees of freedom
Multiple R-squared:  0.6672,    Adjusted R-squared:  0.6663
F-statistic: 779.8 on 1 and 389 DF,  p-value: < 2.2e-16
```

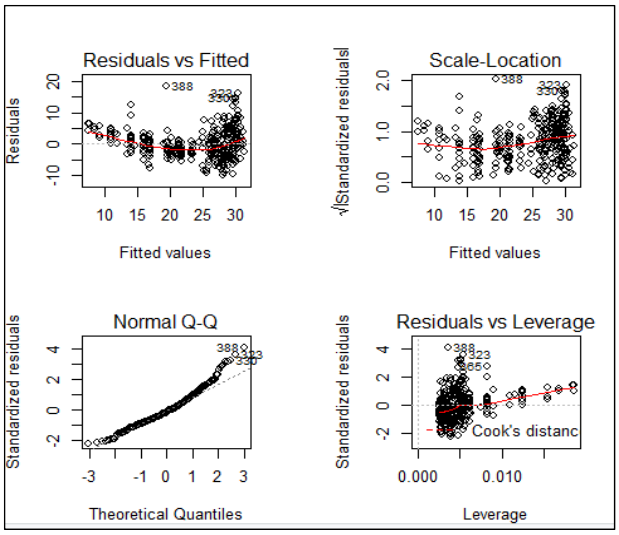
선형회귀식
y(mpg)=35.49-0.0614(disp)

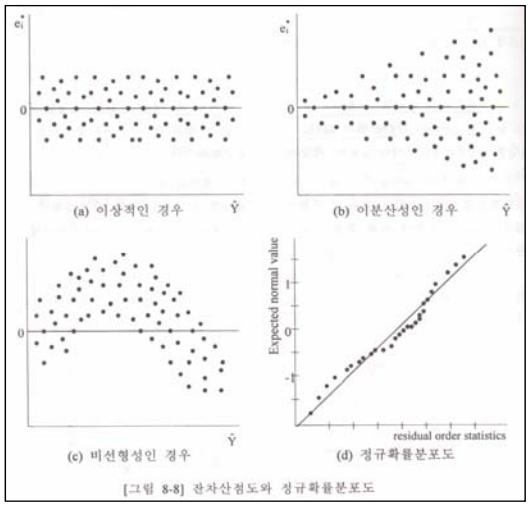
선형회귀식의 결정계수
R^2=0.67

5. 회귀분석 – 잔차의 산점도

• 회귀분석의 가정과 진단

```
# residual diagnostic plot
layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
plot(r2)
```





[그림 8-8] 잔차산점도와 정규확률분포도

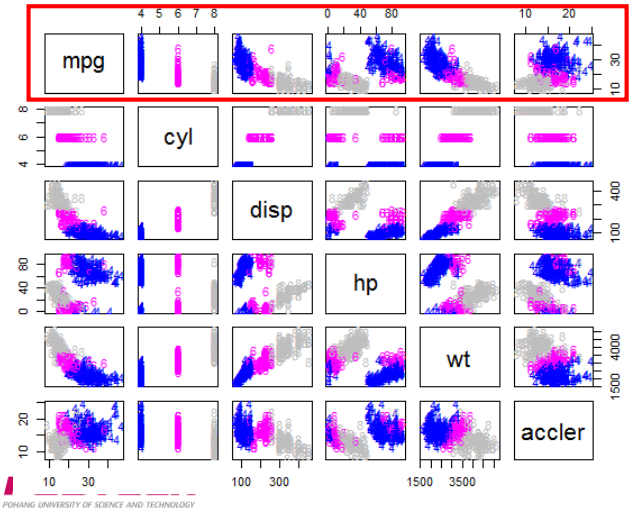
참고 : 전치혁, 정민근, 이혜선, 공학응용통계, 홍릉출판사, 2012

6. 회귀분석

• 다중회귀분석 : 독립변수들이 여러 개인 경우

```
# 3. multiple Regression
r3<-lm(mpg~wt+accler, data=car1)
summary(r3)
anova(r3)
```

• pairwise scatterplot




```
# pariwise plot
# new variable lists
vars1<-c("disp", "wt", "accler", "mpg")
# pariwise plot
pairs(car[vars1], main ="Autompg",cex=1,
```

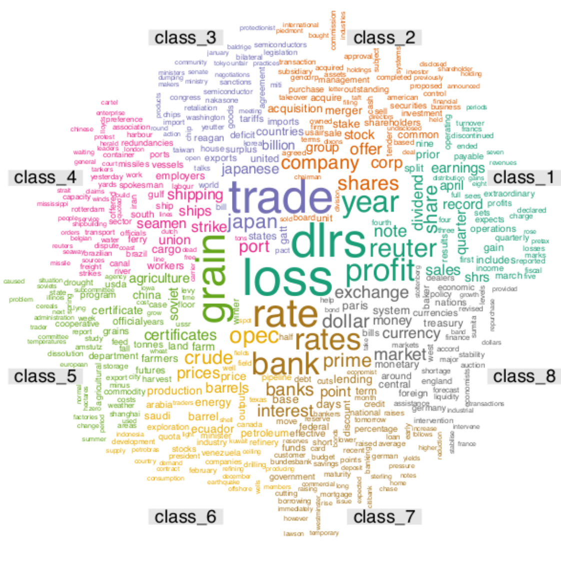
Wk8-3 : 텍스트마이닝 I

1. 텍스트마이닝이란?

텍스트마이닝 (Text mining)

- 텍스트마이닝은 웹페이지, 이메일, 소셜네트워크 기록, 등 전자문서 파일로부터 특정 연관성(동시적으로 빈도가 높은 단어추출)을 분석하는 방법

- 텍스트마이닝은 다양한방식의 알고리즘을 이용하여 대용량의 텍스트문서로부터 트렌드와 관심어를 찾아내는 기법으로 사용



2. 텍스트마이닝에 필요한 패키지

텍스트마이닝을 위한 패키지

```
# 자연어처리
install.packages("NLP")
library(NLP)

# 텍스트마이닝 패키지
install.packages("tm")
library(tm)

# 텍스트마이닝 결과의 시각화
install.packages("wordcloud")
library(wordcloud)
```

그 외 패키지

```
# 한글처리를 위한 패키지
install.packages("KoNLP")
library(KoNLP)

# 트위터의 데이터를 불러오는 패키지
install.packages("twitteR")
library(twitteR)

# color displaying
install.packages('RColorBrewer')
library(RColorBrewer)
```


2. 텍스트마이닝에 필요한 패키지

- 패키지설치와 라이브러리 설정, tm의 예제 데이터 (crude)

```
# lec8_3.R
# Install package NLP, tm, wordcloud

# set working directory
setwd("D:/tempstore/moocr")

# Natural language processing
install.packages('NLP')
# text mining package
install.packages('tm')
# visualizing
install.packages('wordcloud')
# color displaying
install.packages('RColorBrewer')

# set library (set in order)
library(NLP)
library(tm)
library(RColorBrewer)
library(wordcloud)

# 20 new articles from Reuter- 21578 data set
data(crude)
# To know about crude data
help(crude)
```

텍스트마이닝을 위한 추가패키지 설치

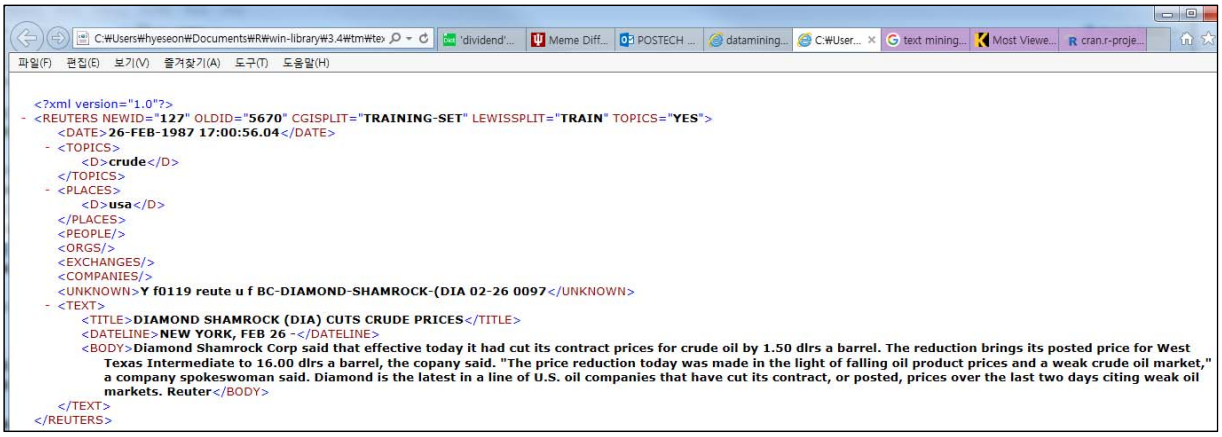
패키지의 라이브러리 설정

tm에 포함된 데이터 crude사용

3. 텍스트마이닝 패키지(tm)의 데이터

- 예제데이터 : crude* 데이터 (로이터통신의 20개의 뉴스기사)

crude[[1]] : 첫번째 기사 (아래와 같은 XML파일 형태로 저장)- tm패키지에 들어있는 예제데이터



```
<?xml version="1.0"?>
- <REUTERS NEWID="127" OLDID="5670" CGISPLIT="TRAINING-SET" LEWISSPLIT="TRAIN" TOPICS="YES">
    <DATE>26-FEB-1987 17:00:56.04</DATE>
  - <TOPICS>
      <D>crude</D>
    </TOPICS>
  - <PLACES>
      <D>usa</D>
    </PLACES>
    <PEOPLE/>
    <ORGS/>
    <EXCHANGES/>
    <COMPANIES/>
    <UNKNOWN>Y f0119 reute u f BC-DIAMOND-SHAMROCK-(DIA 02-26 0097</UNKNOWN>
  - <TEXT>
      <TITLE>DIAMOND SHAMROCK (DIA) CUTS CRUDE PRICES</TITLE>
      <DATELINE>NEW YORK, FEB 26 -</DATELINE>
      <BODY>Diamond Shamrock Corp said that effective today it had cut its contract prices for crude oil by 1.50 dlrs a barrel. The reduction brings its posted price for West
        Texas Intermediate to 16.00 dlrs a barrel, the copany said. "The price reduction today was made in the light of falling oil product prices and a weak crude oil market,"
        a company spokeswoman said. Diamond is the latest in a line of U.S. oil companies that have cut its contract, or posted, prices over the last two days citing weak oil
        markets. Reuter</BODY>
    </TEXT>
</REUTERS>
```

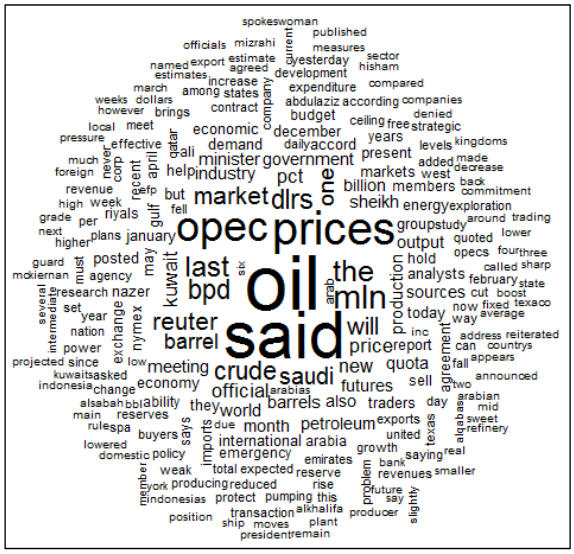
*Reuters-21578 text categorization test collection Distribution 1.0 README file (v 1.2), 1987년의 뉴스기사

4. 텍스트마이닝 예제 : Crude(원유)데이터

lec8_3.R

```
# set library (set in order)
library(NLP)
library(tm)
library(RColorBrewer)
library(wordcloud)

# 20 new articles from Reuter- 21578 data set
data(crude)
# To know about crude data
help(crude)

# remove punctuation in documnet
crude<-tm_map(crude, removePunctuation)
# remove stopwords
crude<-tm_map(crude, removeNumbers)
crude<-tm_map(crude, function(x) removeWords(x,stopwords()))
stopwords()

# from frequency counts
tdm<-TermDocumentMatrix(crude)
m<-as.matrix(tdm)
v<-sort(rowSums(m), decreasing=TRUE)
d<-data.frame(word=names(v), freq=v)

# plot a word cloud
# to see detail help(wordcloud)
par(mfrow=c(1, 1))
wordcloud(d$word, d$freq, random.order=FALSE)
```

텍스트마이닝 결과 : oil, said, price



Ref: https://www.r-bloggers.com/text-mining/

4. 텍스트마이닝 예제

• help(wordcloud)

Plot a word cloud

Usage

```
wordcloud(words,freq,scale=c(4,.5),min.freq=3,max.words=Inf,
        random.order=TRUE, random.color=FALSE, rot.per=.1,
        colors="black",ordered.colors=FALSE,use.r.layout=FALSE,
        fixed.asp=TRUE, ...)
```

Arguments

words	the words
freq	their frequencies
scale	A vector of length 2 indicating the range of the size of the words.
min.freq	words with frequency below min.freq will not be plotted
max.words	Maximum number of words to be plotted. least frequent terms dropped
random.order	plot words in random order. If false, they will be plotted in decreasing frequency
random.color	choose colors randomly from the colors. If false, the color is chosen based on the frequency
rot.per	proportion words with 90 degree rotation
colors	color words from least to most frequent
ordered.colors	if true, then colors are assigned to words in order

5. 텍스트마이닝 : 코퍼스(corpus)

• 코퍼스(corpus, 말뭉치) : 텍스트집합을 의미

(예) 신문기사(html, text), SNS (tweeter, facebook)

• 코퍼스 기반 언어연구*, 빅데이터 시대의 언어분석+

• Source의 종류

DirSource() : 디렉토리
DataframeSource() : R 데이터프레임
VectorSource() : R 벡터
XMLSource() : XML 파일
URISource() : URI

• 문서포맷과 Reader

readPlain : Plain Text
readPDF: pdf 파일
readDOC: MS word 문서
readXML: XML문서

*R을 활용한 코퍼스언어학과 통계학, 이용훈, 한국문화사, 2016
+빅데이터 시대의 언어연구, 이민행, 2015

Wk8-4 : 텍스트마이닝 II

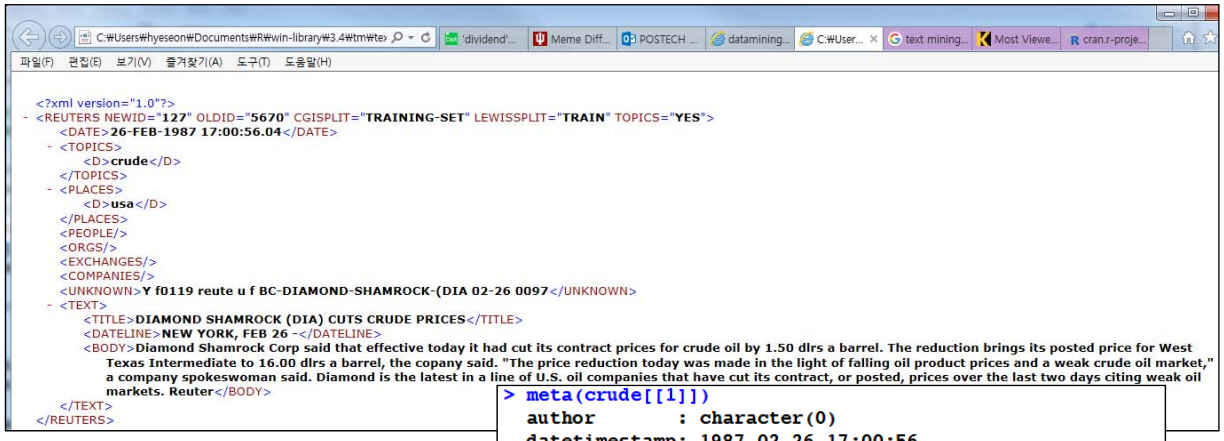
6. 텍스트마이닝 : 함수

- 텍스트마이닝에서 사용되는 함수

함수	설명과 예제코드
str(x[[1]])	데이터의 구조에 대한 정보 (첫번째 파일의 구조) str(crude[[1]])
content(x[[1]])	문서의 내용 (첫번째 문서의 내용)
meta(x)	메타정보 (x에 기록되어있는 저자, 날자, id, 등 정보를 보여줌) meta(crude[[1]])
inspect(x)	코퍼스, 텍스트, 문서행렬 등에 대한 정보를 제공 data(crude) inspect(crude[1:3]) inspect(crude[[1]]) tdm <- TermDocumentMatrix(crude) inspect(tdm)
lapply(x, content)	파일의 내용을 보여줌 lapply(crude, content)

6. 텍스트마이닝 : 함수

crude[[1]] 데이터 : 첫번째 기사 (아래와 같은 XML파일 형태로 저장)



```
<?xml version="1.0"?>
- <REUTERS NEWID="127" OLDID="5670" CGISPLIT="TRAINING-SET" LEWISSPLIT="TRAIN" TOPICS="YES">
    <DATE>26-FEB-1987 17:00:56.04</DATE>
  - <TOPICS>
      <D>crude</D>
    </TOPICS>
  - <PLACES>
      <D>usa</D>
    </PLACES>
    <PEOPLE/>
    <ORGS/>
    <EXCHANGES/>
    <COMPANIES/>
    <UNKNOWN>Y f0119 reute u f BC-DIAMOND-SHAMROCK-(DIA 02-26 0097</UNKNOWN>
  - <TEXT>
      <TITLE>DIAMOND SHAMROCK (DIA) CUTS CRUDE PRICES</TITLE>
      <DATELINE>NEW YORK, FEB 26 -</DATELINE>
      <BODY>Diamond Shamrock Corp said that effective today it had cut its contract prices for crude oil by 1.50 dlrs a barrel. The reduction brings its posted price for West
        Texas Intermediate to 16.00 dlrs a barrel, the copany said. "The price reduction today was made in the light of falling oil product prices and a weak crude oil market,"
        a company spokeswoman said. Diamond is the latest in a line of U.S. oil companies that have cut its contract, or posted, prices over the last two days citing weak oil
        markets. Reuter</BODY>
    </TEXT>
  </REUTERS>
```

```
# information about the
str(crude[[1]])
content(crude[[1]])
meta(crude[[1]])
lapply(crude, content)
```

```
> meta(crude[[1]])
  author       : character(0)
  datetimestamp: 1987-02-26 17:00:56
  description  :
  heading      : DIAMOND SHAMROCK (DIA) CUTS CRUDE PRICES
  id           : 127
  language     : en
  origin       : Reuters-21578 XML
  topics       : YES
  lewissplit   : TRAIN
  cgisplit     : TRAINING-SET
  oldid        : 5670
  places       : usa
  people       : character(0)
  orgs         : character(0)
  exchanges    : character(0)
```


6. 텍스트마이닝 함수

- **inspect 함수**

```
# inspect function
inspect(crude[1:3])
```

각 파일에 char숫자

```
> inspect(crude[1:3])
<<VCorpus>>
Metadata:  corpus specif
Content:  documents: 3

$`reut-00001.xml`
<<PlainTextDocument>>
Metadata:  15
Content:  chars: 527

$`reut-00002.xml`
<<PlainTextDocument>>
Metadata:  15
Content:  chars: 2634

$`reut-00004.xml`
<<PlainTextDocument>>
Metadata:  15
Content:  chars: 330
```

```
# inspect function
inspect(crude[1:3])
inspect(crude[[1]])
```

첫번째 파일의 내용을 보여줌

```
> inspect(crude[[1]])
<<PlainTextDocument>>
Metadata:  15
Content:  chars: 527

Diamond Shamrock Corp said that
effective today it had cut its contract prices for crude oil by
1.50 dlrs a barrel.
    The reduction brings its posted price for West Texas
Intermediate to 16.00 dlrs a barrel, the copany said.
    "The price reduction today was made in the light of falling
oil product prices and a weak crude oil market," a company
spokeswoman said.
    Diamond is the latest in a line of U.S. oil companies that
have cut its contract, or posted, prices over the last two days
citing weak oil markets.
```

7. 텍스트마이닝 전처리 함수

- 텍스트 전처리

함수	설명
tm_map(x, removePunctuation) (예) crude<-tm_map(crude, removePunctuation)	문장부호 제거 (. , “” ‘ ’)
tm_map (x, stripWhitespace) (예) crude<-tm_map(crude,stripWhitespace)	공백문자 제거
tm_map (x, removeNumbers) (예) crude<-tm_map(crude, removeNumbers)	숫자 제거

7. 텍스트마이닝 전처리 함수

- 텍스트 전처리 : 문장부호 없애기- tm_map(x, removePunctuation)

```
str(crude[[1]])
content(crude[[1]])
```

```
> content(crude[[1]])
[1] "Diamond Shamrock Corp said that\neffective today it had cut its con
tract prices for crude oil by\n1.50 dlrs a barrel.\n    The reduction br
ings its posted price for West Texas\nIntermediate to 16.00 dlrs a barre
l, the copany said.\n    \"The price reduction today was made in the lig
ht of falling\noil product prices and a weak crude oil market,\" a compa
ny\nspokeswoman said.\n    Diamond is the latest in a line of U.S. oil c
ompanies that\nhave cut its contract, or posted, prices over the last tw
o days\nciting weak oil markets.\n Reuter"
```

```
# 1. remove punctuation in documnet
crude<-tm_map(crude, removePunctuation)
content(crude[[1]])
```

문장부호 제거 (. , “” ‘ ’)

```
> content(crude[[1]])
[1] "Diamond Shamrock Corp said that\neffective today it had cut its con
tract prices for crude oil by\n150 dlrs a barrel\n    The reduction brin
```

7. 텍스트마이닝 전처리 함수

- 텍스트 전처리 : 숫자 제거- tm_map(x, removeNumbers)

```
# 2. remove numbers
crude<-tm_map(crude, removeNumbers)
content(crude[[1]])
```

숫자 제거

```
> content(crude[[1]])
[1] "Diamond Shamrock Corp said that\neffective today it had cut its con
tract prices for crude oil by\n dlrs a barrel\n    The reduction brings
```

- 텍스트 전처리 : stopwords 제거 – 언어별로 다름, 지정할수 있음

```
# 3. remove stopwords
crude<-tm_map(crude, function(x) removeWords(x,stopwords()))
content(crude[[1]])
```

that, it, had, its, for, by, The 제거

```
> content(crude[[1]])
[1] "Diamond Shamrock Corp said \neffective today   cut  contract prices
  crude oil \n dlrs  barrel\n    The reduction brings  posted price  Wes
```


7. 텍스트마이닝 전처리 함수

- 텍스트전처리 : stopword 리스트

```
# 3. remove stopwords
crude<-tm_map(crude,
content(crude[[1]])

stopwords()
```

```
> stopwords()
  [1] "i"          "me"         "my"         "myself"     "we"
  [6] "our"        "ours"       "ourselves"  "you"        "your"
 [11] "yours"      "yourself"   "yourselves" "he"         "him"
 [16] "his"        "himself"    "she"        "her"        "hers"
 [21] "herself"    "it"         "its"        "itself"     "they"
 [26] "them"       "their"      "theirs"     "themselves" "what"
 [31] "which"      "who"        "whom"       "this"       "that"
 [36] "these"      "those"      "am"         "is"         "are"
 [41] "was"        "were"       "be"         "been"       "being"
 [46] "have"       "has"        "had"        "having"     "do"
 [51] "does"       "did"        "doing"      "would"      "should"
 [56] "could"      "ought"      "i'm"        "you're"     "he's"
```

stopwords("en") 174개
stopwords("SMART") 517개

8. 텍스트마이닝 수행

• 문서행렬을 구성 : TermDocumentMatrix(문서이름)

```
# 4. contsruct term-doucument matrix
tdm<-TermDocumentMatrix(crude)
inspect(tdm)
```

```
> inspect(tdm)
<<TermDocumentMatrix (terms: 1011, documents: 20)>>
Non-/sparse entries: 1770/18450
Sparsity           : 91%
Maximal term length: 16
Weighting          : term frequency (tf)
Sample             :
        Docs
Terms    144 236 237 242 246 248 273 489 502 704
  bpd      4   7   0   0   0   2   8   0   0   0
  crude    0   2   0   0   0   0   5   0   0   0
  dlrs     0   2   1   0   0   4   2   1   1   0
  last     1   4   3   0   2   1   7   0   0   0
  mln      4   4   1   0   0   3   9   3   3   0
  oil     12   7   3   3   5   9   5   4   5   3
  opec    13   6   1   2   1   6   5   0   0   0
  prices   5   5   1   2   1   9   5   2   2   3
  said    11  10   1   3   5   7   8   2   2   4
  the      2   0   1   1   3   1   4   1   2   4
```

crude 문서번호 144에 나오는 단어들의 빈도

```
$`144`
[1] "OPEC may  forced  meet  \nscheduled June session  readdress  production
cutting\nagreement  organization wants  halt  current slide\n oil prices oil
 industry analysts said\n    The movement  higher oil prices  never    easy\n
 OPEC thought They may need  emergency meeting  sort \n problems said Daniel
Yergin director  Cambridge Energy\nResearch Associates CERA\n    Analysts  oi
l industry sources said  problem OPEC\nfaces  excess oil supply  world oil ma
rkets\n    OPECs problem    price problem  production\nissue  must  addresse
d   way said Paul Mlotok oil\nanalyst  Salomon Brothers Inc\n    He said  mar
kets earlier optimism  OPEC  \nability  keep production  control  given way
\npessimistic outlook  organization must address soon \n wishes  regain  ini
tiative  oil prices\n    But  analysts  uncertain  even \nemergency meeting
 address  problem  OPEC production\n  158 mln bpd quota set last December\n
  OPEC  learn    buyers market   \ndeemed quotas fixed prices  set differen.
.. <truncated>
```


8. 텍스트마이닝 수행

• 문서행렬을 행렬로 변환

```
# 5. read tdm as a matrix
m<-as.matrix(tdm)
head(m)
```

```
> m<-as.matrix(tdm)
> head(m)
           Docs
Terms       127 144 191 194 211 236 237 242 246 248 273 349 352 353 368 489
  abdulaziz   0   0   0   0   0   0   0   0   5   0   0   0   0   0   0   0
  ability     0   2   0   0   0   3   0   0   0   0   1   0   0   0   0   0
  able        0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
  abroad      0   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0
  accept      0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
  accord      0   0   0   0   0   0   0   0   0   5   1   0   2   0   0   0
```

행렬로 변환된 m에는 962 개 단어의 빈도가 20개 파일에 대해 계산되어 있음

```
> dim(m)
[1] 962  20
```

• 단어의 빈도 순서로 정렬

```
# 6. sorting in high frequency to low
v<-sort(rowSums(m), decreasing=TRUE)
v[1:10]
```

각 단어에 대한 빈도의 합(행기준)
가장높은것부터 [1:10]번까지

```
> v[1:10]
  oil  said prices  opec   mln   the  last   bpd  dlrs crude
   85    73    48    42    31    26    24    23    23    21
```

8. 텍스트마이닝 수행

- 단어 이름과 빈도를 결합한 행렬을 데이터 프레임으로 저장
- crude관련기사 파일로부터 962개의 단어들을 분류하여 빈도를 계산

```
# 7. match with freq and word names
d<-data.frame(word=names(v), freq=v)
head(d)
d[957:962, ]
```

가장 빈도가 높은 단어 6개

```
> head(d)
          word freq
oil        oil   85
said      said   73
prices  prices   48
opec      opec   42
mln        mln   31
the        the   26
```

가장 빈도가 낮은 단어 6개

```
> d[957:962, ]
                word freq
whether      whether    1
wishes        wishes    1
worldwide  worldwide    1
xon              xon    1
yergin        yergin    1
yesterdays yesterdays    1
```

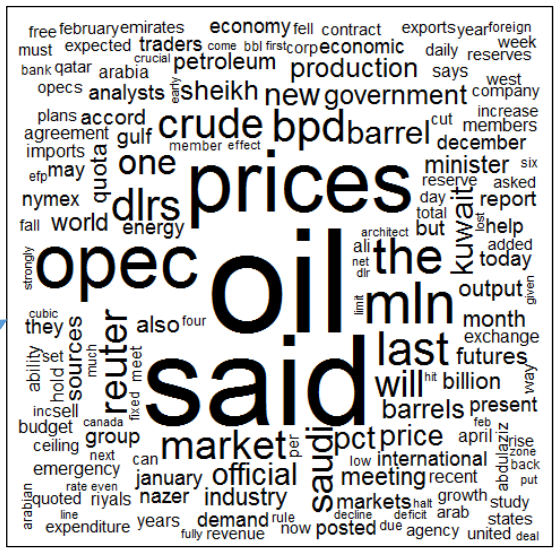

9. 텍스트마이닝 결과 그리기

- 빈도가 가장 높은 단어부터 그리기

```
# 7-1. Now lets try it with frequent words plotted first
wordcloud(d$word,d$freq,c(8,.5),2,,FALSE,.1)
```

가장 빈도가 높은 단어 6개

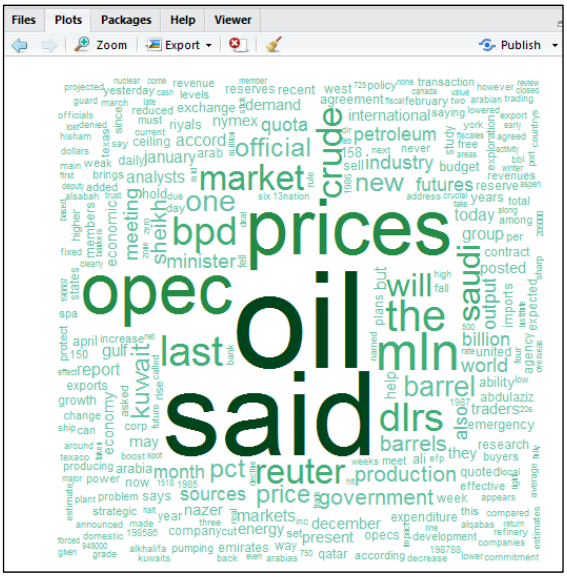
```
> head(d)
          word freq
oil        oil   85
said      said   73
prices  prices   48
opec      opec   42
mln        mln   31
the        the   26
```



9. 텍스트마이닝 결과 그리기

- 빈도가 가장 높은 단어부터 그리기 (색상 넣기)

```
# 7-2. color plot with frequent words plotted first
pal <- brewer.pal(9,"BuGn")
pal <- pal[-(1:4)]
wordcloud(d$word,d$freq,c(8,.3),2,,FALSE,,.15,pal)
```



아래의 패키지 필요

```
# color displaying
#install.packages('RColorBrewer')
library(RColorBrewer)
```


9. 텍스트마이닝 : 한글

- 한글문서에서 텍스트마이닝

```
# 9. Korean text #####
# takes quite long for installing
install.packages("KoNLP")
library(KoNLP)
# Korean dictionary
useSejongDic (backup=T)
```

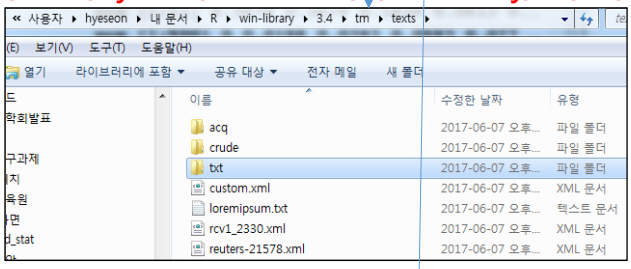
- 설치에 시간이 약간 걸림
- 단계별로 filtering하는 방법은 동일
- 한글문서로 텍스트마이닝

10. 텍스트마이닝 데이터 저장소 : 참고

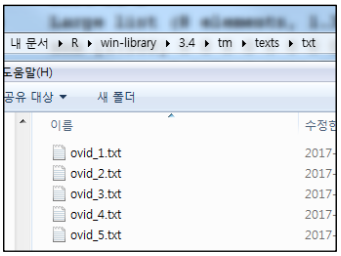
- tm패키지의 예제데이터와 저장소

```
tm.dir <- system.file ("texts", "txt", package = "tm")
tm.dir
#C:/Users/hyeseon/Documents/R/win-library/3.4/tm/texts/txt"
dir (tm.dir)
#[1] "ovid_1.txt" "ovid_2.txt" "ovid_3.txt" "ovid_4.txt" "ovid_5.txt
```

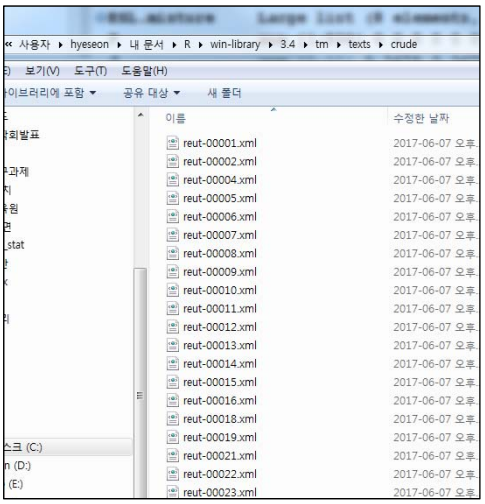
C:/Users/hyeseon/Documents/R/win-library/3.4/tm/texts



C:/Users/hyeseon/Documents/R/win-library/3.4/tm/texts/txt



Crude 데이터 (tm패키지에 들어있는 예제데이터)



8-4. 텍스트마이닝 II

10. 텍스트마이닝 데이터 저장소 : 참고

- crude 데이터

Crude 데이터 (tm패키지에 들어있는 예제데이터)