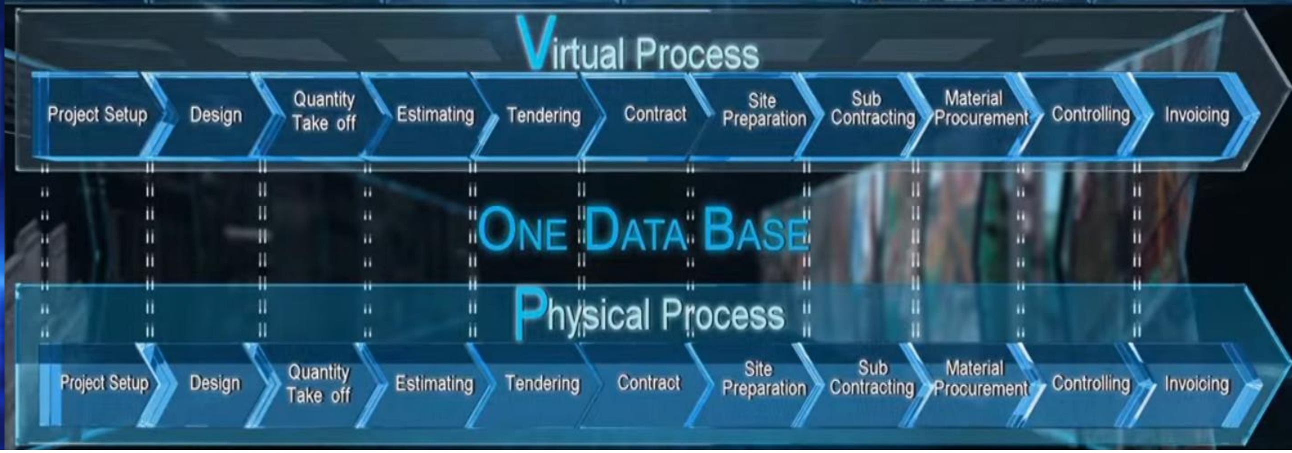
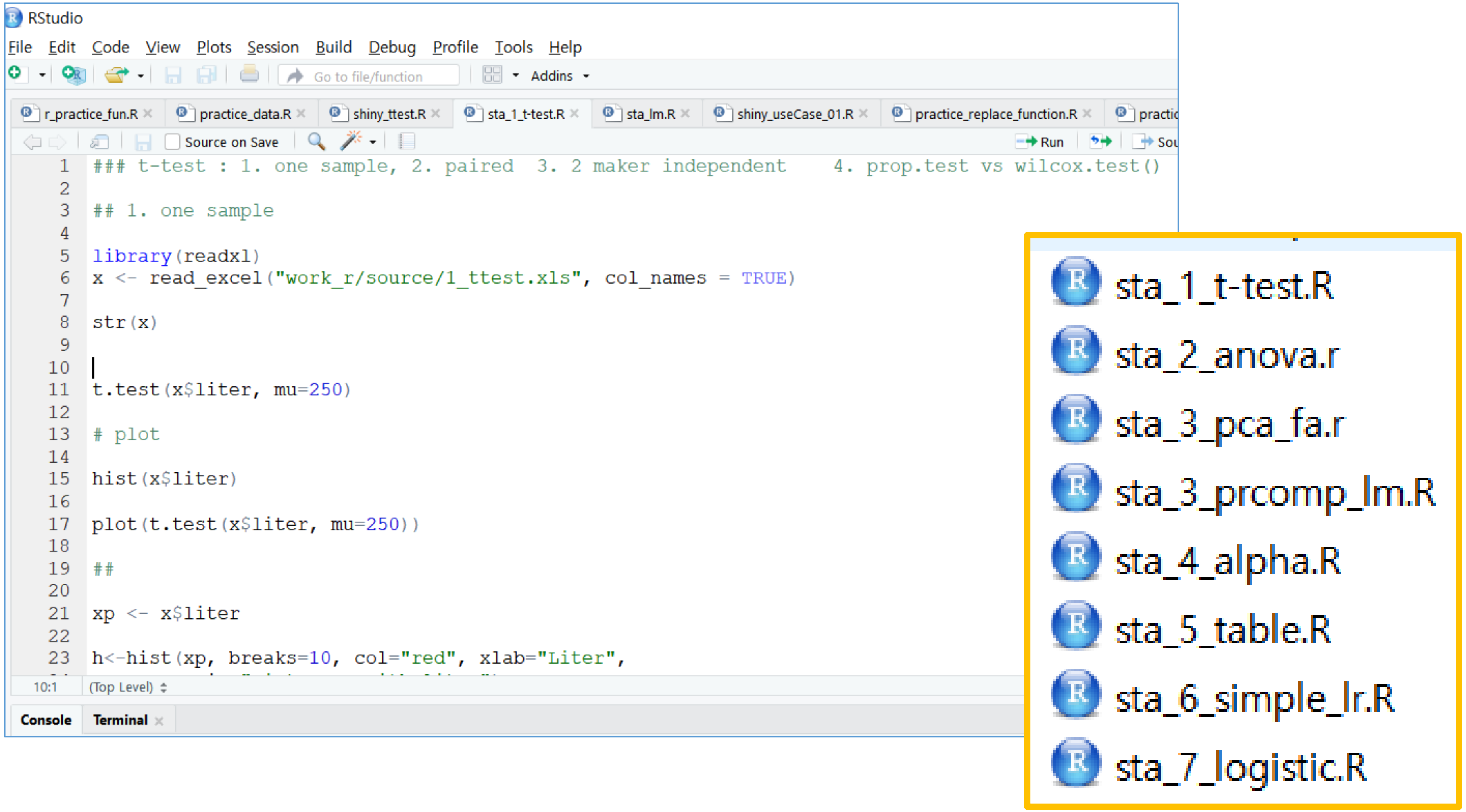


(312) 통계분석 w/ R

- t 검정, 분산분석, 타당성과 신뢰성
- 요인분석, 회귀분석
- 군집분석

R DEMO

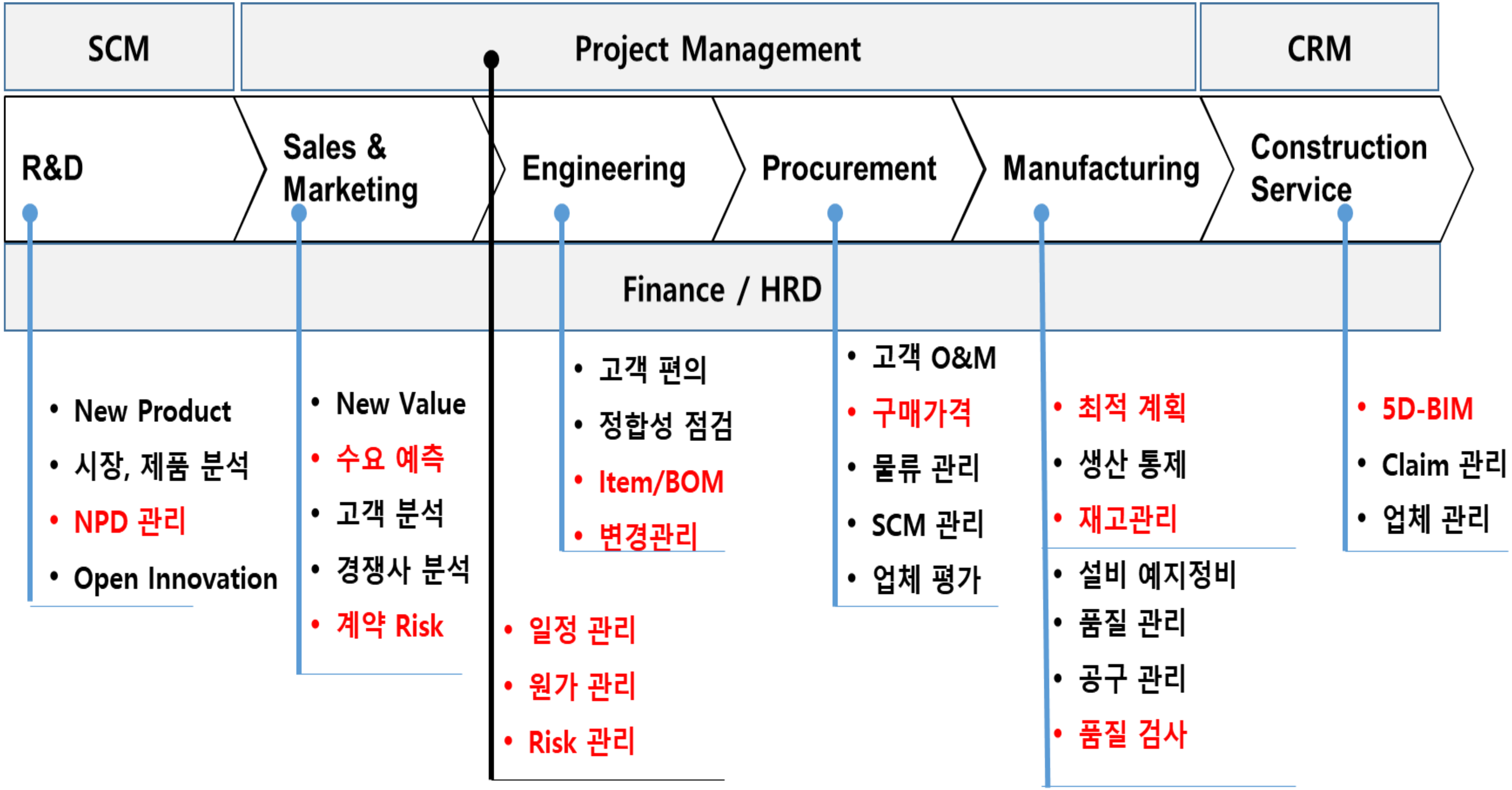


```r
### t-test : 1. one sample, 2. paired  3. 2 maker independent    4. prop.test vs wilcox.test()

## 1. one sample

library(readxl)
x <- read_excel("work_r/source/1_ttest.xls", col_names = TRUE)

str(x)


t.test(x$liter, mu=250)

# plot

hist(x$liter)

plot(t.test(x$liter, mu=250))

##

xp <- x$liter

h<-hist(xp, breaks=10, col="red", xlab="Liter",
```

- sta_1_t-test.R
- sta_2_anova.r
- sta_3_pca_fa.r
- sta_3_prcomp_lm.R
- sta_4_alpha.R
- sta_5_table.R
- sta_6_simple_lr.R
- sta_7_logistic.R

Source : shiny_ttest.r

분석 대상 분야 (R, ML 동일)



SCM

Project Management

CRM

R&D

Sales & Marketing

Engineering

Procurement

Manufacturing

Construction Service

Finance / HRD

- New Product
- 시장, 제품 분석
- NPD 관리
- Open Innovation

- New Value
- 수요 예측
- 고객 분석
- 경쟁사 분석
- 계약 Risk

- 일정 관리
- 원가 관리
- Risk 관리

- 고객 편의
- 정합성 점검
- Item/BOM
- 변경관리

- 고객 O&M
- 구매가격
- 물류 관리
- SCM 관리
- 업체 평가

- 최적 계획
- 생산 통제
- 재고관리
- 설비 예지정비
- 품질 관리
- 공구 관리
- 품질 검사

- 5D-BIM
- Claim 관리
- 업체 관리

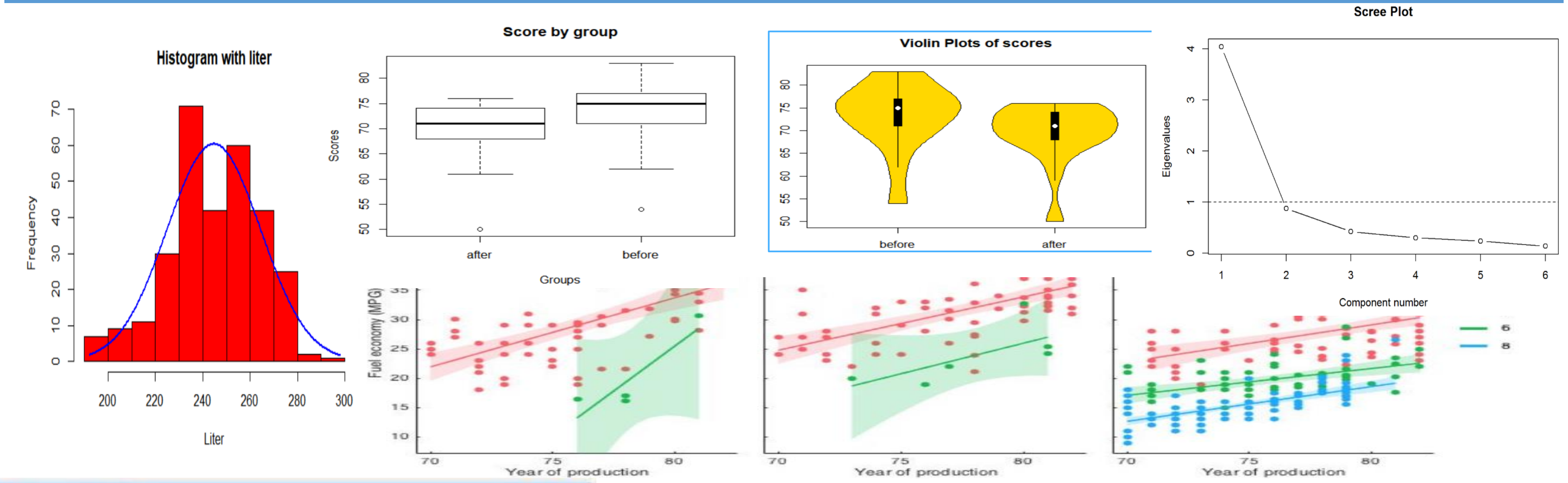
Source :

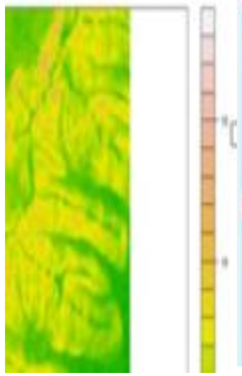
R 통계 분석 및 결과 시각화 영역

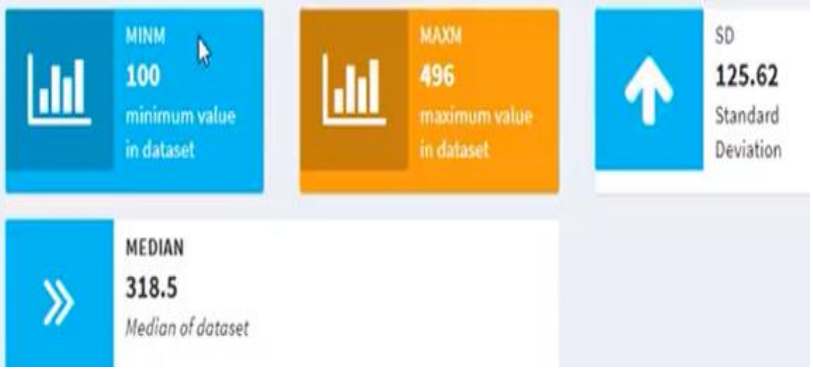
R 통계 분석

- 공분산
- t-Test (일표본, 대응, 독립)
- ANOVA (one-way, two-way, MANOVA)
- 요인분석 (PCA/FA)
- 상관분석, 신뢰도 분석
- 단순/다중 회귀분석
- 로지스틱
- 판별분석, 군집분석

분석 결과 시각화 영역







MINM
100
minimum value in dataset

MAXM
496
maximum value in dataset

SD
125.62
Standard Deviation

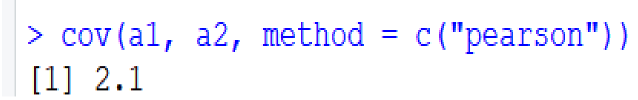
MEDIAN
318.5
Median of dataset

Source :

공분산, 상관계수

목적 및 절차

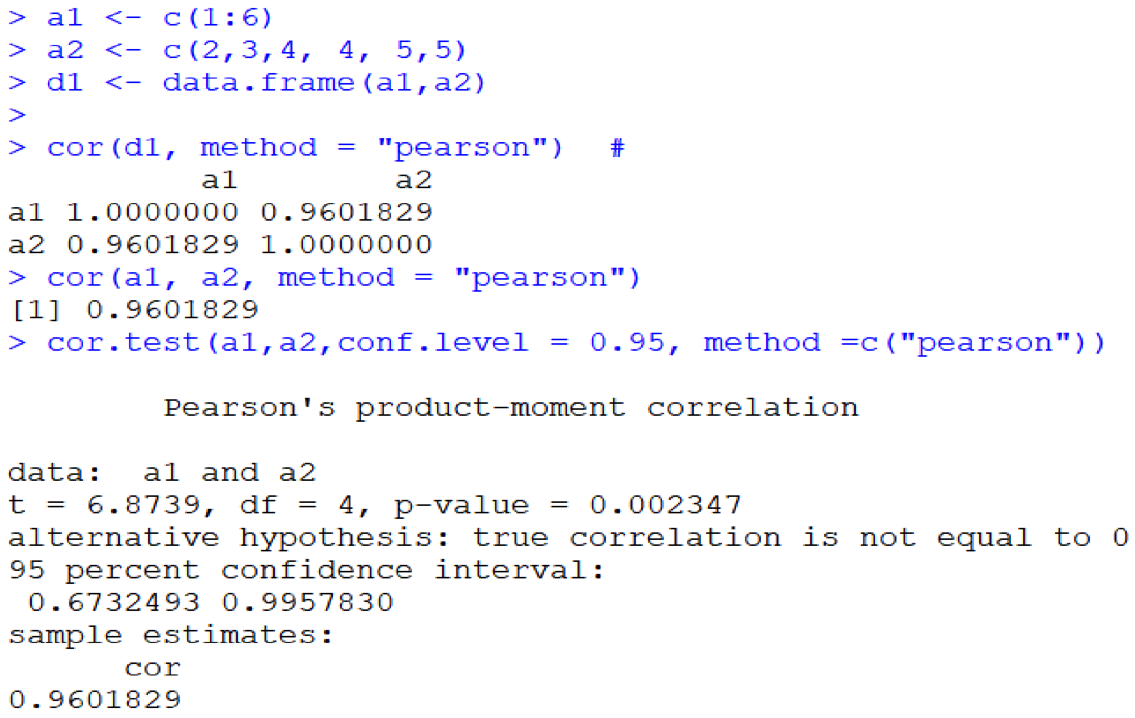
- library(MASS)
- x <- Cars93$MPG.highway
- y <- Cars93$Weight
- cov(x, y, method = c("pearson"))
- 결측치 확인 sum(is.na(x)) / sum(is.na(y))
- cor(x,y, method = c("pearson"))



```
> cov(a1, a2, method = c("pearson"))
[1] 2.1
```

분석내용 및 결과해석

- 통계량 및 검정결과

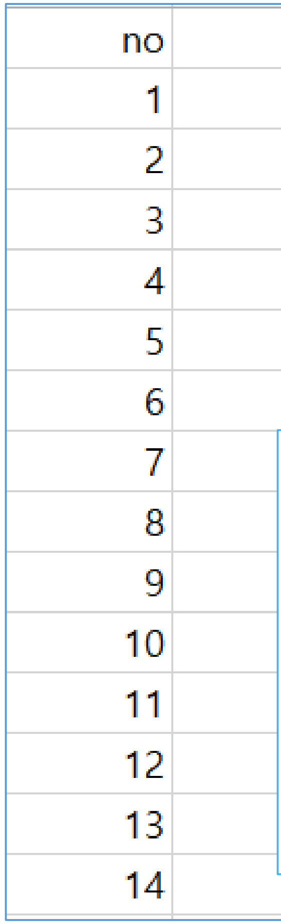


```
> a1 <- c(1:6)
> a2 <- c(2,3,4, 4, 5,5)
> d1 <- data.frame(a1,a2)
>
> cor(d1, method = "pearson")   #
          a1        a2
a1 1.0000000 0.9601829
a2 0.9601829 1.0000000
> cor(a1, a2, method = "pearson")
[1] 0.9601829
> cor.test(a1,a2,conf.level = 0.95, method =c("pearson"))

        Pearson's product-moment correlation

data:  a1 and a2
t = 6.8739, df = 4, p-value = 0.002347
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6732493 0.9957830
sample estimates:
      cor
0.9601829
```

Source :

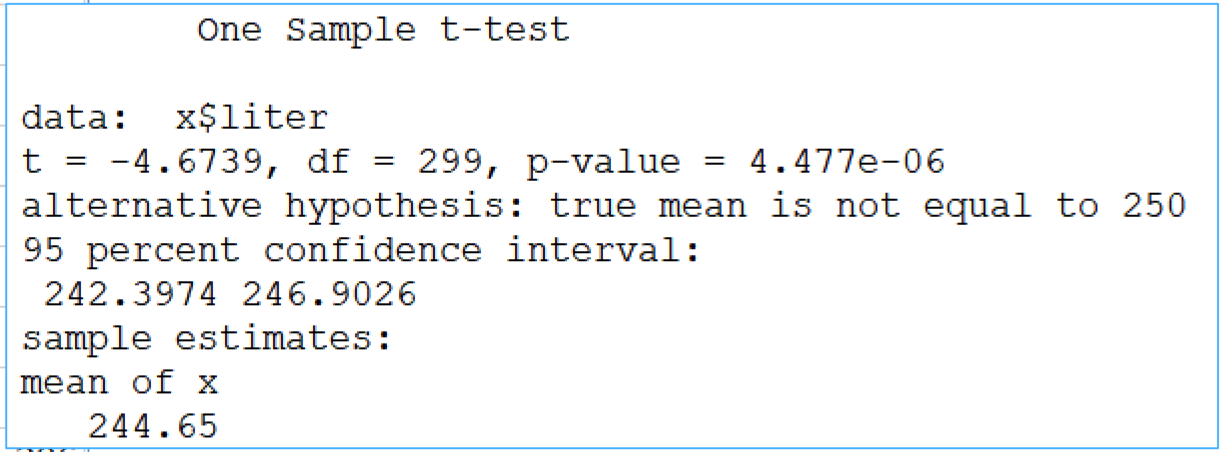
t-test (일표본)

- **Data**

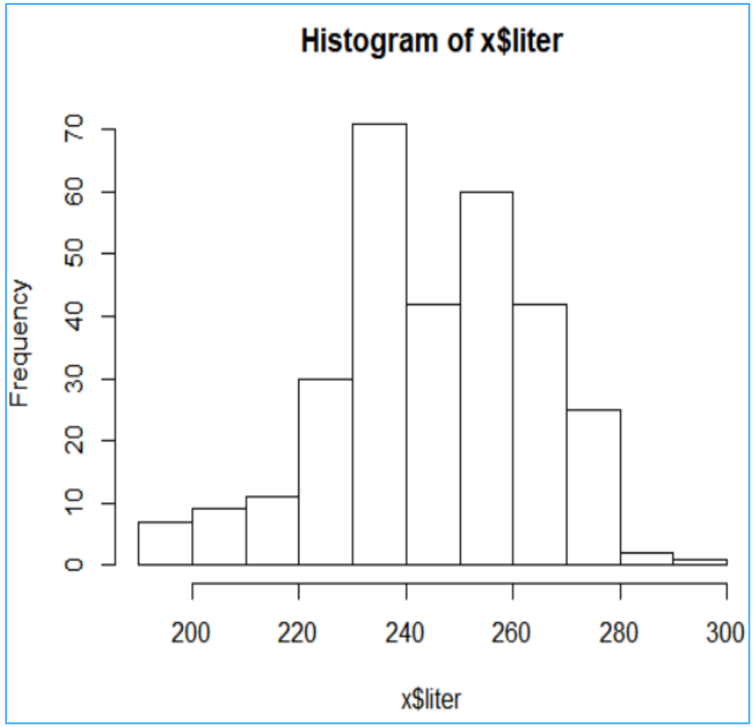


no	liter
1	252
2	271
3	282
4	257
5	240
6	242
7	
8	
9	
10	
11	
12	
13	
14	206

- **기술통계**
  - ➢ library(readxl)
  - ➢ x <- read_excel("work_r/source/1_ttest.xls", col_names = TRUE)
  - ➢ str(x)
  - ➢ t.test(x$liter, mu=250)
  - ➢ hist(x$liter)



```
        One Sample t-test

data:  x$liter
t = -4.6739, df = 299, p-value = 4.477e-06
alternative hypothesis: true mean is not equal to 250
95 percent confidence interval:
 242.3974 246.9026
sample estimates:
mean of x
   244.65
```

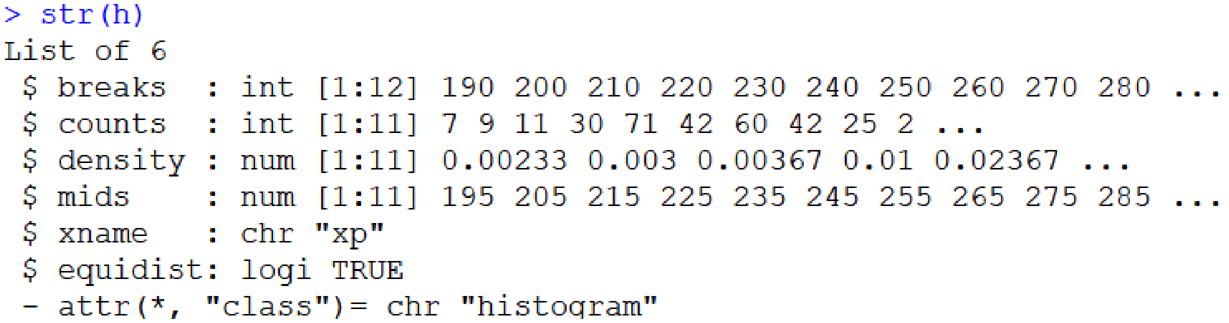


Source : sta_1_t_test

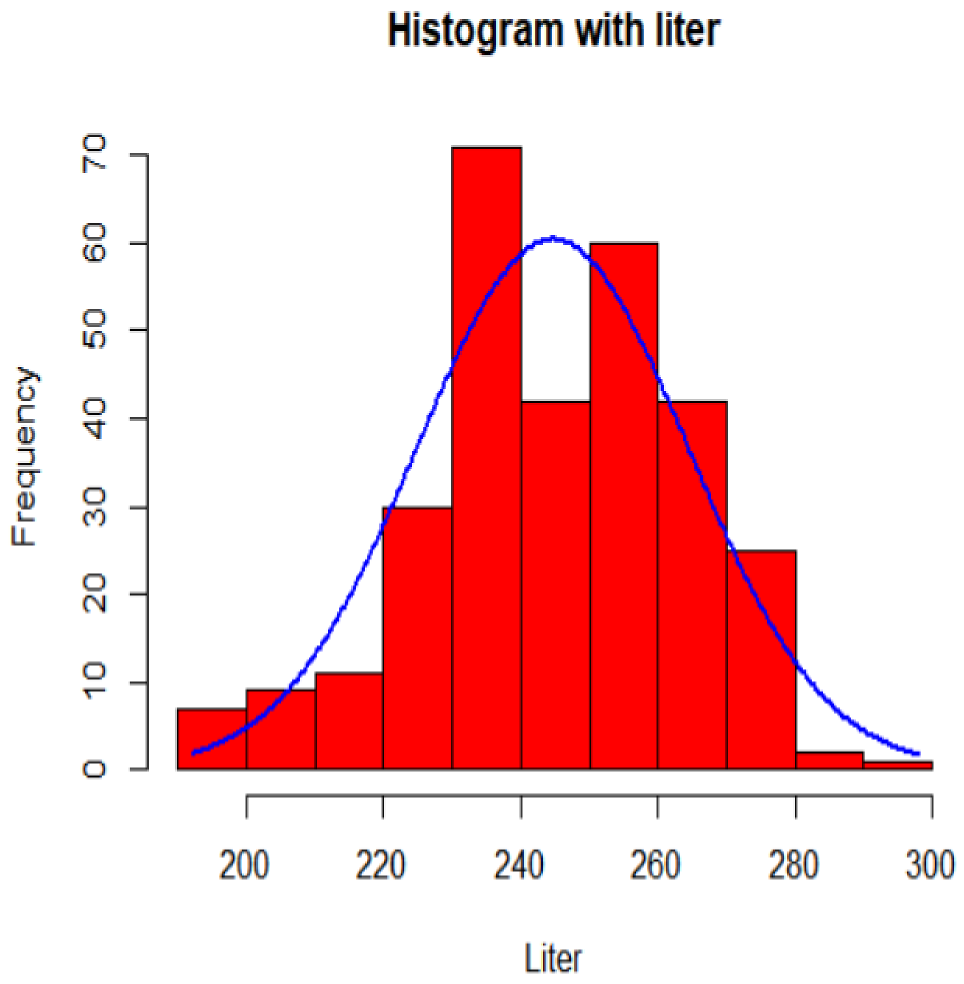
t-test (일표본) – 정규 분포 곡선

```
xp <- x$liter

h <- hist(xp, breaks=10, col="red", xlab="Liter",
        main="Histogram with liter")
```

```
xfit<-seq(min(xp),max(xp),length=300)

yfit<-dnorm(xfit,mean=mean(xp),sd=sd(xp))
yfit <- yfit*diff(h$mids[1:2])*length(xp) # consider frequency

lines(xfit, yfit, col="blue", lwd=2)
```



```
> str(h)
List of 6
 $ breaks  : int [1:12] 190 200 210 220 230 240 250 260 270 280 ...
 $ counts  : int [1:11] 7 9 11 30 71 42 60 42 25 2 ...
 $ density : num [1:11] 0.00233 0.003 0.00367 0.01 0.02367 ...
 $ mids    : num [1:11] 195 205 215 225 235 245 255 265 275 285 ...
 $ xname   : chr "xp"
 $ equidist: logi TRUE
 - attr(*, "class")= chr "histogram"
```

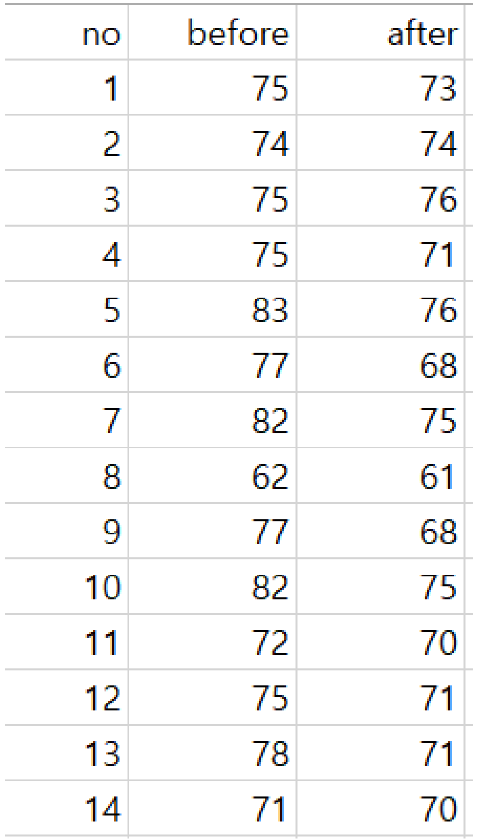


Source : sta_1_t_test

https://www.statmethods.net/graphs/density.html

t-test (대응표본)

- **Data**



no	before	after
1	75	73
2	74	74
3	75	76
4	75	71
5	83	76
6	77	68
7	82	75
8	62	61
9	77	68
10	82	75
11	72	70
12	75	71
13	78	71
14	71	70

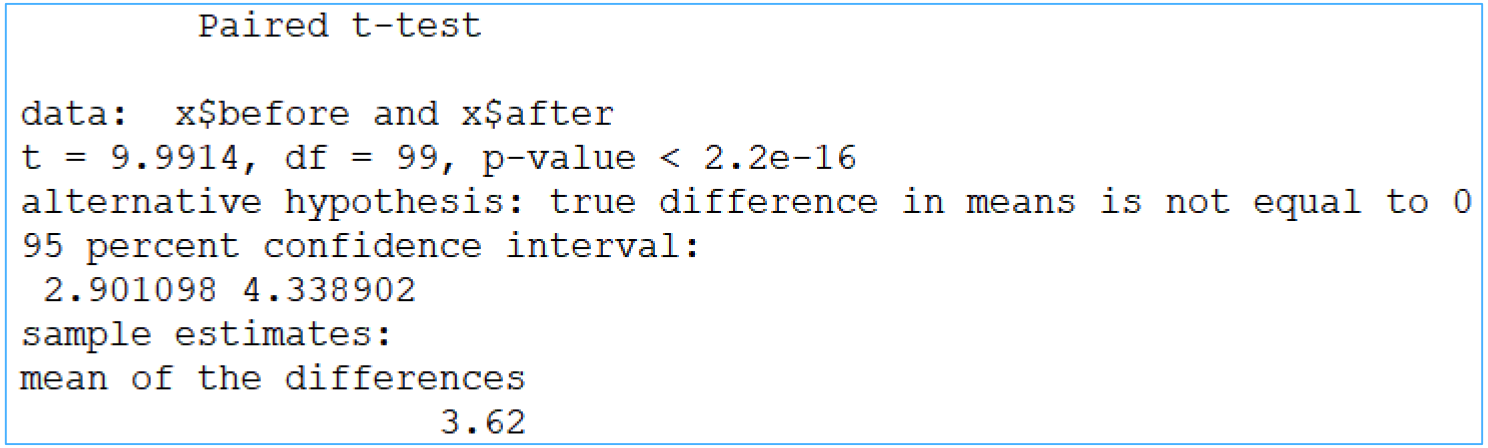
- **기술통계**

➢ library(readxl)

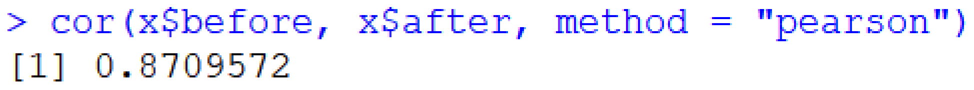
➢ x <- read_excel("work_r/source/2_pttest.xls", col_names = TRUE)

➢ t.test(x$before, x$after, var.equal=T, paired=T)

➢ cor(x$before, x$after, method = "pearson")



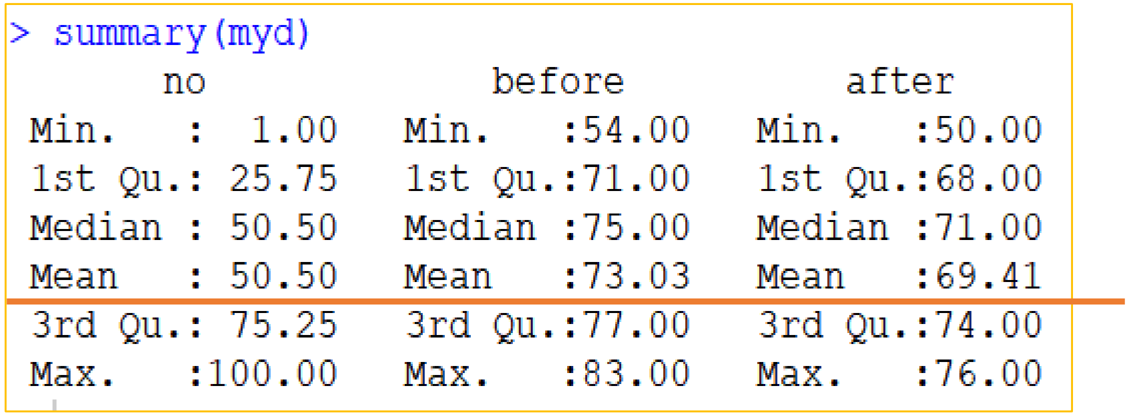
```
        Paired t-test

data:  x$before and x$after
t = 9.9914, df = 99, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 2.901098 4.338902
sample estimates:
mean of the differences
                   3.62
```



```
> cor(x$before, x$after, method = "pearson")
[1] 0.8709572
```

Source : sta_1_t_test , 2_pttest.xls

t-test (대응표본)

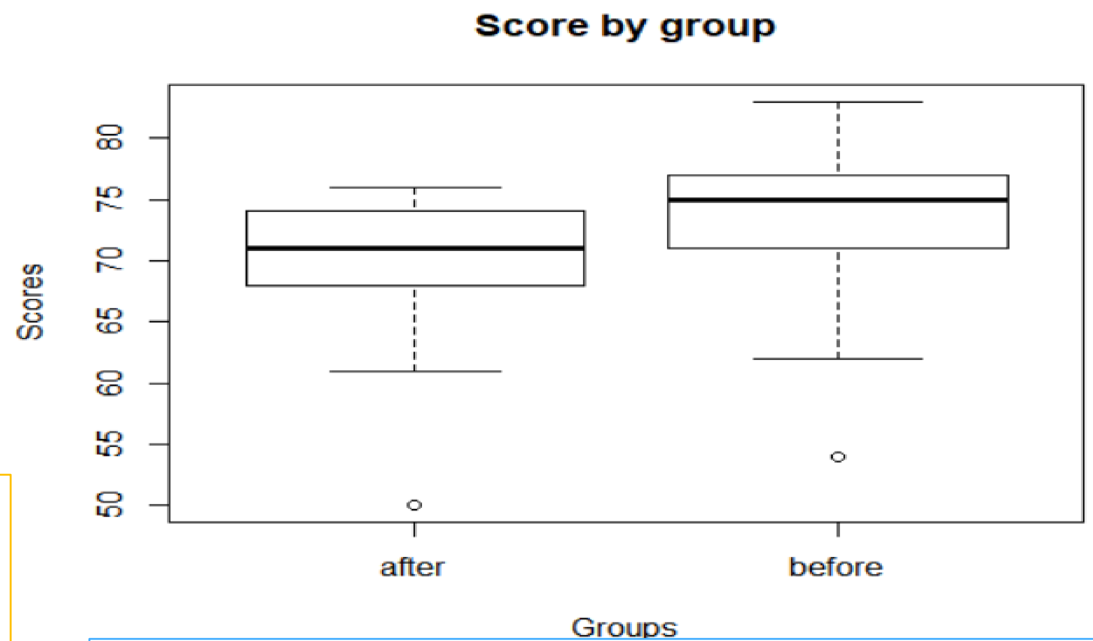


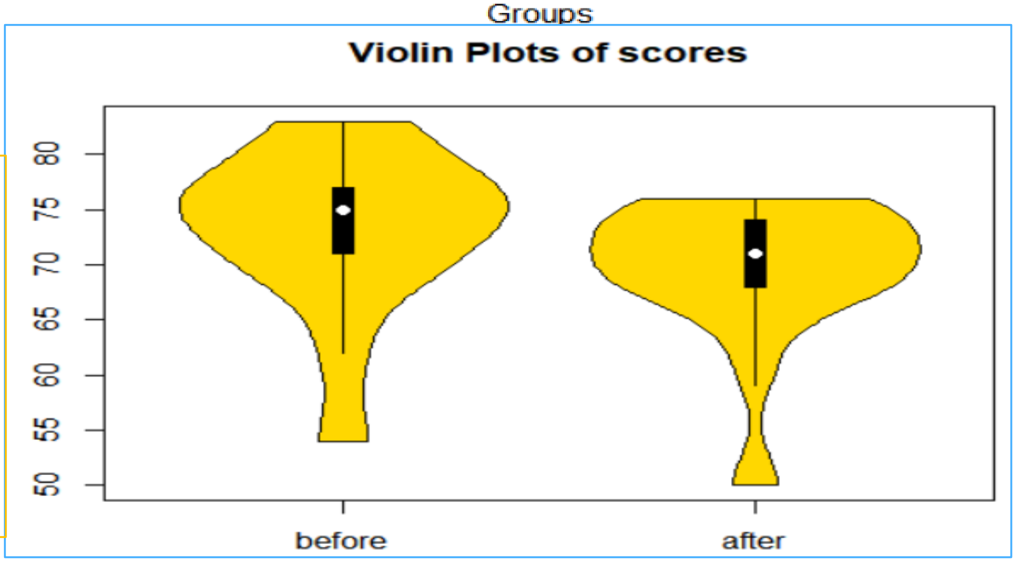
```
> summary(myd)
       no             before          after      
 Min.   :  1.00   Min.   :54.00   Min.   :50.00  
 1st Qu.: 25.75   1st Qu.:71.00   1st Qu.:68.00  
 Median : 50.50   Median :75.00   Median :71.00  
 Mean   : 50.50   Mean   :73.03   Mean   :69.41  
 3rd Qu.: 75.25   3rd Qu.:77.00   3rd Qu.:74.00  
 Max.   :100.00   Max.   :83.00   Max.   :76.00  
```

```
➢ x1 <- myd[, 2]
➢ str(x1)
➢ x1$group = "before"
➢ names(x1)<-c("score","group")
➢ str(x1)
```

```
➢ x2 <- myd[, 3]
➢ x2$group = "after"
➢ names(x2)<-c("score","group")

➢ x12 <- rbind(x1, x2)
```

```
➢ boxplot(score~group,data=x12, main="Score by group",  xlab="Groups",
  ylab="Scores")

➢ library(vioplot)
➢ x1 <- myd$before
➢ x2 <- myd$after
➢ vioplot(x1, x2, names=c("before", "after"), col="gold")
➢ title("Violin Plots of scores")
```

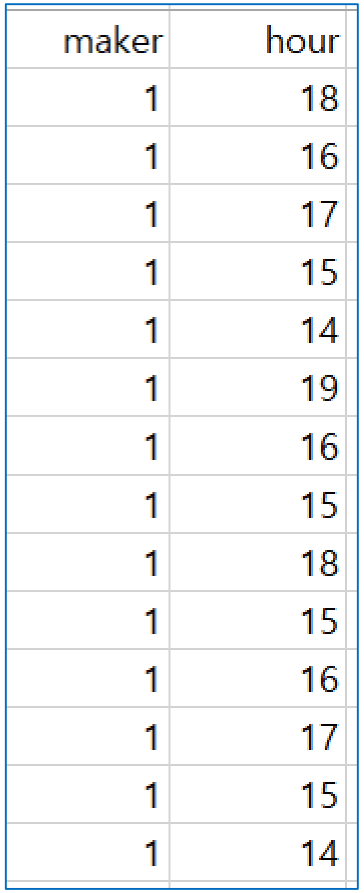




Source : sta_1_t_test / https://www.statmethods.net/graphs/boxplot.html

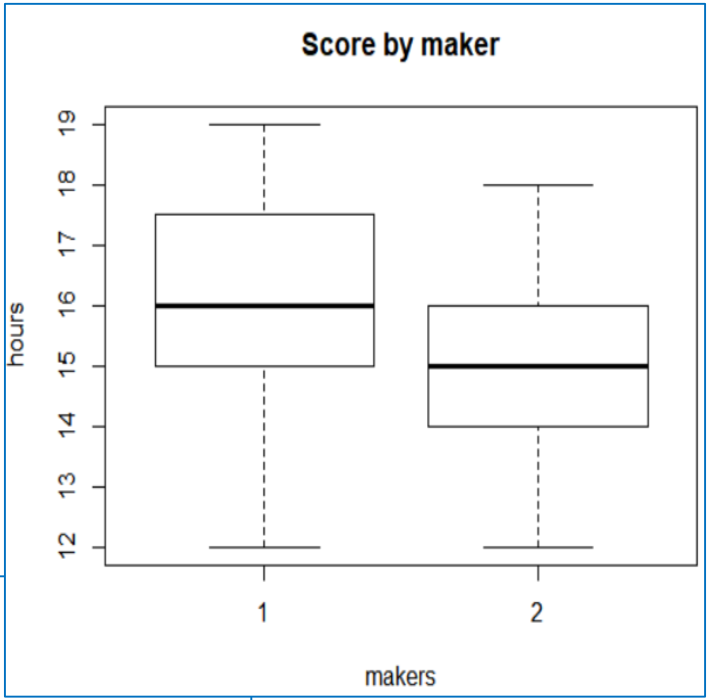
2 t-test (독립표본)

- **Data : maker (1, 2)**



maker	hour
1	18
1	16
1	17
1	15
1	14
1	19
1	16
1	15
1	18
1	15
1	16
1	17
1	15
1	14

```
x <- read_excel("work_r/source/3_2ittest.xls", col_names = TRUE)

t.test (x$hour ~ x$maker)


boxplot(hour~maker,data=x, main="Score by maker",
        xlab="makers", ylab="hours")
```



```
        Welch Two Sample t-test

data:  x$hour by x$maker
t = 6.3744, df = 197.98, p-value = 1.265e-09
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.9737946 1.8462054
sample estimates:
mean in group 1 mean in group 2
          16.35           14.94
```

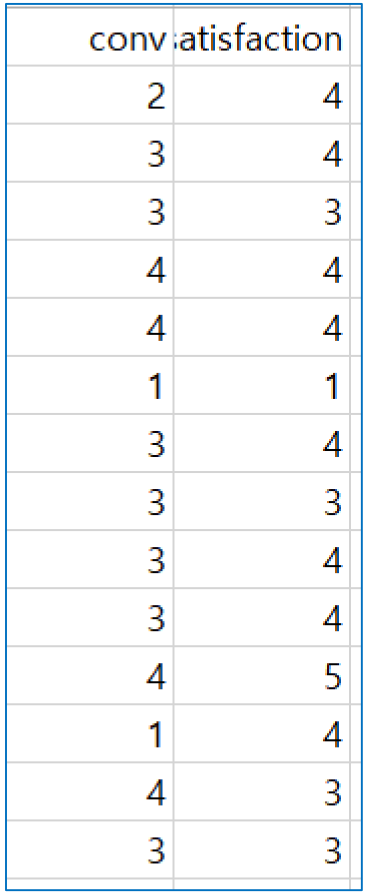
Source : sta_1_t_test , 3_2ittest.xls

t-test distribution curve

Source : sta_1_t_test , 2_pttest.xls

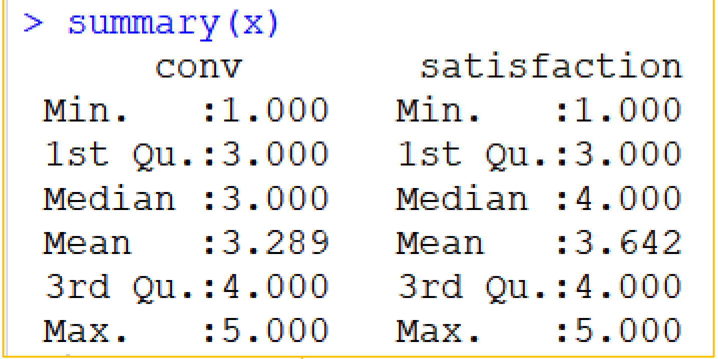
one-way ANOVA

- **Data : satisfaction by convenient**

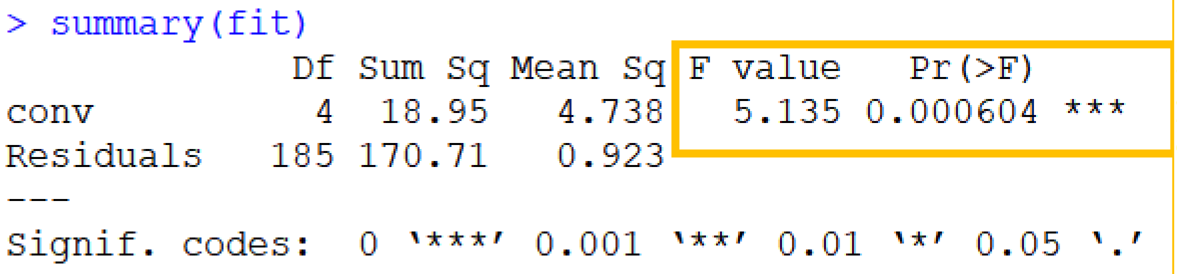


conv	satisfaction
2	4
3	4
3	3
4	4
4	4
1	1
3	4
3	3
3	4
3	4
4	5
1	4
4	3
3	3

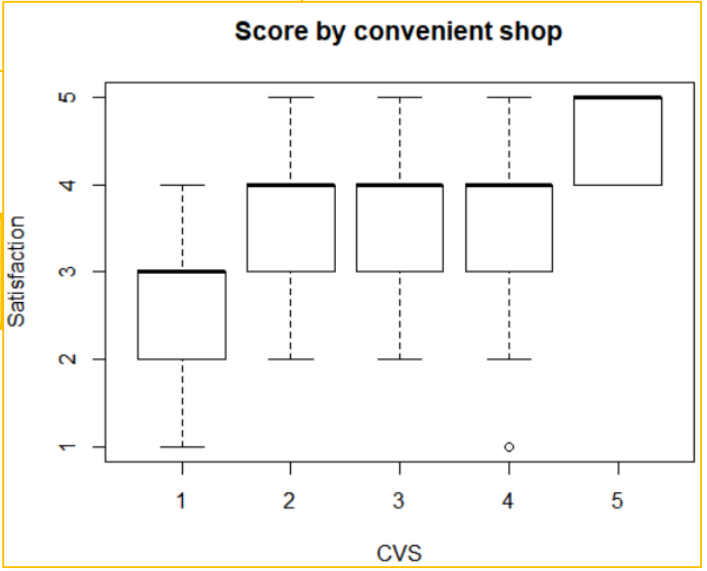
```
x <- read_excel("work_r/source/4_oneway_anova.xlsx", col_names = TRUE)
summary(x)
x$conv = as.factor(x$conv)
str(x)
fit <- aov(satisfaction ~ conv, data = x)
summary(fit)
boxplot(satisfaction ~ conv, data = x, main="Score by convenient shop",
        xlab="CVS", ylab="Satisfaction")
```



```
> summary(x)
      conv        satisfaction
 Min.   :1.000   Min.   :1.000
 1st Qu.:3.000   1st Qu.:3.000
 Median :3.000   Median :4.000
 Mean   :3.289   Mean   :3.642
 3rd Qu.:4.000   3rd Qu.:4.000
 Max.   :5.000   Max.   :5.000
```



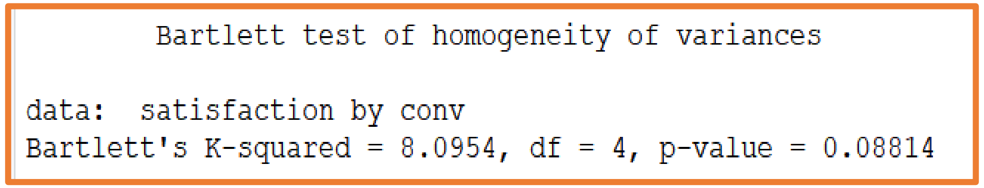
```
> summary(fit)
             Df Sum Sq Mean Sq F value   Pr(>F)    
conv          4  18.95   4.738   5.135 0.000604 ***
Residuals   185 170.71   0.923                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
```



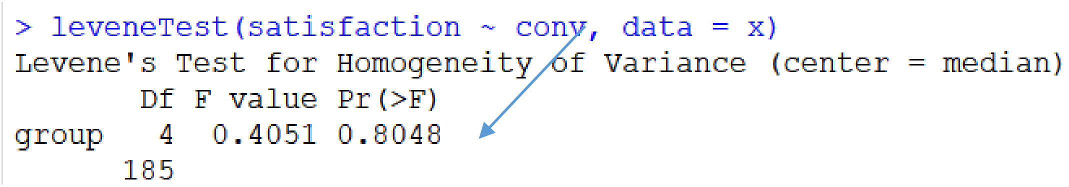
Source : sta_2_anova.R / 4_oneway_anova.xlsx

one-way ANOVA

```
bartlett.test(satisfaction ~ conv, data = x)
##
library(car)
leveneTest(satisfaction ~ conv, data = x)
```



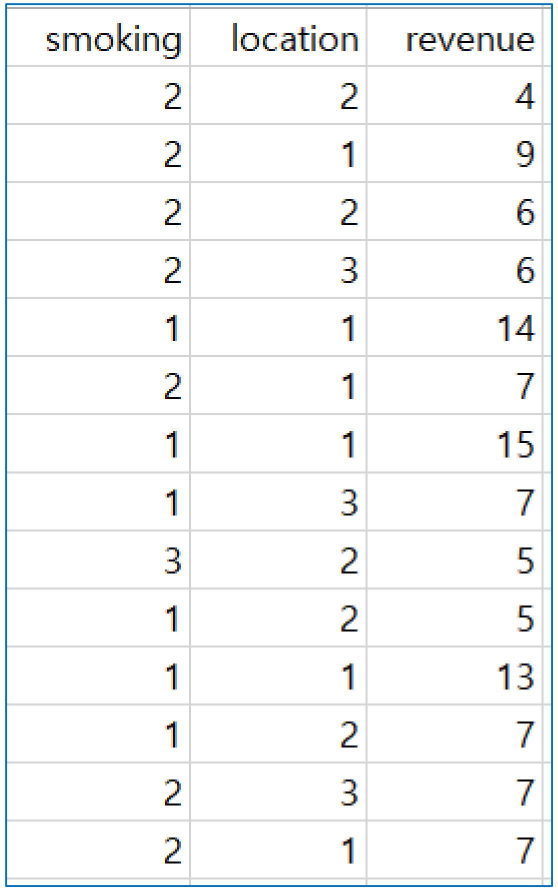
```
        Bartlett test of homogeneity of variances

data:  satisfaction by conv
Bartlett's K-squared = 8.0954, df = 4, p-value = 0.08814
```



```
> leveneTest(satisfaction ~ conv, data = x)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   4  0.4051 0.8048
      185
```

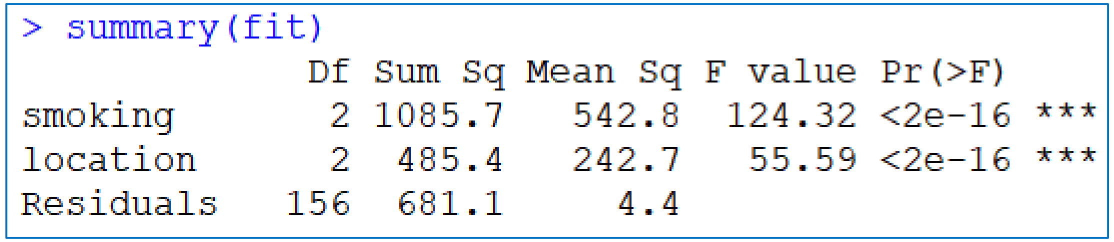
Source : sta_2_anova.R / 4_oneway_anova.xlsx

Two-way ANOVA



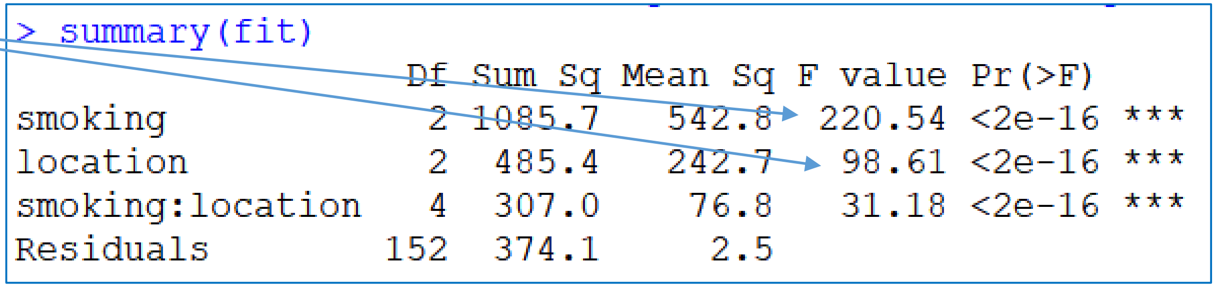
smoking	location	revenue
2	2	4
2	1	9
2	2	6
2	3	6
1	1	14
2	1	7
1	1	15
1	3	7
3	2	5
1	2	5
1	1	13
1	2	7
2	3	7
2	1	7

```
x <- read_excel("work_r/source/5_twoway_anova.xlsx", col_names = TRUE)
summary(x)

x$smoking = as.factor(x$smoking)
x$location = as.factor(x$location)
str(x)

fit <- aov(revenue ~ smoking + location, data = x)
summary(fit)

fit <- aov(revenue ~ smoking + location + smoking * location , data = x)
```



```
> summary(fit)
             Df Sum Sq Mean Sq F value Pr(>F)    
smoking       2 1085.7   542.8  124.32 <2e-16 ***
location      2  485.4   242.7   55.59 <2e-16 ***
Residuals   156  681.1     4.4                   
```

수치 상이함



```
> summary(fit)
                  Df Sum Sq Mean Sq F value Pr(>F)    
smoking            2 1085.7   542.8  220.54 <2e-16 ***
location           2  485.4   242.7   98.61 <2e-16 ***
smoking:location   4  307.0    76.8   31.18 <2e-16 ***
Residuals        152  374.1     2.5                   
```

Source : sta_2_anova.R / 5_twoway_anova.xlsx

요인분석 (PCA/FA)

General methods for principal component analysis

There are two general methods to perform PCA in R :

Spectral decomposition which examines the covariances / correlations between variables

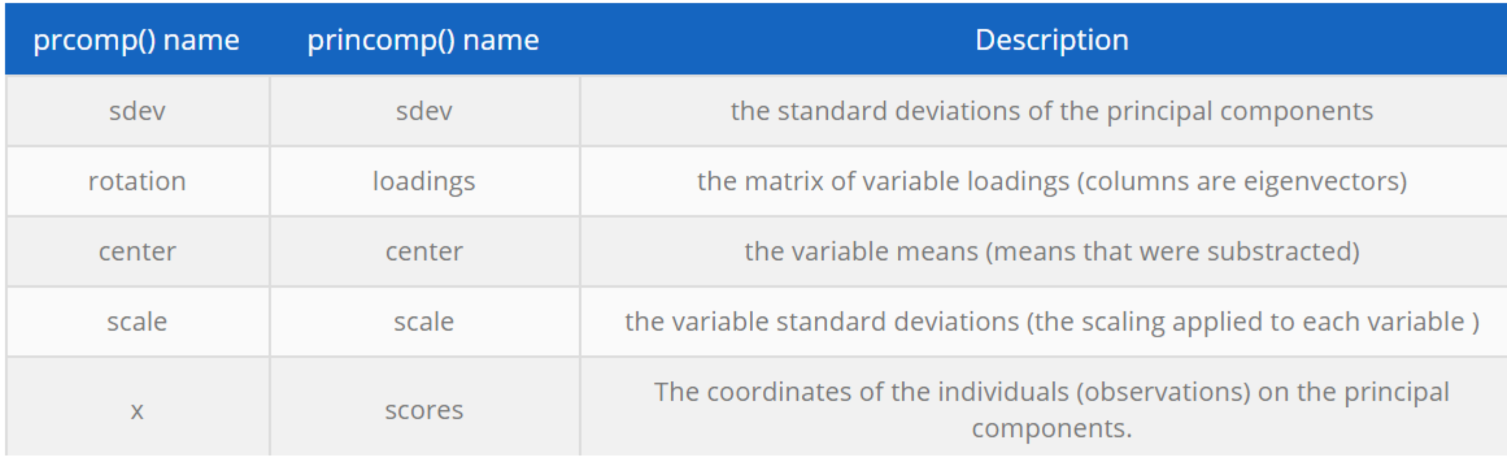
Singular value decomposition which examines the covariances / correlations between individuals

The function **princomp**() uses the spectral decomposition approach. The functions **prcomp**() and **PCA**()[FactoMineR] use the singular value decomposition (SVD).

prcomp() and princomp() functions

The simplified format of these 2 functions are :

prcomp(x, scale = FALSE) princomp(x, cor = FALSE, scores = TRUE)

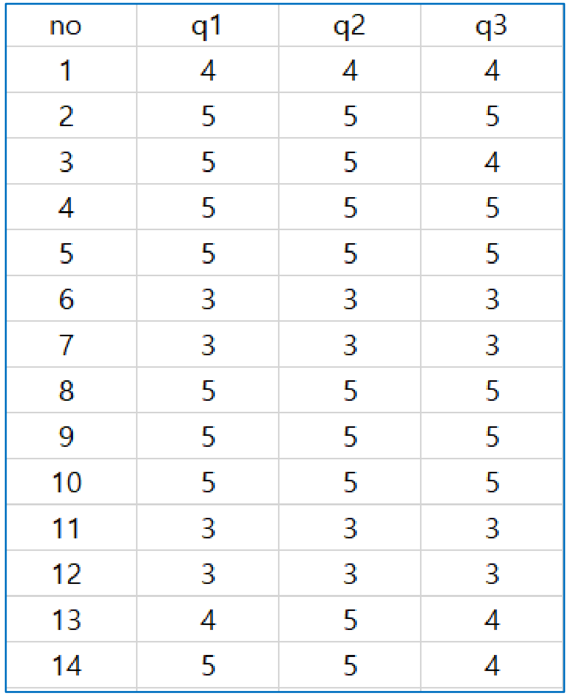


prcomp() name	princomp() name	Description
sdev	sdev	the standard deviations of the principal components
rotation	loadings	the matrix of variable loadings (columns are eigenvectors)
center	center	the variable means (means that were substracted)
scale	scale	the variable standard deviations (the scaling applied to each variable)
x	scores	The coordinates of the individuals (observations) on the principal components.

Source : http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/118-principal-component-analysis-in-r-prcomp-vs-princomp/

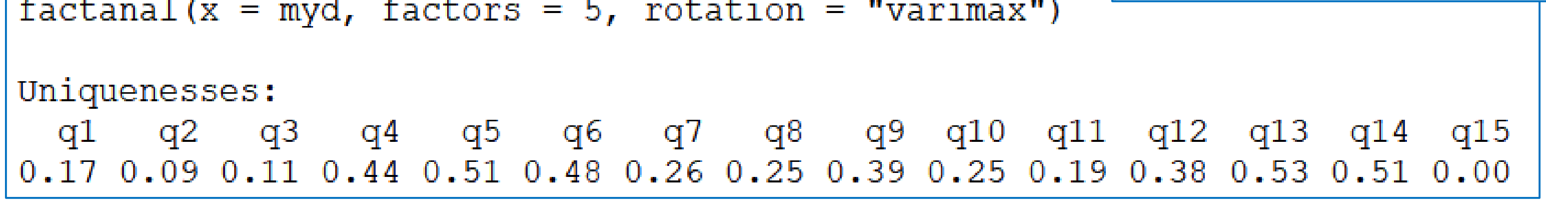
요인분석 (PCA/FA) – factanal

- **Data : q1 ~ q15**

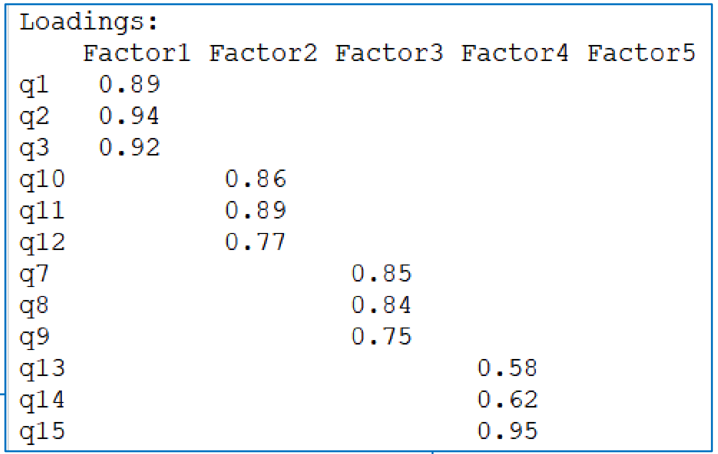


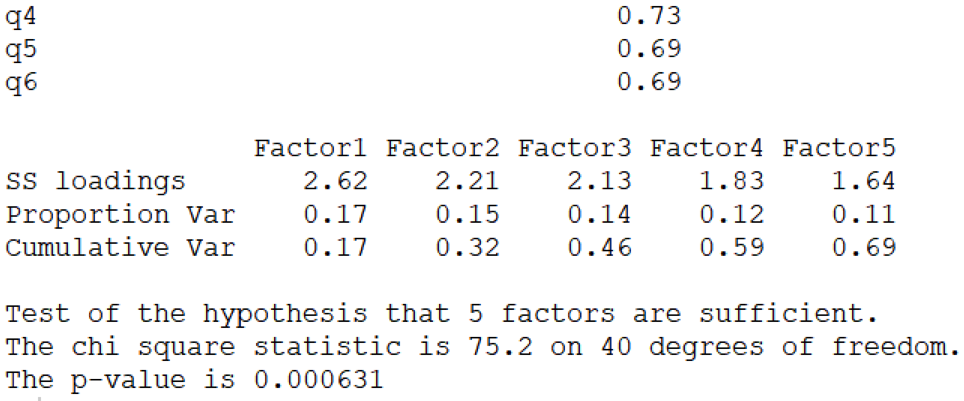
no	q1	q2	q3
1	4	4	4
2	5	5	5
3	5	5	4
4	5	5	5
5	5	5	5
6	3	3	3
7	3	3	3
8	5	5	5
9	5	5	5
10	5	5	5
11	3	3	3
12	3	3	3
13	4	5	4
14	5	5	4

```
myd <- read_excel("work_r/source/6_pca.xls", col_names = TRUE)
str(myd)

library(dplyr)

myd <- myd %>% select(2 : 16)

fit <- factanal(myd, 5, rotation="varimax")

print(fit, digits=2, cutoff=.3, sort=TRUE)
```



```
Call:
factanal(x = myd, factors = 5, rotation = "varimax")

Uniquenesses:
  q1   q2   q3   q4   q5   q6   q7   q8   q9  q10  q11  q12  q13  q14  q15 
0.17 0.09 0.11 0.44 0.51 0.48 0.26 0.25 0.39 0.25 0.19 0.38 0.53 0.51 0.00 
```



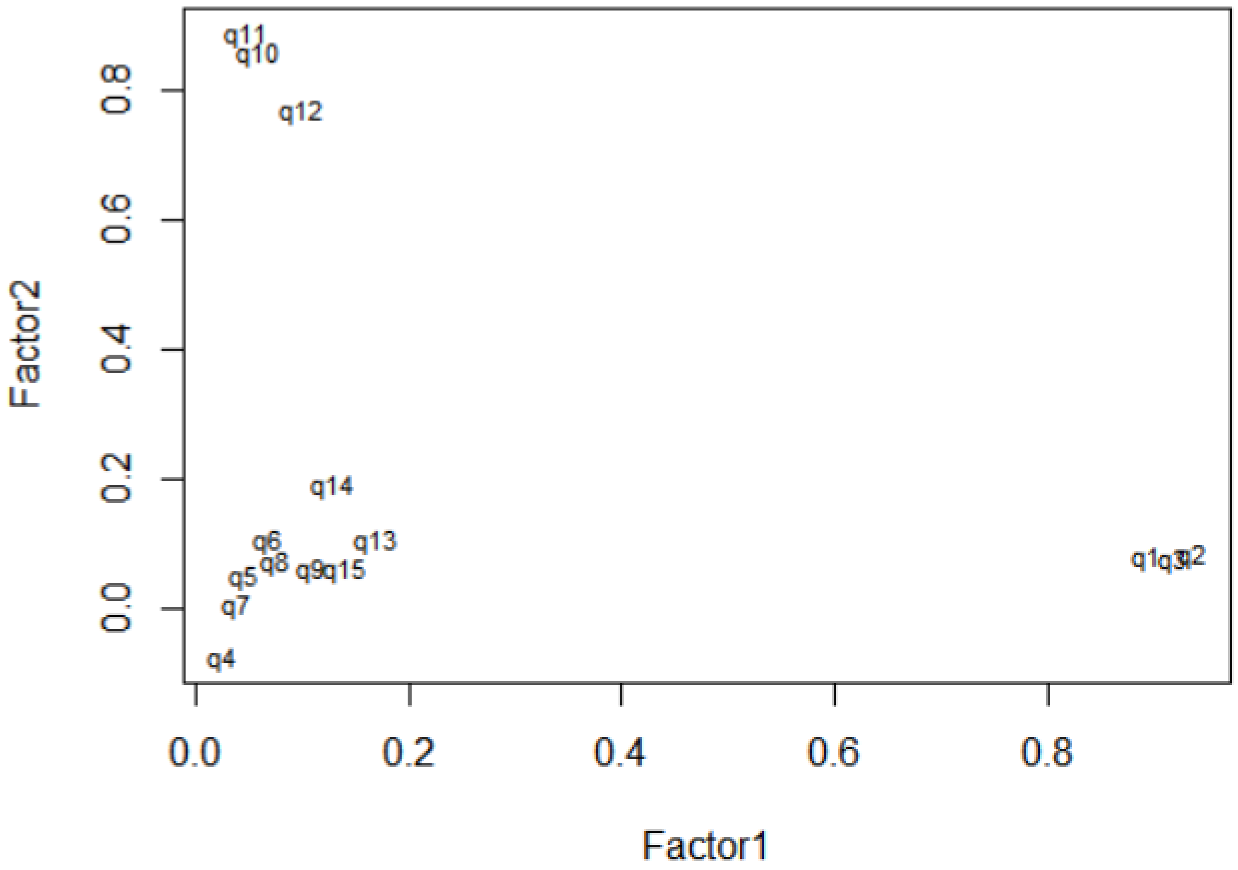


```
Loadings:
    Factor1 Factor2 Factor3 Factor4 Factor5
q1   0.89                                  
q2   0.94                                  
q3   0.92                                  
q10          0.86                          
q11          0.89                          
q12          0.77                          
q7                   0.85                  
q8                   0.84                  
q9                   0.75                  
q13                          0.58          
q14                          0.62          
q15                          0.95          
q4                                   0.73  
q5                                   0.69  
q6                                   0.69  

               Factor1 Factor2 Factor3 Factor4 Factor5
SS loadings       2.62    2.21    2.13    1.83    1.64
Proportion Var    0.17    0.15    0.14    0.12    0.11
Cumulative Var    0.17    0.32    0.46    0.59    0.69

Test of the hypothesis that 5 factors are sufficient.
The chi square statistic is 75.2 on 40 degrees of freedom.
The p-value is 0.000631
```

Source : sta_3_pca_fa.R , 6_pca.xls

요인분석 (PCA/FA) – factanal

```
# plot factor 1 by factor 2
load <- fit$loadings[,1:2]
plot(load,type="n") # set up plot
text(load,labels=names(myd),cex=.7) # add variable names
```

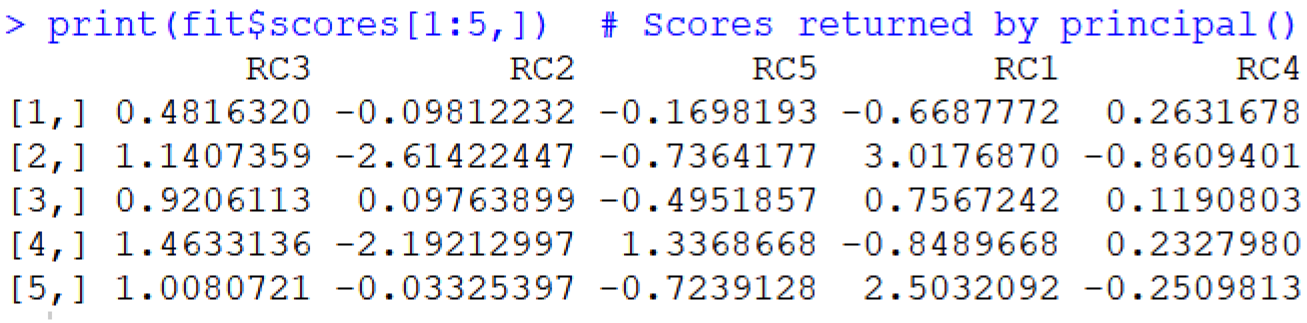


Source : sta_3_pca_fa.R , 6_pca.xls

요인분석 (PCA/FA) – principal

fit <- psych::principal(myd, rotate="varimax", nfactors=5, scores=TRUE)

print(fit$scores[1:5,]) # Scores returned by principal()

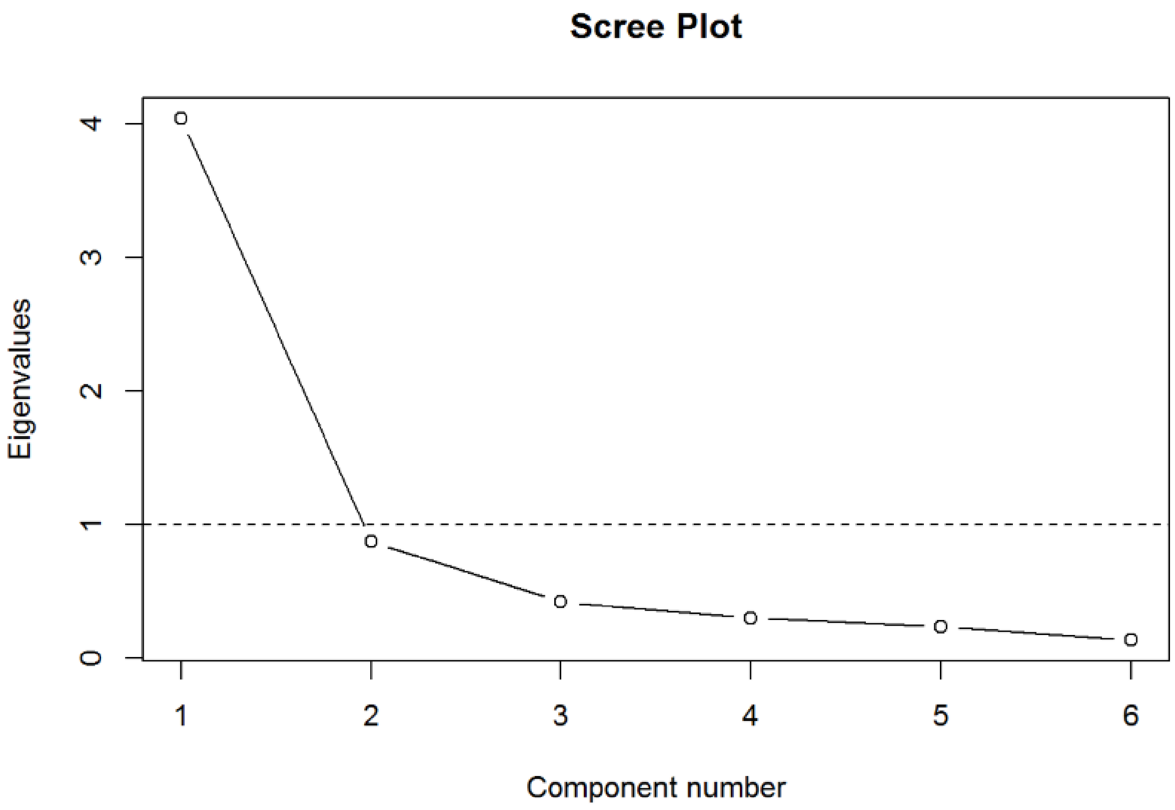


```
> print(fit$scores[1:5,])  # Scores returned by principal()
           RC3         RC2        RC5        RC1        RC4
[1,] 0.4816320 -0.09812232 -0.1698193 -0.6687772  0.2631678
[2,] 1.1407359 -2.61422447 -0.7364177  3.0176870 -0.8609401
[3,] 0.9206113  0.09763899 -0.4951857  0.7567242  0.1190803
[4,] 1.4633136 -2.19212997  1.3368668 -0.8489668  0.2327980
[5,] 1.0080721 -0.03325397 -0.7239128  2.5032092 -0.2509813
```

Source : sta_3_pca_fa.R , 6_pca.xls

요인분석 (PCA/FA) – principal

Principal Components Analysis using R



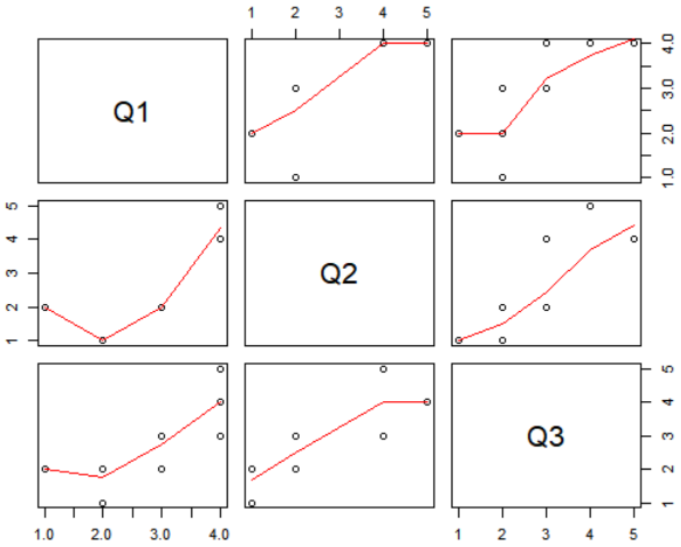
Source : http://faculty.missouri.edu/huangf/data/mvnotes/pca_in_r_2.html

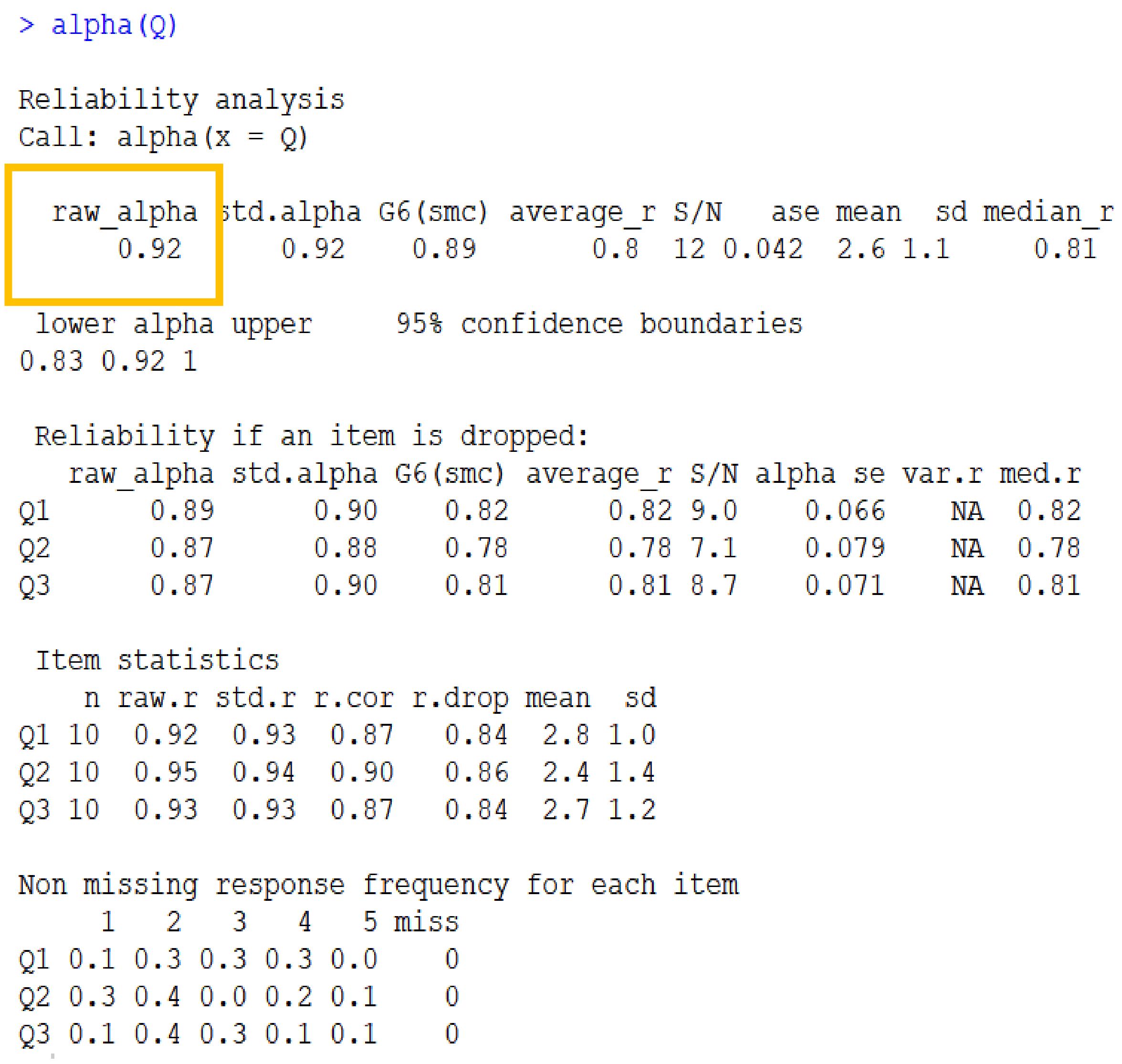
요인분석 (PCA/FA) – principal

Source : http://faculty.missouri.edu/huangf/data/mvnotes/pca_in_r_2.html

신뢰도 분석

```
Q <- data.frame(
    Q1=c(1,4,2,3,4,2,3,4,3,2),
    Q2=c(2,4,1,2,4,1,2,5,2,1),
    Q3=c(2,5,1,3,3,2,3,4,2,2))
pairs(Q, panel=panel.smooth)
# 2. cronbach : install.packages("psy")  /  alpha ()
library(psy)
cronbach(Q)

library(psych)
alpha(Q)
a <- alpha(Q)
str(a)
a$total
```



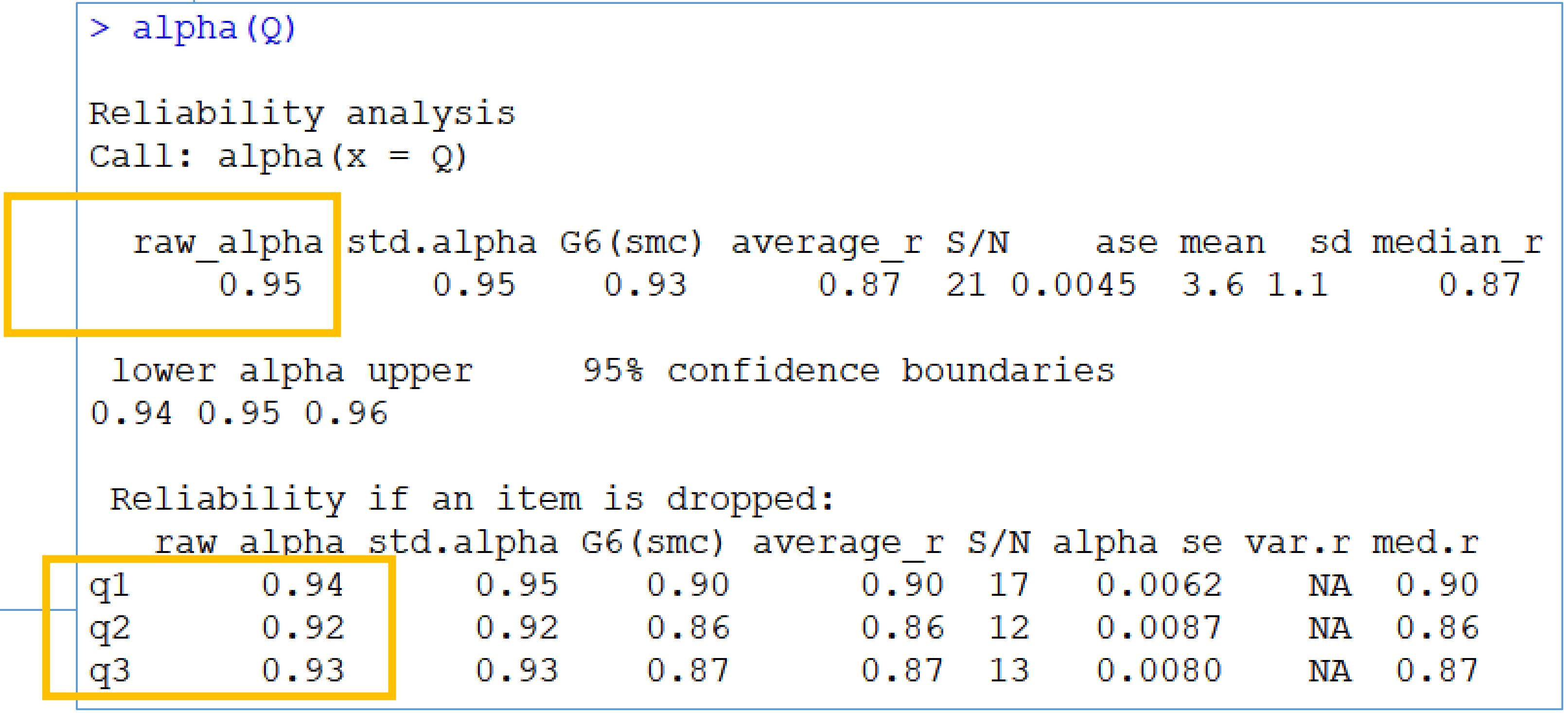


```
> alpha(Q)

Reliability analysis
Call: alpha(x = Q)

  raw_alpha std.alpha G6(smc) average_r S/N   ase mean  sd median_r
      0.92      0.92     0.89       0.8  12 0.042  2.6 1.1      0.81

 lower alpha upper     95% confidence boundaries
0.83 0.92 1

 Reliability if an item is dropped:
   raw_alpha std.alpha G6(smc) average_r S/N alpha se var.r med.r
Q1      0.89      0.90    0.82      0.82 9.0    0.066    NA  0.82
Q2      0.87      0.88    0.78      0.78 7.1    0.079    NA  0.78
Q3      0.87      0.90    0.81      0.81 8.7    0.071    NA  0.81

 Item statistics
    n raw.r std.r r.cor r.drop mean  sd
Q1 10  0.92  0.93  0.87   0.84  2.8 1.0
Q2 10  0.95  0.94  0.90   0.86  2.4 1.4
Q3 10  0.93  0.93  0.87   0.84  2.7 1.2

Non missing response frequency for each item
     1   2   3   4   5 miss
Q1 0.1 0.3 0.3 0.3 0.0    0
Q2 0.3 0.4 0.0 0.2 0.1    0
Q3 0.1 0.4 0.3 0.1 0.1    0
```

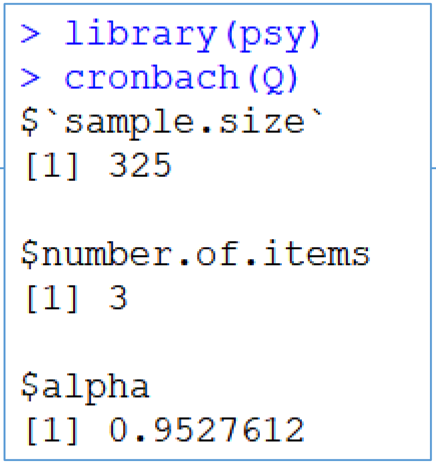
Source : sta_4_alpha.R , 6_pca.xls

신뢰도 분석

```
myd <- read_excel("work_r/source/6_pca.xls",
col_names = TRUE)

myd <- myd %>% select(2 : 16)   # 15

pairs(myd, panel=panel.smooth)

Q <- myd[, 1 : 3]   # q1 ~ q3

alpha(Q)

a <- alpha(Q)
str(a)

a$total
```



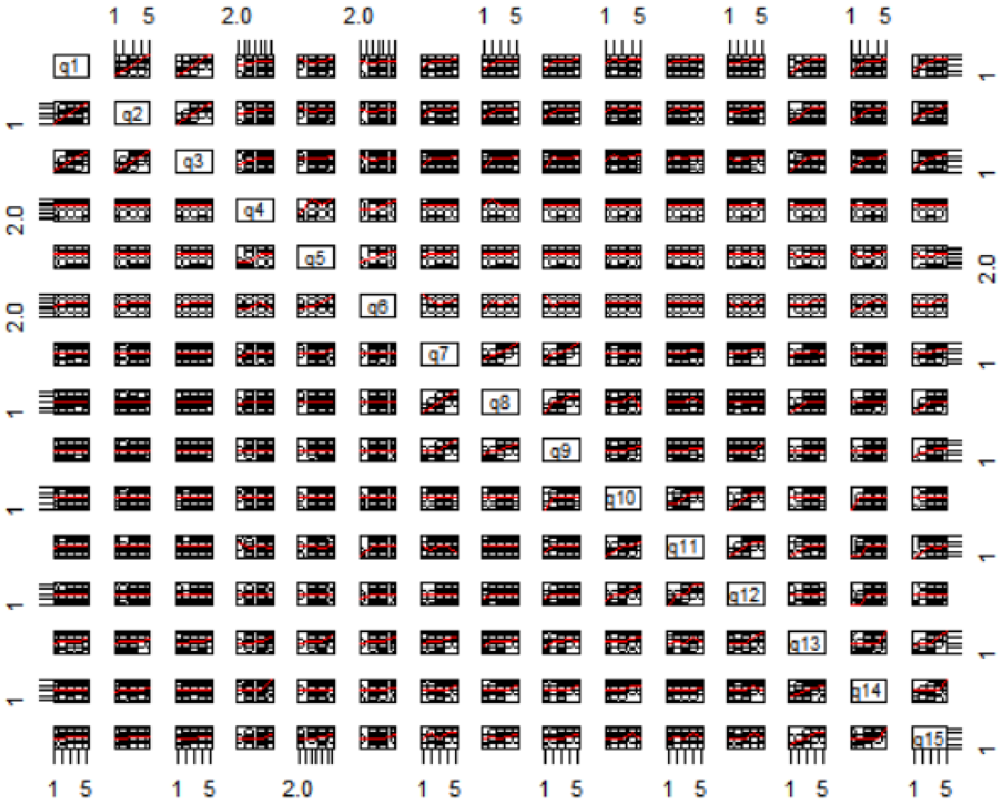
```
> alpha(Q)

Reliability analysis
Call: alpha(x = Q)

  raw_alpha std.alpha G6(smc) average_r S/N    ase mean  sd median_r
      0.95      0.95    0.93      0.87  21 0.0045  3.6 1.1     0.87

 lower alpha upper     95% confidence boundaries
0.94 0.95 0.96

 Reliability if an item is dropped:
   raw_alpha std.alpha G6(smc) average_r S/N alpha se var.r med.r
q1      0.94      0.95    0.90      0.90  17   0.0062    NA  0.90
q2      0.92      0.92    0.86      0.86  12   0.0087    NA  0.86
q3      0.93      0.93    0.87      0.87  13   0.0080    NA  0.87
```

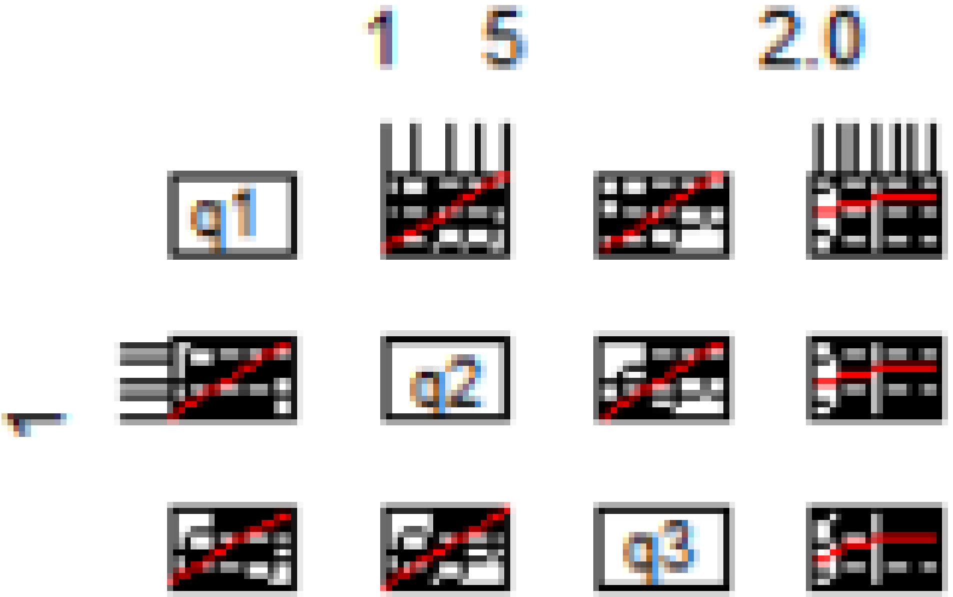


```
> library(psy)
> cronbach(Q)
$`sample.size`
[1] 325

$number.of.items
[1] 3

$alpha
[1] 0.9527612
```

Source : sta_4_alpha.R , 6_pca.xls

신뢰도 분석

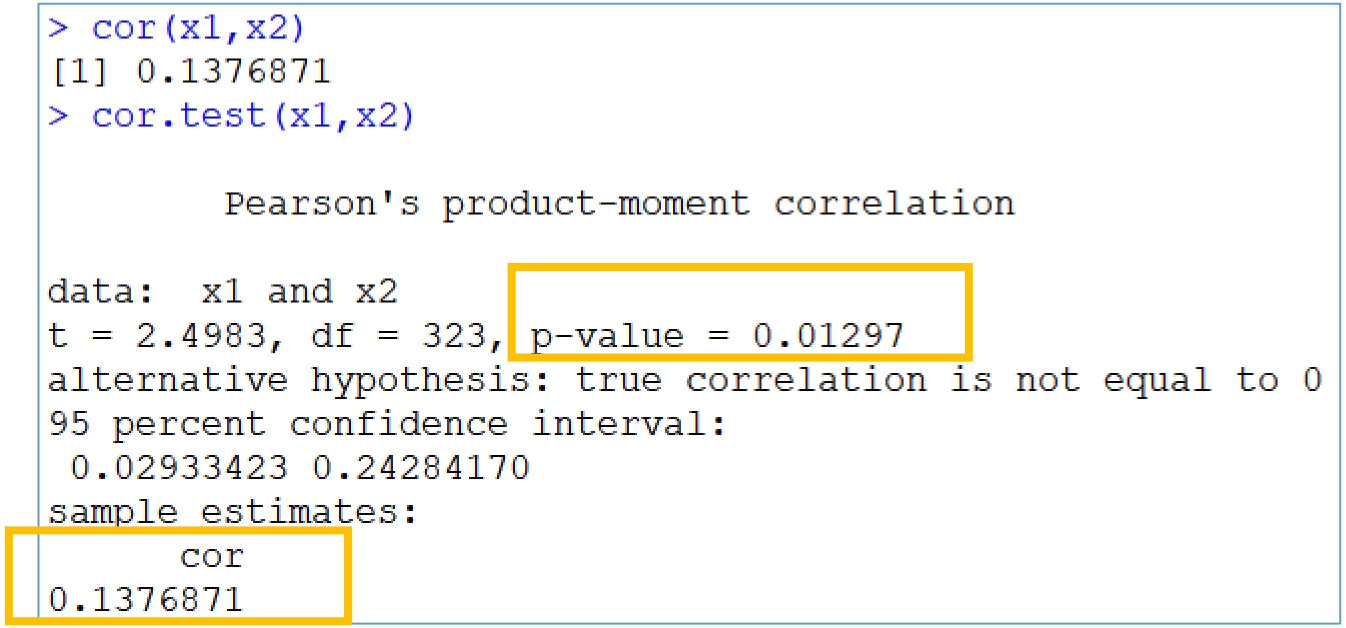




Source : sta_4_alpha.R , 6_pca.xls

상관분석 – 피어슨

```
myd <- read_excel("work_r/source/6_pca.xls",
col_names = TRUE)

myd <- myd %>% select(2 : 16)

x1 <- rowMeans(myd[, 1:3])
x2 <- rowMeans(myd[, 4:6])

cor(x1,x2)
cor.test(x1,x2)
```

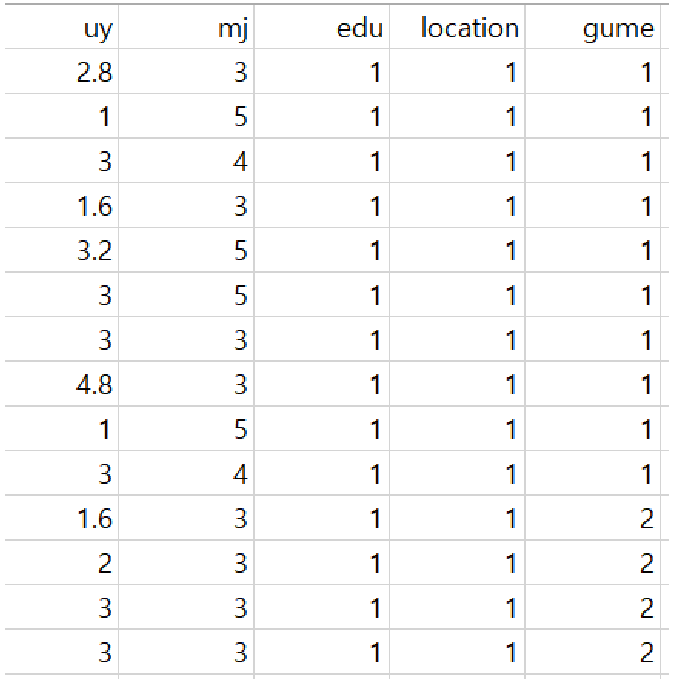


```
> cor(x1,x2)
[1] 0.1376871
> cor.test(x1,x2)

        Pearson's product-moment correlation

data:  x1 and x2
t = 2.4983, df = 323, p-value = 0.01297
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.02933423 0.24284170
sample estimates:
      cor
0.1376871
```

Source : sta_4_alpha.R , 6_pca.xls

교차분석, chi square

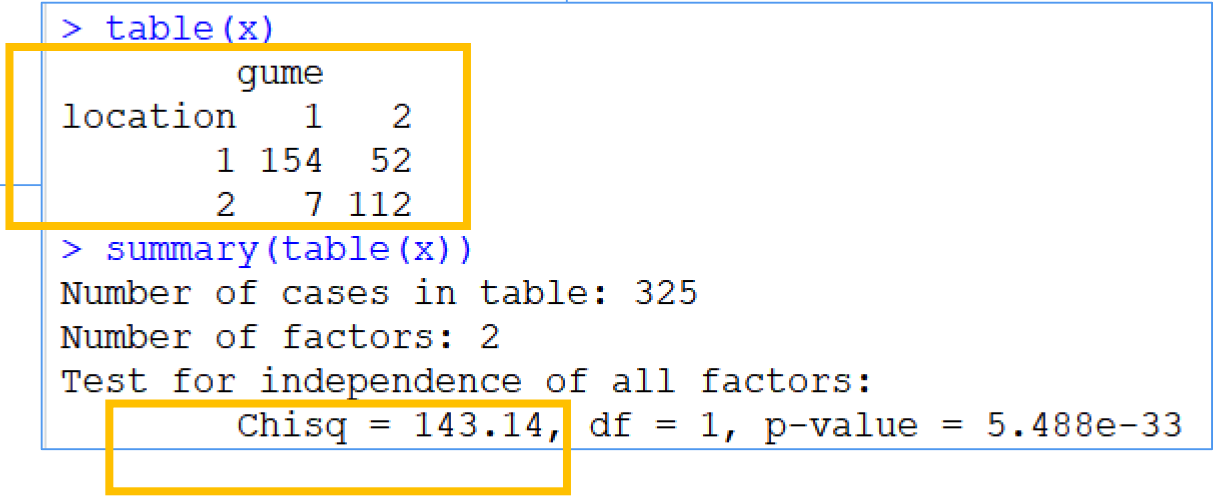
- **location 1,2 의 구매의사 비교**



uy	mj	edu	location	gume
2.8	3	1	1	1
1	5	1	1	1
3	4	1	1	1
1.6	3	1	1	1
3.2	5	1	1	1
3	5	1	1	1
3	3	1	1	1
4.8	3	1	1	1
1	5	1	1	1
3	4	1	1	1
1.6	3	1	1	2
2	3	1	1	2
3	3	1	1	2
3	3	1	1	2

```
library(readxl)
myd <-
read_excel("work_r/source/7_table_chisquare.xls",
col_names = TRUE)

str(myd)

x <- myd[c("location", "gume")]
str(x)

table(x)

summary(table(x))
```

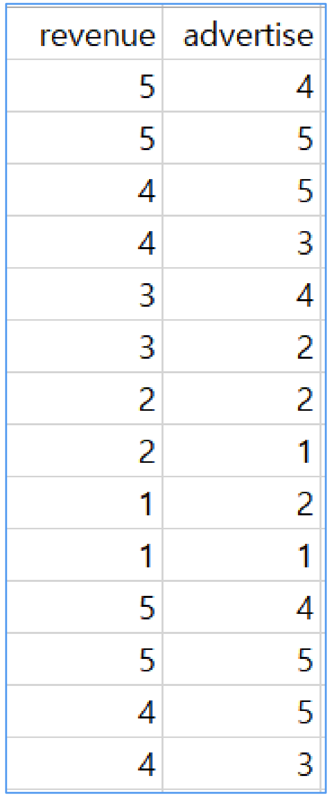
* CrossTable



```
> table(x)
        gume
location   1   2
       1 154  52
       2   7 112
> summary(table(x))
Number of cases in table: 325
Number of factors: 2
Test for independence of all factors:
        Chisq = 143.14, df = 1, p-value = 5.488e-33
```

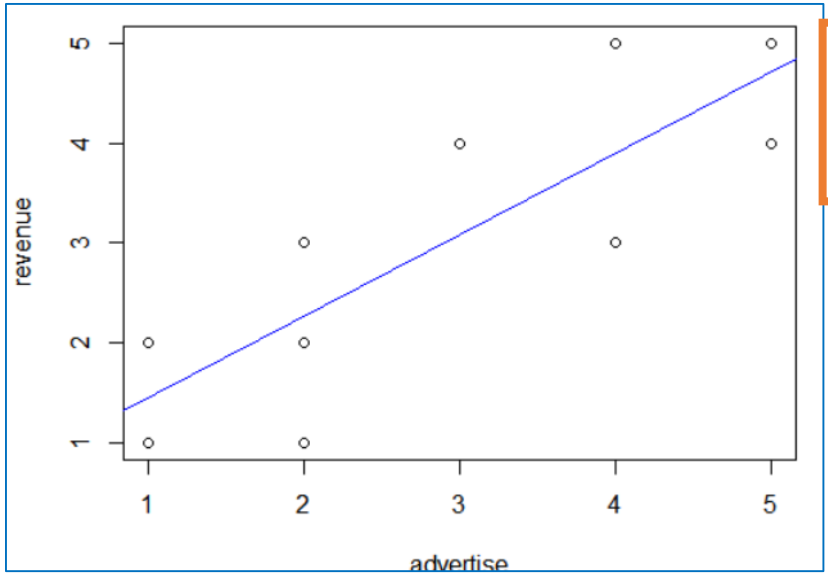
Source : sta_5_table.R , 7_table_chisquare.xls

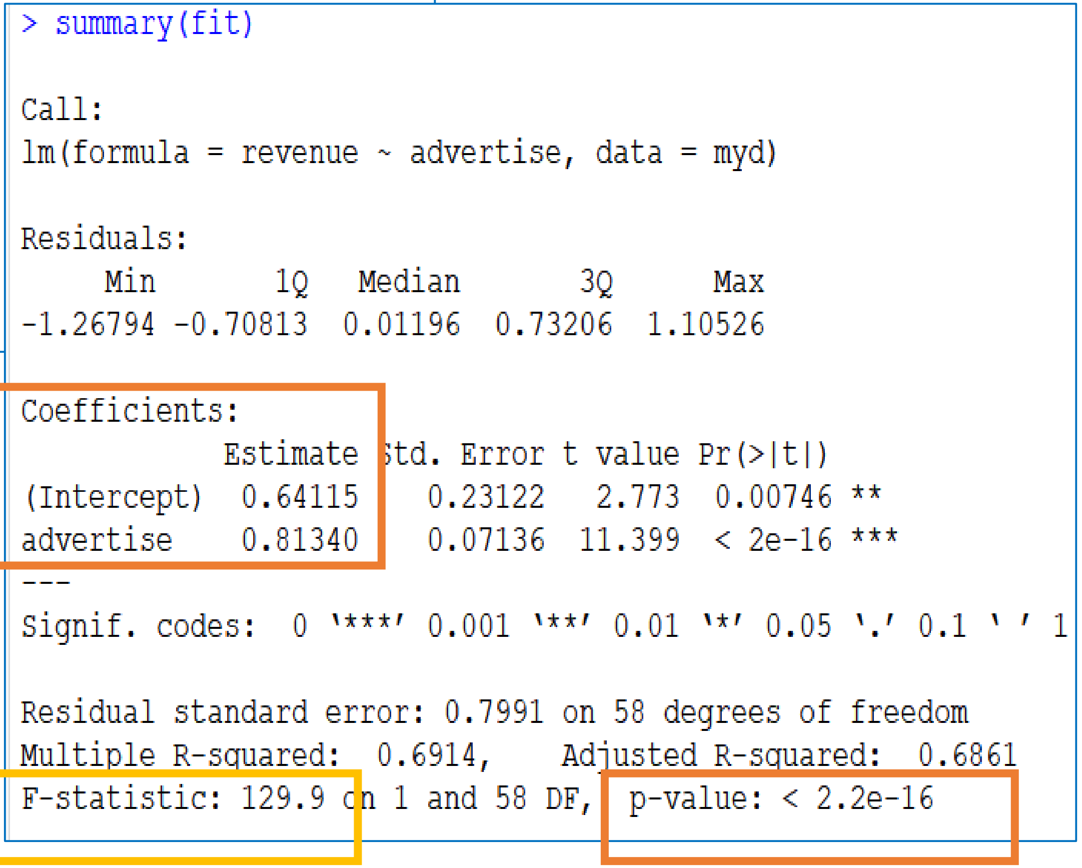
단순 회귀분석



revenue	advertise
5	4
5	5
4	5
4	3
3	4
3	2
2	2
2	1
1	2
1	1
5	4
5	5
4	5
4	3

```
library(readxl)

myd <- read_excel("work_r/source/8_simple_lr.xls", col_names = TRUE)

fit <- lm( revenue ~ advertise, data=myd)

summary(fit)

plot(revenue ~ advertise, data=myd)

abline(fit,col="blue")
```

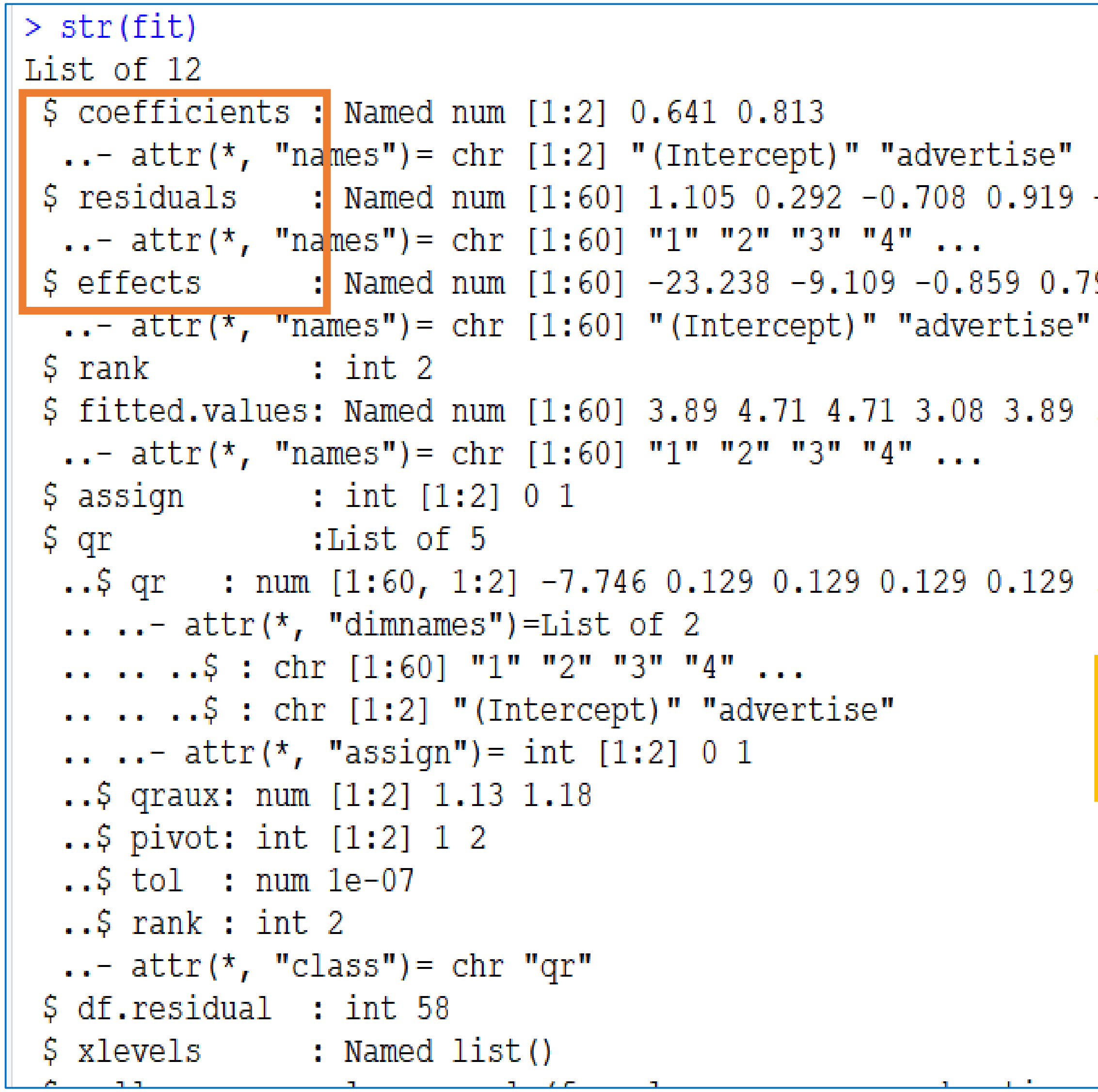




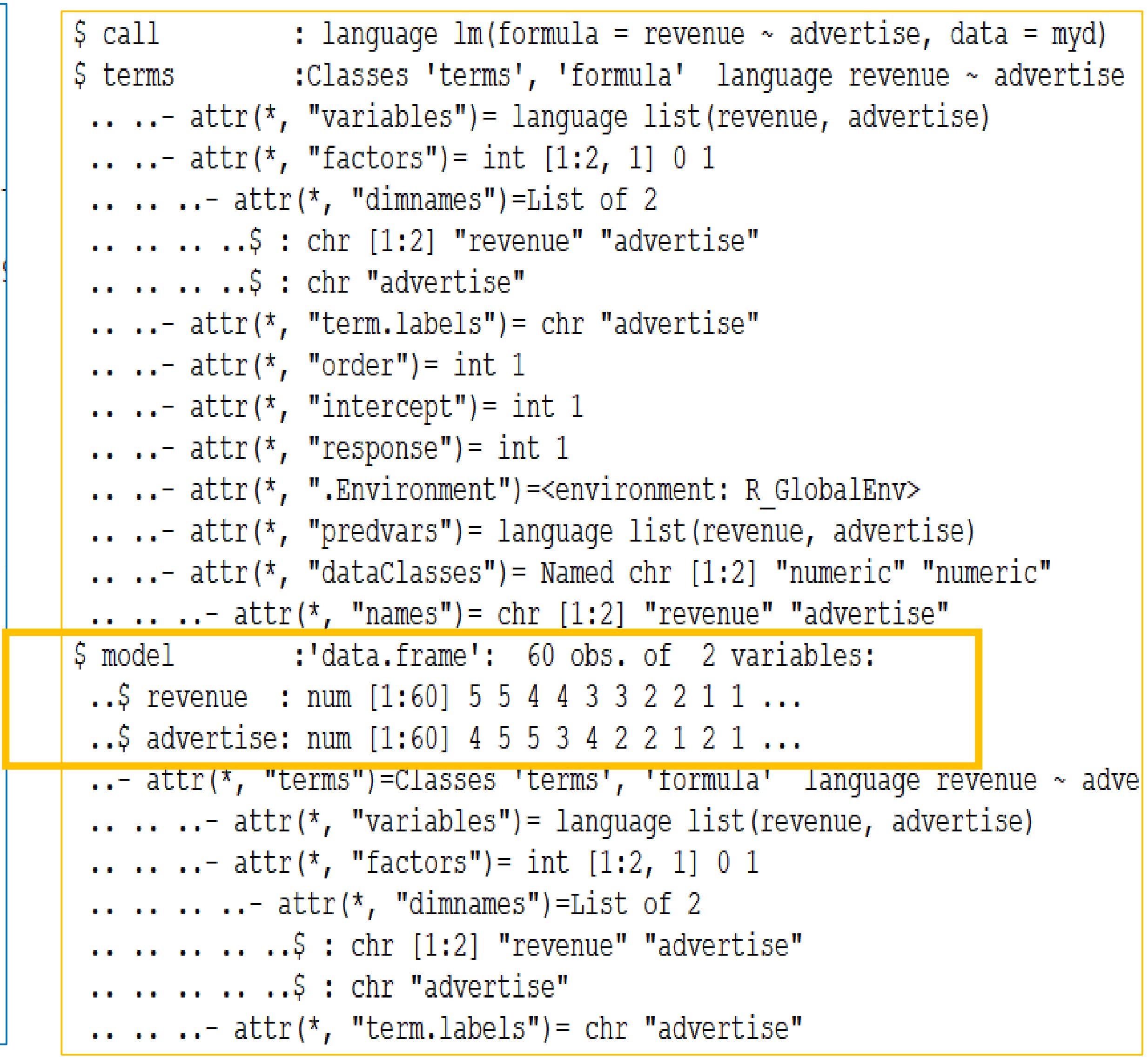
```
> summary(fit)

Call:
lm(formula = revenue ~ advertise, data = myd)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.26794 -0.70813  0.01196  0.73206  1.10526 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.64115    0.23122   2.773  0.00746 ** 
advertise    0.81340    0.07136  11.399  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7991 on 58 degrees of freedom
Multiple R-squared:  0.6914,	Adjusted R-squared:  0.6861 
F-statistic: 129.9 on 1 and 58 DF,  p-value: < 2.2e-16
```

Source : sta_6_simple_lr.R , 8_simple_lr.xls

단순 회귀분석 – fit



```
> str(fit)
List of 12
 $ coefficients : Named num [1:2] 0.641 0.813
  ..- attr(*, "names")= chr [1:2] "(Intercept)" "advertise"
 $ residuals    : Named num [1:60] 1.105 0.292 -0.708 0.919 -
  ..- attr(*, "names")= chr [1:60] "1" "2" "3" "4" ...
 $ effects      : Named num [1:60] -23.238 -9.109 -0.859 0.7
  ..- attr(*, "names")= chr [1:60] "(Intercept)" "advertise"
 $ rank         : int 2
 $ fitted.values: Named num [1:60] 3.89 4.71 4.71 3.08 3.89
  ..- attr(*, "names")= chr [1:60] "1" "2" "3" "4" ...
 $ assign       : int [1:2] 0 1
 $ qr           :List of 5
  ..$ qr   : num [1:60, 1:2] -7.746 0.129 0.129 0.129 0.129
  .. ..- attr(*, "dimnames")=List of 2
  .. .. ..$ : chr [1:60] "1" "2" "3" "4" ...
  .. .. ..$ : chr [1:2] "(Intercept)" "advertise"
  .. ..- attr(*, "assign")= int [1:2] 0 1
  ..$ qraux: num [1:2] 1.13 1.18
  ..$ pivot: int [1:2] 1 2
  ..$ tol  : num 1e-07
  ..$ rank : int 2
  ..- attr(*, "class")= chr "qr"
 $ df.residual  : int 58
 $ xlevels      : Named list()
```

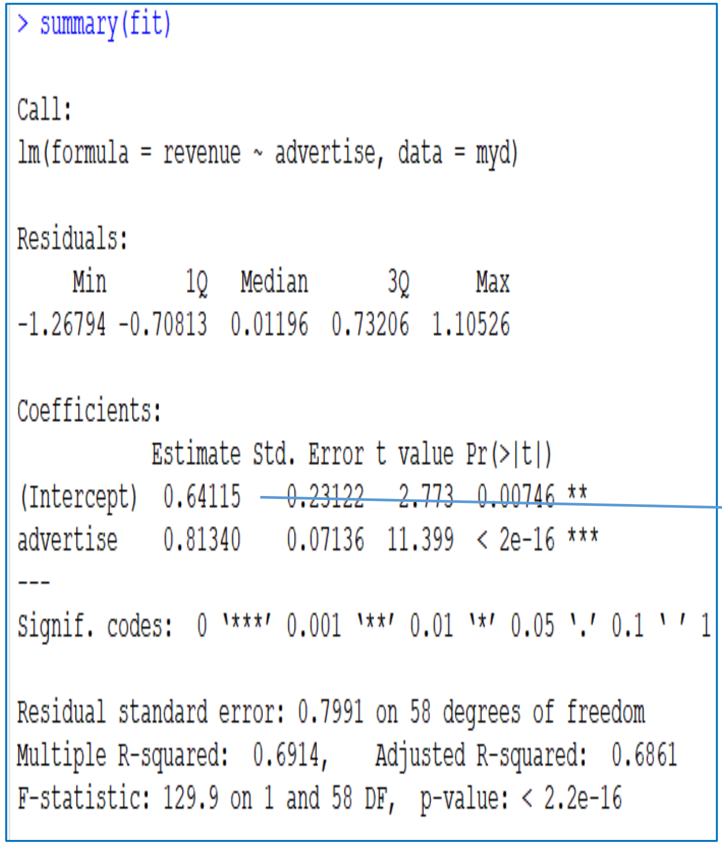


```
$ call         : language lm(formula = revenue ~ advertise, data = myd)
$ terms        :Classes 'terms', 'formula'  language revenue ~ advertise
 .. ..- attr(*, "variables")= language list(revenue, advertise)
 .. ..- attr(*, "factors")= int [1:2, 1] 0 1
 .. .. ..- attr(*, "dimnames")=List of 2
 .. .. .. ..$ : chr [1:2] "revenue" "advertise"
 .. .. .. ..$ : chr "advertise"
 .. ..- attr(*, "term.labels")= chr "advertise"
 .. ..- attr(*, "order")= int 1
 .. ..- attr(*, "intercept")= int 1
 .. ..- attr(*, "response")= int 1
 .. ..- attr(*, ".Environment")=<environment: R_GlobalEnv>
 .. ..- attr(*, "predvars")= language list(revenue, advertise)
 .. ..- attr(*, "dataClasses")= Named chr [1:2] "numeric" "numeric"
 .. .. ..- attr(*, "names")= chr [1:2] "revenue" "advertise"
$ model        :'data.frame':   60 obs. of  2 variables:
 ..$ revenue  : num [1:60] 5 5 4 4 3 3 2 2 1 1 ...
 ..$ advertise: num [1:60] 4 5 5 3 4 2 2 1 2 1 ...
 ..- attr(*, "terms")=Classes 'terms', 'formula'  language revenue ~ adve
 .. .. ..- attr(*, "variables")= language list(revenue, advertise)
 .. .. ..- attr(*, "factors")= int [1:2, 1] 0 1
 .. .. .. ..- attr(*, "dimnames")=List of 2
 .. .. .. .. ..$ : chr [1:2] "revenue" "advertise"
 .. .. .. .. ..$ : chr "advertise"
 .. .. ..- attr(*, "term.labels")= chr "advertise"
```

Source : sta_6_simple_lr.R , 8_simple_lr.xls

회귀분석 – 회귀분석 통계량

- **회귀식 분산분석 : p < .05 유의하다**



```
> summary(fit)

Call:
lm(formula = revenue ~ advertise, data = myd)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.26794 -0.70813  0.01196  0.73206  1.10526 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.64115    0.23122   2.773  0.00746 ** 
advertise    0.81340    0.07136  11.399  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7991 on 58 degrees of freedom
Multiple R-squared:  0.6914,	Adjusted R-squared:  0.6861 
F-statistic: 129.9 on 1 and 58 DF,  p-value: < 2.2e-16
```

분산분석 : anova (fit)

계수 : fit $ coef

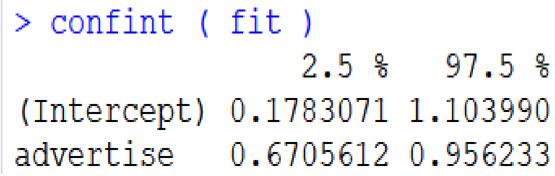
잔차제곱합 : deviance(fit)

직교효과들로 이루어진 벡터 : effects (fit)

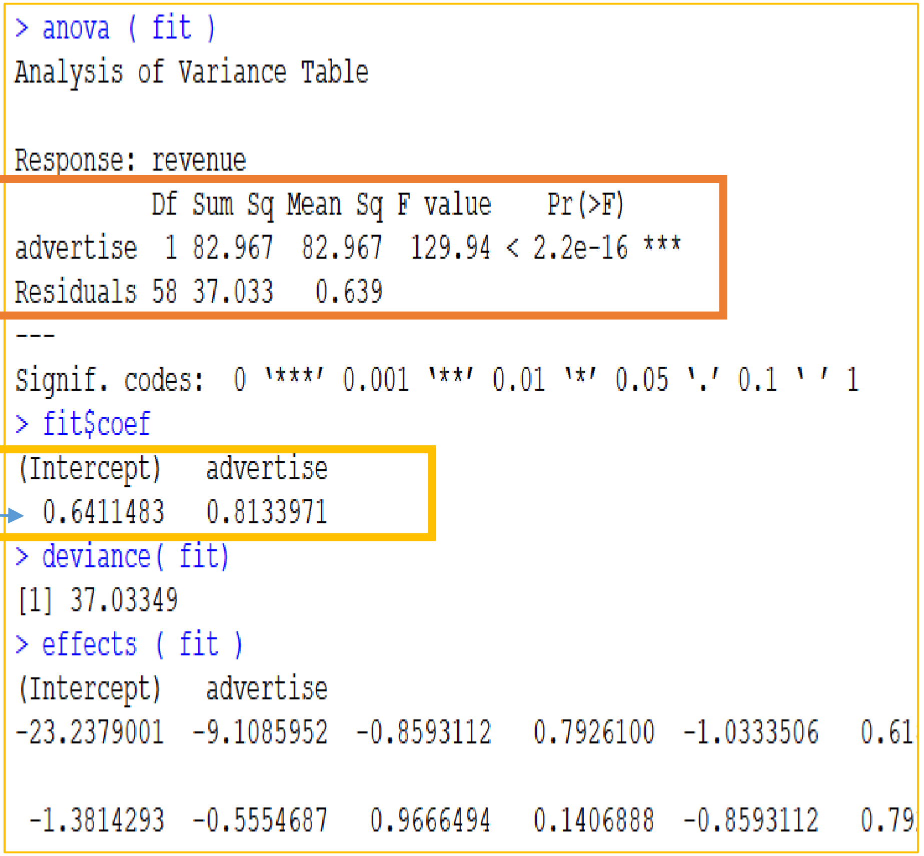
적합된 y값으로 이루어진 벡터 : fitted (fit)

주 매개변수들의 분산-공분산 행렬 : vcov (fit)

신뢰구간 : confint (fit)



```
> confint ( fit )
                2.5 %   97.5 %
(Intercept) 0.1783071 1.103990
advertise   0.6705612 0.956233
```

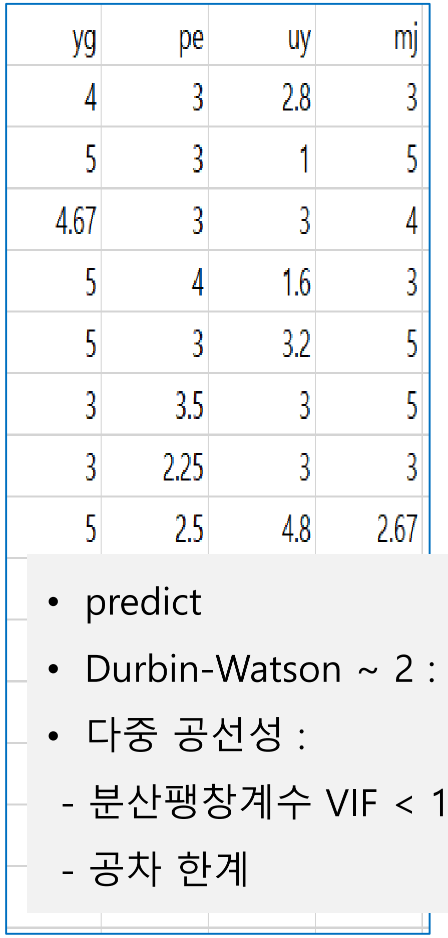


```
> anova ( fit )
Analysis of Variance Table

Response: revenue
          Df Sum Sq Mean Sq F value    Pr(>F)    
advertise  1 82.967  82.967  129.94 < 2.2e-16 ***
Residuals 58 37.033   0.639                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> fit$coef
(Intercept)   advertise 
  0.6411483   0.8133971 
> deviance( fit)
[1] 37.03349
> effects ( fit )
(Intercept)   advertise                                                 
-23.2379001  -9.1085952  -0.8593112   0.7926100  -1.0333506   0.61

 -1.3814293  -0.5554687   0.9666494   0.1406888  -0.8593112   0.79
```

Source : sta_6_simple_lr.R , 9_multivar_lr.xlsx

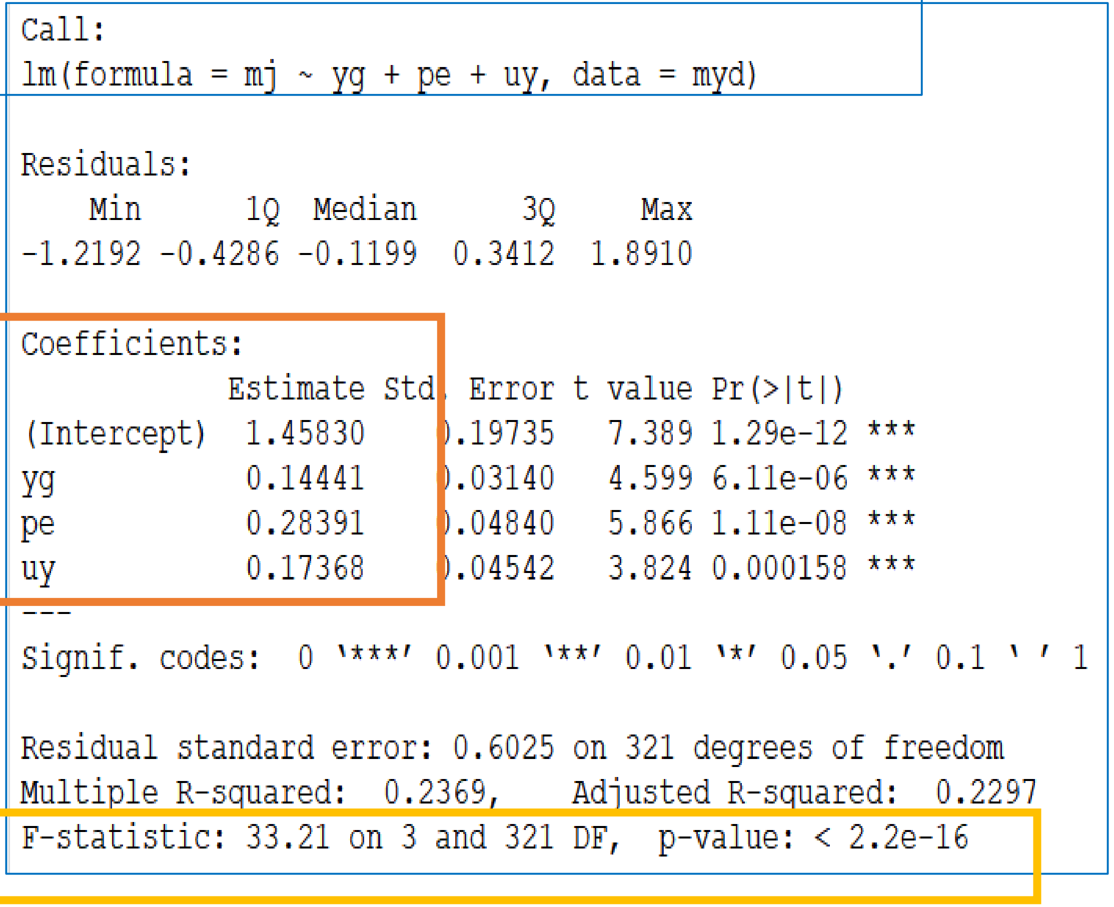
다중 회귀분석

- **외관, 편의, 유용성과 만족감 관계**



yg	pe	uy	mj
4	3	2.8	3
5	3	1	5
4.67	3	3	4
5	4	1.6	3
5	3	3.2	5
3	3.5	3	5
3	2.25	3	3
5	2.5	4.8	2.67

```
myd <- read_excel("work_r/source/9_multivar_lr.xlsx", col_names = TRUE)

fit <- lm( mj ~ yg + pe + uy, data=myd)

summary(fit)
```



```
Call:
lm(formula = mj ~ yg + pe + uy, data = myd)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.2192 -0.4286 -0.1199  0.3412  1.8910 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.45830    0.19735   7.389 1.29e-12 ***
yg           0.14441    0.03140   4.599 6.11e-06 ***
pe           0.28391    0.04840   5.866 1.11e-08 ***
uy           0.17368    0.04542   3.824 0.000158 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6025 on 321 degrees of freedom
Multiple R-squared:  0.2369,	Adjusted R-squared:  0.2297 
F-statistic: 33.21 on 3 and 321 DF,  p-value: < 2.2e-16
```

- predict
- Durbin-Watson ~ 2 : 잔차의 독립성
- 다중 공선성 :

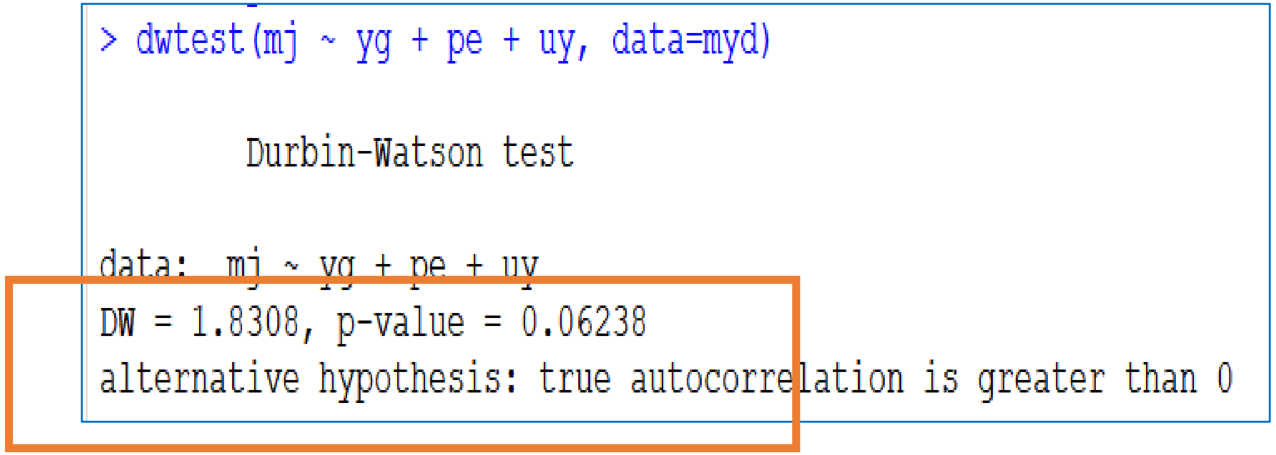
  - 분산팽창계수 VIF < 10

  - 공차 한계

Source : sta_6_simple_lr.R , 9_multivar_lr.xlsx

다중 회귀분석 - Durbin-Watson / p-p / scatter plot

```
library(lmtest)

dwtest(mj ~ yg + pe + uy, data=myd)
```



```
> dwtest(mj ~ yg + pe + uy, data=myd)

        Durbin-Watson test

data:  mj ~ yg + pe + uy
DW = 1.8308, p-value = 0.06238
alternative hypothesis: true autocorrelation is greater than 0
```

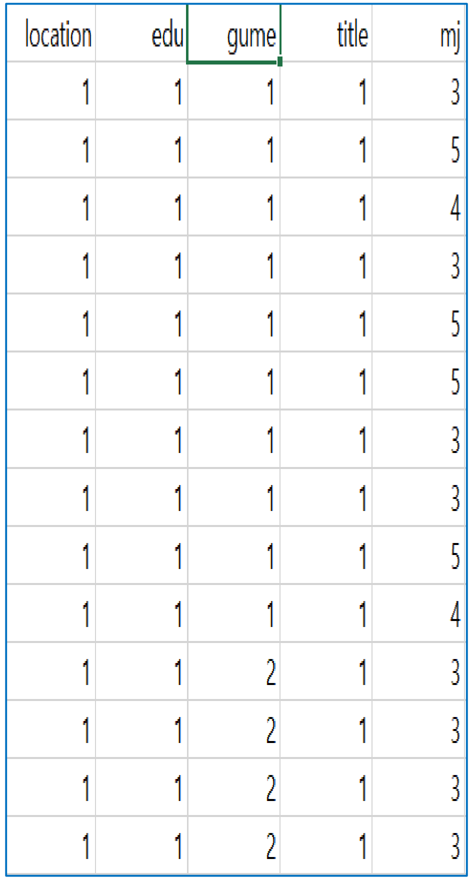
Source : sta_6_simple_lr.R , 9_multivar_lr.xlsx

단계적 회귀분석 -

Source : sta_6_simple_lr.R , 9_multivar_lr.xlsx

logistic

- **지역, 학력, 구매의사 (1 구매, 2 안함)**



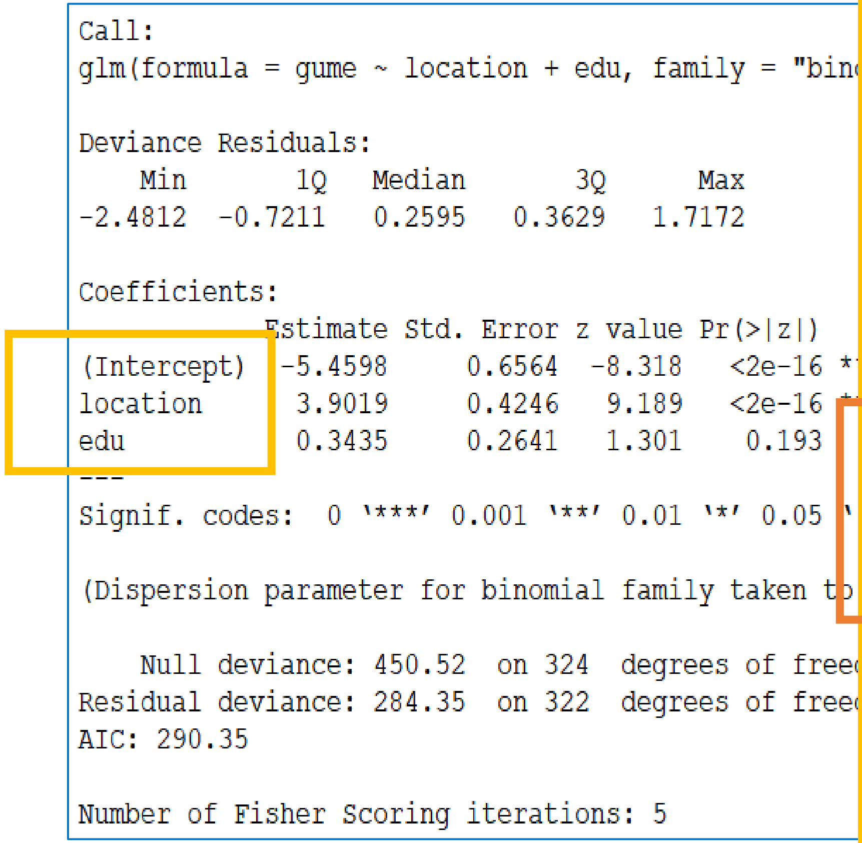
location	edu	gume	title	mj
1	1	1	1	3
1	1	1	1	5
1	1	1	1	4
1	1	1	1	3
1	1	1	1	5
1	1	1	1	5
1	1	1	1	3
1	1	1	1	3
1	1	1	1	5
1	1	1	1	4
1	1	2	1	3
1	1	2	1	3
1	1	2	1	3
1	1	2	1	3

```
library(readxl)

myd <- read_excel("work_r/source/10_logistic.xls", col_names = TRUE)

# data cleansing : norminal
library(dplyr)

myd2 <- myd %>% select(gume, location, edu)
str(myd2)

myd2$location <- as.factor(myd$location)
myd2$edu <- as.factor(myd$edu)
myd2$gume <- as.factor(myd$gume)
```

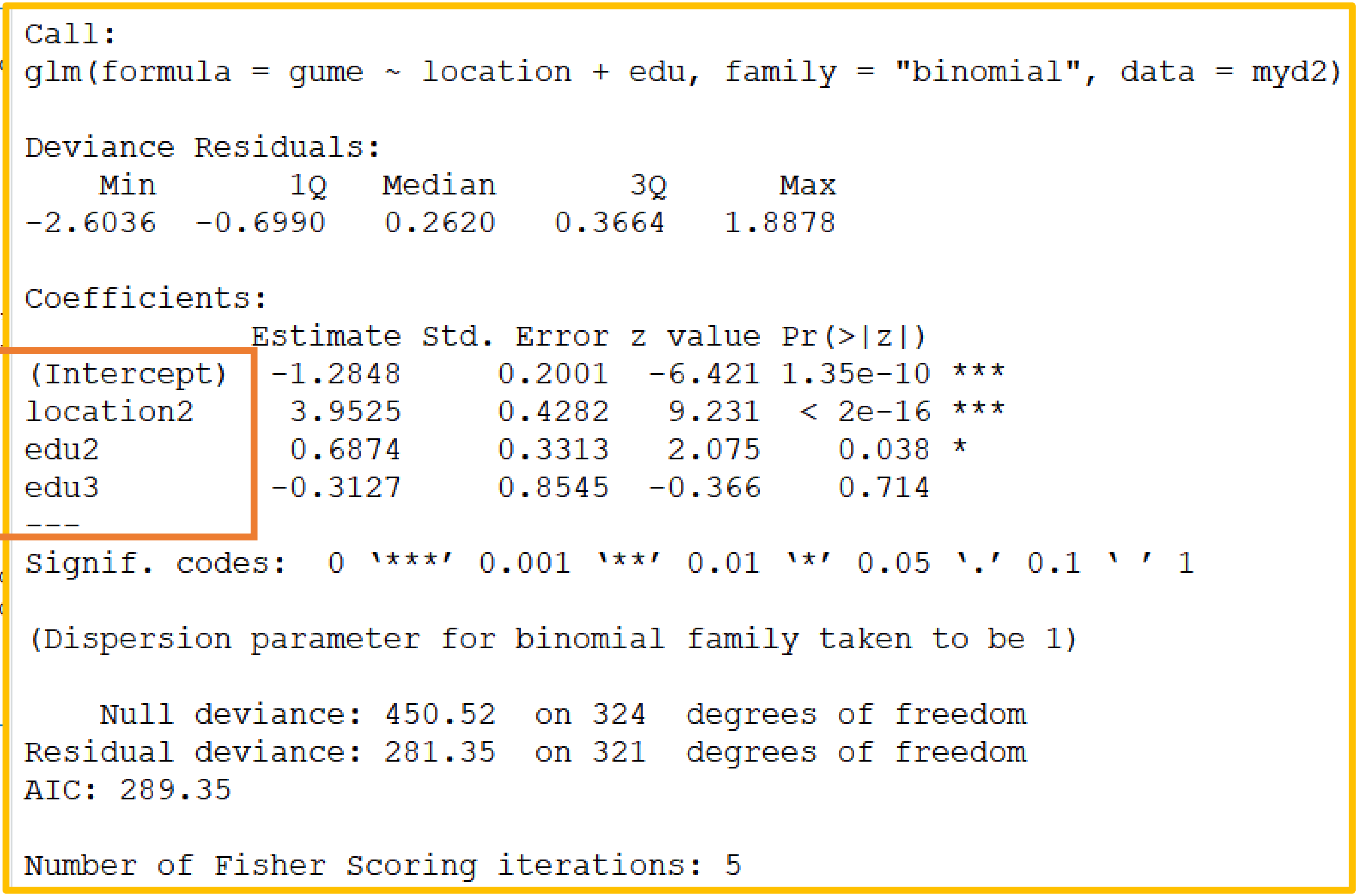
```
head(myd2)

# gume 0, 1

myd2$gume <- ifelse (myd2$gume == 1 , 0, 1)
myd2$gume <- as.factor(myd2$gume)

str(myd2)
summary(myd2)


fit <- glm( gume ~ location + edu, data=myd2,
family = "binomial")

str(fit)

summary(fit)
```

Source : sta_7_logistic.R , 10_logistic.xls

logistic – 더미변수 / 범주형



```
Call:
glm(formula = gume ~ location + edu, family = "bin

Deviance Residuals:
    Min       1Q   Median       3Q      Max  
-2.4812  -0.7211   0.2595   0.3629   1.7172  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.4598     0.6564  -8.318   <2e-16 *
location      3.9019     0.4246   9.189   <2e-16 *
edu           0.3435     0.2641   1.301    0.193
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '

(Dispersion parameter for binomial family taken to

    Null deviance: 450.52  on 324  degrees of freed
Residual deviance: 284.35  on 322  degrees of freed
AIC: 290.35

Number of Fisher Scoring iterations: 5
```



```
Call:
glm(formula = gume ~ location + edu, family = "binomial", data = myd2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max  
-2.6036  -0.6990   0.2620   0.3664   1.8878  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)  -1.2848     0.2001  -6.421 1.35e-10 ***
location2     3.9525     0.4282   9.231  < 2e-16 ***
edu2          0.6874     0.3313   2.075    0.038 *  
edu3         -0.3127     0.8545  -0.366    0.714    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 450.52  on 324  degrees of freedom
Residual deviance: 281.35  on 321  degrees of freedom
AIC: 289.35

Number of Fisher Scoring iterations: 5
```

Source : sta_7_logistic.R , 10_logistic.xls

군집

Source : sta_1_t_test , 2_pttest.xls

분산 분석

Source : sta_1_t_test , 2_pttest.xls

독립성 검정 (chi square test)

Source :

연관분석 상관분석

Source :

판별분석

Source :