Analysis of Twins

Team member: Kaini Liu, Chunmei Gao, Haoqi Sun, Hongtao Li

1. Background(Introduction)

In this project, we apply various tests to analyze the National Merit Twin Study dataset (J. C. Loehlin & R. C. Nichols, 1976). One of the main purposes of this study was to identify the effect of sex, type of twins, mother education level, father education level, and family income towards intelligence performance among members of 839 late-adolescent twin pairs. The results we expected to get will help explain the effect of growth environment and natural trait toward intelligence performance of twins. Understanding of these relationships could be important for educators and twins' parents.

2. Data Exploration and Visualization

The data file contains both characters and numeric variables. There are totally 1678 observations. Each case is an individual; and they are actually paired twins, which means that there are 839 pairs of twins. In the dataset, each of the twins has the same sex, moed, faed and faminc. It means each pair of twins has family environment. The scores of intelligence have about the same range.

Variable	Description	Value
Sex	Gender	1: male; 2: female
Zygosity	Type of twins	1: identical; 2: fraternal
Moed	Mother education level	1: failed to complete 8th grade; 2: part high school 3: high school graduate; 4: part college 5: college graduate; 6: graduate degree
Faed	Father education level	Same to mother education
Faminc	Annual family income level	1: < $5000; 2: $5000 ~ $7499; 3: $7500 ~ $9999; 4: $10000 ~ $14999; 5: $15000 ~ $19999; 6: $20000 ~ $24999; 7: >= $25000
English	English	Ranging from 0-34
Math	Mathematics	Ranging from 0-34
Socsci	Social Science	Ranging from 0-34
Natsci	Natural Science	Ranging from 0-34
Vocab	Vocabulary	Ranging from 0-34

Zygosity is the degree of similarity of the alleles for a trait in an organism. In our dataset there are two types of zygosity: identical and fraternal. Two separate sperm, resulting in fraternal twins, while one single egg is fertilized and then divides into two separate embryos resulting in identical twins, fertilize two separate eggs.

There are totally 38 moed, 48 faed and 124 faminc missing in the dataset. The missing values are imputed with a randomly generated value from original distribution.

There are few observations' squared generalized distances slightly larger than chis-square distance. No evidence of mistake-filling or extreme outliers, we decided to use the original data to continue our analysis.

3. Data Analysis and Result

The analysis is based on various hypothesis tests, which are generally based on following assumptions:

1.The samples are random.

2.The random samples from different populations are independent.
3.All populations are multivariate normal and have a common covariance.
Since our original data has 1678 observations which are 839 pairs twins, two observations may related if they are twins. Therefore, we first transformed the data by calculating the mean of two individual in each pair of twins to make sure that all observations are random and independent.
We assumed that given the English, Math, Social Science, Natural Science and Vocabulary score are multivariate normal. To verify this normality assumption, we plotted normal Q-Q plots for five scores, shown in Figure 3-1. These Q- Q plots indicate that the marginal distributions of the five mean score covariates are approximately normal. The multivariate normal distributional assumption for these covariates is exam by Chi-Square plot. Therefore, these indispensable results made our assumption hold.

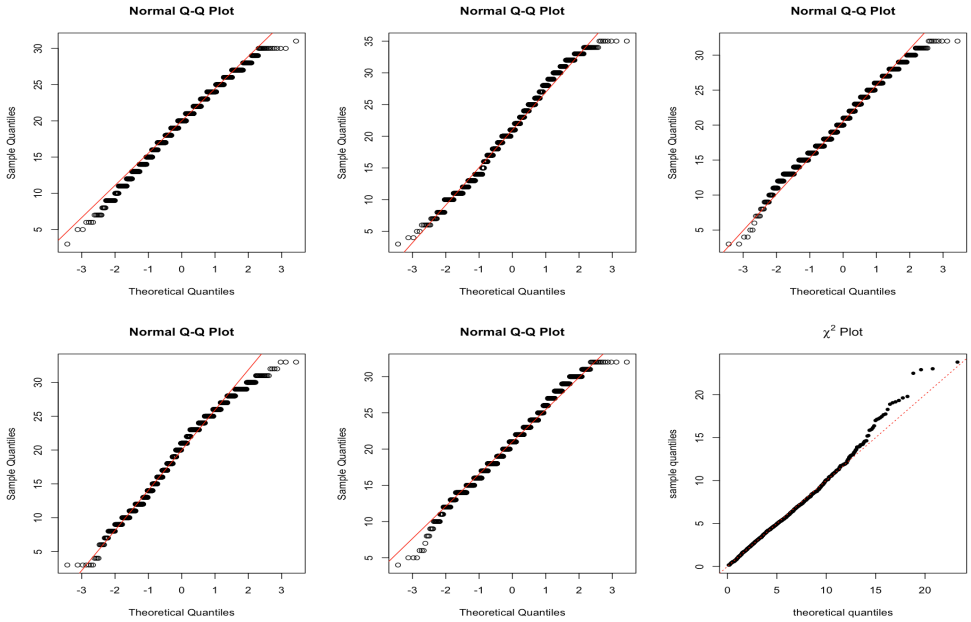

Figure 3-1. Q-Q Plots for each score and Chi-Square Plot for multivariate normal (last)

To test the equality of covariance matrices, we conducted the Box's Test for the equality of covariance matrices of male and female twins ($\Sigma_{male} = \Sigma_{female}$), also those of identical and fraternal twins ($\Sigma_{identical} = \Sigma_{fraternal}$).
Covariance matrices in different gender group are the same. For identical and fraternal twins, covariance matrices are different. By knowing the equality of covariance matrices, we will be able to choose the eligible tests for our following analysis.

Sex Differences in Mean Score

First, we were interested in whether there are intelligence differences between male and female. So, we grouped our transformed data by male and female. Then, we carried out a test that compare mean vector of male and female, each vector contain five mean score of five courses. The null hypothesis of this test is $H_0: \mu_{female} = \mu_{male}$. The square of the statistical distance, T^2 Test statistics (313.1) is greater than critical value (11.1) means that there exist significant difference between male score and female group mean score.
Moreover, we desire to know which group perform better in each subject. This can be seen by output Bonferroni 95% confidence interval ($\mu_{female} - \mu_{male}$) for each subject in Table 3-1. Confidence interval of vocabulary mean score contains 0. Except that, female group gets slightly higher score in English. Male would perform better in math, social science and natural science.

Subject	95% Confidence Interval

English	(0.32, 1.85)
Math	(-4.48, -2.52)
Social Science	(-1.94, -0.35)
Natural Science	(-3.42 -1.60)
Vocabulary	(-0.92, 0.71)

Table 3-1. Bonferroni 95% CI for $(\mu_{female} - \mu_{male})$

Zygosity Differences

This test is also focus on the differences between the mean vectors of two groups: identical twins and fraternal twins. Similar with the sex group test, the null hypothesis is $\mu_{\text{Identical}} = \mu_{\text{fraternal}}$. This null hypothesis was not being rejected due to the test statistics (8.3) is smaller than the critical value (11.8). This indicates that the mean score of each course are almost the same between identical twins and fraternal twins.

This result also corresponds to what we got by calculating the Bonferroni 95% confidence intervals. All the confidence intervals of five subjects contain 0, which imply that the two type twins have no significant intelligent differences.

Family Background Effect

Now we begin to consider the effect of family environment: family income and parents' education level. We are interested in the relationship between twins and their family. To solve these, we carried out MANOVA test to examine the three effects of Income, Father Education and Mother Education separately. Results show that all these three family environment variables showed significant effect towards twins' intelligence.

Moreover, we also interested in how these variables effects the score of twins. So, simultaneous confidence intervals for family effects were calculated.

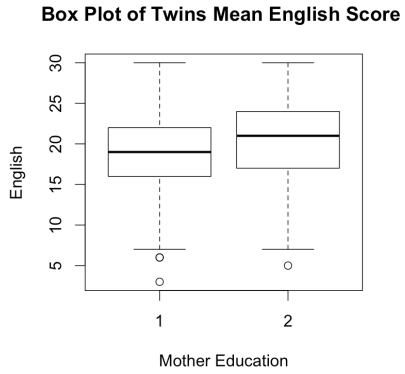

Figure 3-2 Box plot of twins mean English score

For example, for studying the effect of mother education, the score of twins was divided into two groups based on six level of mother education level. Then, the score differences of two mother education groups were calculated ($\mu_{\text{lower group}} - \mu_{\text{higher group}}$). All the confidence intervals are below 0, which means higher mother education level yield to higher mean score in twins. Similarly, the high father education and family income also have positive effect to twin pairs' scores.

Interaction Effect

From the result above we can see that, except zygosity, all other factor could influence the five mean score. Interaction term among each pair of factor are not significant to five scores in MANOVA test. In other word, two or more factors are not interacted with each other give the impact towards five score.

Genes Effect in Absolute score difference

We have known that there is no difference between identical twins and fraternal twins' mean score. However, we are also curious about whether the gap between two individual of fraternal twins is larger than the difference between identical twins since the genes are more different in fraternal twins than identical twins. The more similar two people in genes, the more similar in their abilities, regardless of the other influential factors. If there were no significant difference in absolute difference between two types of twins, the genes effect would yield to the environmental effect. To convince our hypothesis, first of all, the absolute difference score between two individuals who are pair of twins are calculated.

We assume identical twins which genes are much more similar in genes than fraternal twins. For the environmental factor level, every pairs of twins share the same nurture.

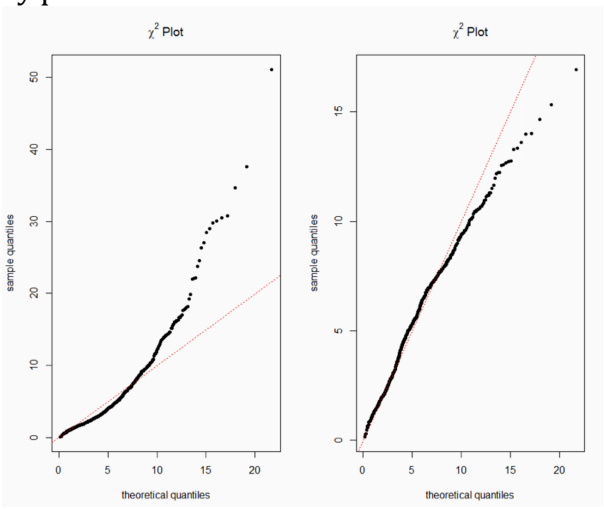

Figure 3-3. Chi-Square plot for absolute score difference and transformed data

Firstly, before testing the equality of absolute score difference, multivariate normal distributional assumption should be satisfied. Chi-square plot (left) of the absolute difference of scores data is given; the multivariate normality is not quite good. Data transform is inevitable here. Figure 3-3 (right) is the chi-square plot of transformed data after takes natural logarithm transformation. In order to avoid the infinity number, each value are added 1 before take the logarithm transformation. After the transformation, the dots in plot follow the straight line much better.

The covariate matrix equality is tested. The T^2 (102.3) is larger than the critical value (11.8), which tells us that the difference in absolute difference scores in two types of twins is significantly different. The difference in identical twins is significantly smaller than that between fraternal twins, which implies the positive effect of genes. The conclusion based on the reasoning before can be draw that the gene does affect the intelligence of individuals.

Table 3-2. The comparison of deviation within two types of twins in test scores

	Score difference between identical twins(N=509 pairs)	Score difference between fraternal twins(N=330 pairs)	t-test p-value
English	2.567780	3.433333	1.401e-07***
Math	3.418468	4.954545	5.154e-09***
Social science	2.601179	3.769697	4.272e-10***
Natural science	3.434185	4.145455	0.001972**

vocabulary	1.982318	2.996970	1.658e-10***

*df1=5,df2=833,p<0.001****

Within each pair of fraternal twins, the expected score gap is larger than each pair identical twins. However, the mean score in fraternal twins pairs has no difference compare to identical twins pairs.

Interaction Effect in absolute difference

It is also interesting to figure out that if there exists in interaction between sex and zygosity. That is to say, if there are two different patterns of genes effect in female and male.

To confirm the hypothesis, two-factor MANOVA are carried out. We obtained a significant interaction effect with p-value 0.009(<0.05). Therefore, we reject null hypothesis that there's no interaction effect between sex and zygosity, and draw a conclusion that the difference in female and male identical twins is significantly different from that in female and male fraternal twins.

Each factor only has two levels. The absolute differences of scores are shown below, as the mean vectors, which we are interested in.

Mean Vector	Female	Male
Identical	(2.3, 3.4, 2.5, 3.6, 1.9)	(2.9, 3.4, 2.8, 3.3, 2.1)
Fraternal	(3.6, 5.0, 3.5, 4.2, 2.8)	(3.2, 4.9, 4.2, 4.1, 3.3)

Table 3-3. Mean Vectors for Zygosity and Sex interaction

To visualize the result, we use the English score for example to make an interaction plot. In the testing, the five scores are used to be a 5 dimensional matrix.

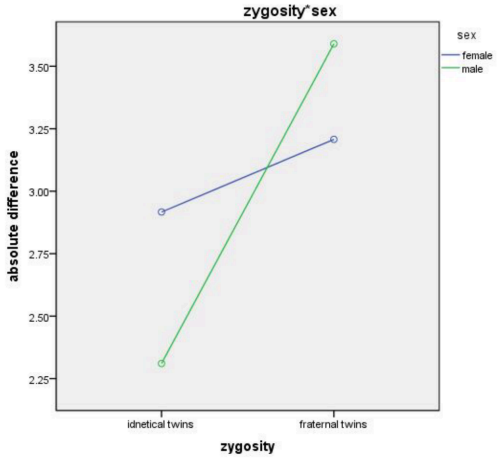

Figure 3-4. Expected absolute difference score with interaction

For the English scores in the table, both female (2.3) and male (2.9) have a higher similarity scores for identical twins, which has been mentioned in the last test. However, the slope is different in female and male lines, which implies a different pattern in two gender groups. The difference in male twins is larger than the female twins.

Twins Effect

We also test the twin effect, which refers to the difference of variance of within twins groups and variance of between twins groups. The within groups variance is significantly different from the between groups

variance which means scores' deviation of twins is less compare with non-twins. Then we divided the individual into two types of zygosity groups. In identical twins groups, scores' deviation in twins is also less than non-twins. While the fraternal group twins' effect are not significant. So the twins' effect is significant in identical twins but not in fraternal twins.

Principal Components Analysis (PCA)

There may exist some kind of relationship between each pair of scores. For example, from the common sense we could usually observe that a person with a good sense of the solving mathematical problems may holds a strong learning ability of nature science. So is that common sense true? One way is to detect the relationships between the 5 scores is to get their correlations and analyze them. From the correlation matrix plot, we could see that the largest correlation is 0.77, which is from the pair of vocabulary and social science that means if a child has a large amount of vocabulary, then she or he also would gain an ability to learn social science well. The correlations are positive with each other in relatively high level. It follows the common trend that the student performs good in one subject will also perform good in the other subject since they all represent a sense of understanding and learning.

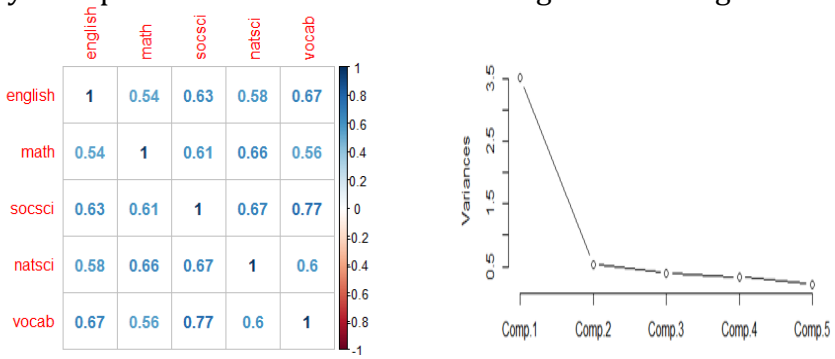

Figure 3-5 (a) Correlation Plot (b) Scree Plot

For the relationship of each subject, we want to form new variables that represent the five scores with fewer dimensions. A principal component analysis is concerned with explaining the variance-covariance structure of a set of variables through a few linear combinations of these variables. First principal component explains 70.3% of the total sample variance. The first two principal components, cumulatively, explain 81.1% of the total sample variance. Based on elbow bend in scree plot and proportion variance explained by components, sample variation is summarized very well by the first two principal components.

Given the component coefficients, the first principal component appears to be evenly weighted difference between different scores, while the second principal component appears to be a weighted mainly sum of the math and nature science.

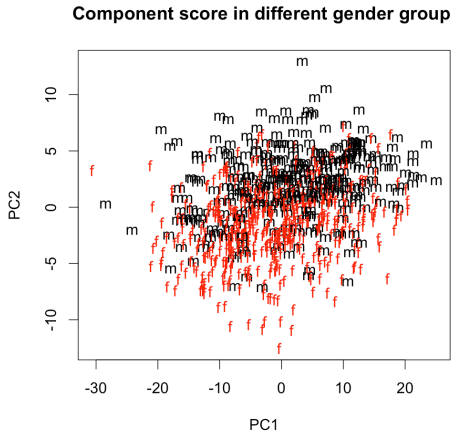

Figure 3-6. Biplot of PC1 and PC2

From Figure 3-6, the male and female samples are obviously presented. This plot shows that the calculated PCA could distinguish the gender of twins.

3. Conclusions

Based on our analysis above, conclusions of the different factors' impact on the intelligence performance among twins are given below:

First, we found that there exists significant difference in the intelligence between male and female. Moreover, we could find that males are much better at math and sciences in general, while females have a good sense on English. It meets the common sense that boys are good at logical thinking and girls have a more sensitive sense of learning language or literature. By using the same way, we find that zygosity does not influence the mean score.

As for the family background effect to the scores of twins, we got the results that these three factors are all have significant effect towards twins' intelligence. Twins in highly educated families or families with good conditions would obtain better intelligence performance.

Within each pair of fraternal twins, the expected score gap is larger than each pair identical twins. However, the mean score in fraternal twins pairs has no difference compare to identical twins pairs.

Interaction of sex and zygosity also affect scores. The gap between male twins fraternal group and identical group is larger than the female twins difference, which implies that female twins tend to be less divergent than male twin group.

Twins have similar performance compare to non-twins.

In the last, we formed two new variables which are linear combination of 5 scores variables. These two components can be used to differentiate gender groups.

Reference

Data source: National Merit Twin Study: Documentation - see Loehlin, J.C. & Nichols, R.C. (1976). Genes, Environment and Personality. Austin TX: University of Texas Press.

Roles

All of us shared almost the same amount of the work.

Appendix

Original Dataset (first 5 rows)

pairnum	sex	zygosity	moed	faed	faminc	english	math	socsci	natsci	vocab
1	2	1	3	4	2	14	13	17	18	14
1	2	1	3	4	2	11	14	15	10	12
4	2	1	1	1	1	20	20	16	16	13
4	2	1	1	1	1	17	19	13	13	14
5	2	1	1	1	1	11	8	15	16	12

R Code

```{r}
library(ggplot2)
library(plyr) # for rename
library(corrplot)

# load the data
twins <- read.csv("~/Documents/6215-Applied Multivariance Model/project/nmtwins.csv")
names(twins)
```

```{r}
#seperate the data into two groups
#identical <- twins[twins$zygosity==1,]
#fraternal <- twins[twins$zygosity==2,]

# describe the data:
#histgram
hist(twins$sex)
hist(twins$zygosity)
hist(twins$moed)
hist(twins$faed)
hist(twins$faminc)
hist(twins$english)

table(twins$moed)
qplot(twins$moed, geom="histogram",main = "mother education", xlab = "mother education",xlim=c(0,7))

#boxplot
boxplot(twins$english~twins$zygosity,data=twins, main="Twins",
        xlab="zygosity", ylab="English")
boxplot(twins$math~twins$moed,data=twins, main="Twins",
        xlab="mother education", ylab="English")
odd <- seq(1, by = 2, len = 839)
# checking the fraternal twins sex
same_zygosity <- c()
for(i in 1:839){
  same_zygosity[i] <- twins[(2*i-1),"zygosity"]!=twins[2*i,"zygosity"]
}
sum(same_zygosity ) # if sum is 0, means all two twins in a pair have the same zygosity

# from the result we can see that if two people are twins, their sex, zygosity, mother education, father education, family income are the same. Except sex, other variables are reasonalble to be the same for a pair of twins as we assume the pair of twins are in the same family environment.
```

```{r}
# missing data: random sample impute
sum(is.na(twins$moed)) # 38/1678 is missing. Only 19 pairs is missing

hist(twins$faed)
sum(is.na(twins$faed)) # 48/1678  is missing. Only 24 pairs are missing

hist(twins$faminc)
sum(is.na(twins$faminc)) #124/1678  is missing. Only 62 pairs are missing

# missing value impute sampling in probability. Instead using the mean or mode for imputation, missing value is imputed using the distribution of the origional data.
# random
set.seed(1992)
twins[which(is.na(twins$moed)),"moed"] <- sample(twins[-which(is.na(twins$moed)),"moed"],size = sum(is.na(twins$moed)))
twins[which(is.na(twins$faed)),"faed"] <- sample(twins[-which(is.na(twins$faed)),"faed"],size = sum(is.na(twins$faed)))
twins[which(is.na(twins$faminc)),"faminc"] <- sample(twins[-
```

```
which(is.na(twins$faminc)),"faminc"],size =
sum(is.na(twins$faminc)))
```

```{r}
# checking normality distribution for each score
variables for the assumption of
qqnorm(twins$english)
qqline(twins$english, col="red")
qqnorm(twins$math)
qqline(twins$math, col="red")
qqnorm(twins$socsci)
qqline(twins$socsci, col="red")
qqnorm(twins$natsci)
qqline(twins$natsci, col="red")
qqnorm(twins$vocab)
qqline(twins$vocab, col="red")
# from the qq plot for each score variables, we didint
see a significant diviation from normal distribution.


# test multivariate normal distribution: QQ plot(for
sigle variable)  Chi-square (for multi variables)
source("http://www.stat.wmich.edu/wang/561/code
s/R/chisqplot.R")
chisqplot(twins[,7:11])
# a few points above the line indicate light positive
skewness.
## why the multi-normal has some divations?
correlation? outlier?
# To check the observation with large distances are
outliers or mistake filling.

 #This is an example for positive skewness looks like in
histogram and QQ plot
 N <- 10000
 x <- rnbinom(N, 10, .5)
 hist(x,
 xlim=c(min(x),max(x)), probability=T, nclass=max(x)-
min(x)+1,
   col='lightblue', xlab=' ', ylab=' ', axes=F,
   main='Positive Skewed')
lines(density(x,bw=1), col='red', lwd=3)
 chisqplot(x)


# squared generalized distances
score <- twins[,7:11]
mean <- colMeans(score)
S=cov(score)
S.inv=solve(S)
ssd=matrix(1:nrow(twins),nrow=1)
n <- nrow(twins)
for(i in 1:n)
{ssd[i]=as.matrix(score[i,]-
mean)%*%as.matrix(S.inv)%*%t(score[i,]-mean)}
# from the chi-square plot, we can see that diviation
stars after squared generalized distances is greater
than 15.So we find the corresponding score.

qchi=c(1:n)
for(i in 1:n)
{qchi[i]=qchisq(p=((i-1/2)/n),df=5)}

ssdsort <- sort.int(ssd,index.return = TRUE)
distance <- ssdsort$x-qchi
hist(distance)
which(ssdsort$x-qchi>2)
(ssdsort$ix[c(1661,1668,1669,1670,1671,1675,1676,1
677)]) # shows which observation has large distance
ssdsort$x[c(1144, 1501,  114,  943,  21,  113, 1142,
645)] # shows squared generalized distances of
observation large form chisquare distance

twins$Colour="black"
# Set new column values to appropriate colours
twins$Colour[c(1661,1668,1669,1670,1671,1675,167
6,1677)]="red"
plot(ssdsort$x~qchi, col=twins$Colour)

score[c(1144, 1501,  114,  943,  21,  113, 1142,
645),]# shows the score

# Also , we calculate the 95% confidence interval of
mean score to see whether the potential outlier is out
of 95% confidence interval(2 standard distance from
the mean)
lower <- mean-2*sqrt(diag(S))
higher <- mean+2*sqrt(diag(S))
(score[c(1144, 1501,  114,  943,  21,  113, 1142,
645),]<lower)+(score[c(1144, 1501,  114,  943,  21,
113, 1142,  645),]>higher)
# except 1144,113, all others people's at least one
score is out of 95% CI. There is no evidence that they
are misfilling. We cannot directly delete them.

twins$Colour="black"
# Set new column values to appropriate colours
twins$Colour[c(1661,1675)]="red" # 1661,1675
corespond to the two observation(1144,113) whose
score is not falling outside the 95%CI
plot(ssdsort$x~qchi, col=twins$Colour)
score[c(1144,113),]
##????
twins$Colour <- NULL # remove the color attribute
```


```{r}
###outlier
```

```
score=as.matrix(score)
family=as.matrix(family)
fit=lm(score~family)
summary(fit)
install.packages("outliers")
library(outliers)
plot(fit,which=4)
chisq.out.test(score,variance=var(score),
opposite=FALSE)

fit <- lm(english~sex+moed+faed+zygosity,data=twins)
# Assessing Outliers
outlierTest(fit) # Bonferonni p-value for most extreme
obs
qqPlot(fit, main="QQ Plot") #qq plot for studentized
resid
leveragePlots(fit) # leverage plots
```

```{r}
##### Important: we have to calculate the mean score
of each pair of twins. Because each observation is not
independent. otherwise the variance matrix can not be
inversed. Like assume a pair of twins is one person.
index=rep(c(1,2),times=839)# 1678/2
twins=data.frame(twins,index)
twins_young=twins[twins$index==1,]
twins_old=twins[twins$index==2,]
twins_pair_raw <- data.frame(twins_young,twins_old)
twins_pair_raw2=twins_pair_raw[,-
c(12,13,14,15,16,17,18,24)]
twins_pair<-
rename(twins_pair_raw2,c("sex.1"="sex2","moed.1"="
moed2","faed.1"="faed2",
"faminc.1"="faminc2","english.1" = "english2",
"math.1" = "math2", "socsci.1" =
"socsci2","natsci.1"="natsci2", "vocab.1"="vocab2"))
mean_English <- c()
mean_math <- c()
mean_socsci <- c()
mean_natsci <- c()
mean_vocab <- c()
for(i in 1:nrow(twins_pair)){
  mean_English[i] <- (twins_pair[i,"english"] +
twins_pair[i,"english2"])/2
  mean_math[i] <- (twins_pair[i,"math"] +
twins_pair[i,"math2"])/2
  mean_socsci[i] <- (twins_pair[i,"socsci"] +
twins_pair[i,"socsci2"])/2
  mean_natsci[i] <- (twins_pair[i,"natsci"] +
twins_pair[i,"natsci2"])/2
  mean_vocab[i] <- (twins_pair[i,"vocab"] +
twins_pair[i,"vocab2"])/2
}
twins_pair <-
data.frame(twins_pair,mean_English,mean_math,mean
_socsci,mean_natsci,mean_vocab)
names(twins_pair)
```

```{r}
#sex/gender: different sex can cause score difference?
Now we are viewing the mean_score as generated by
independent indivaidual. There is no twin effects.

# we can check the covariance matrix equality CH 6
textbook page310
# one of the assumption made when comparing two or
more multivariate mean vectors is the covariance
matrics of the potentially different group are the same
# H0: equal covariance matrices
# test: Box's M-test  (testbook page310)
#install.packages("biotools")
library(biotools)
twins_pair[2:6,17:21]
boxM(twins_pair[,17:21],twins_pair[,2]) # test
different sex groups have same covariance matrix.
# result: Chi-Sq (approx.) = 12.2032, df = 15, p-value =
0.6636
# P-value is greater than 0.5 which indicate that there
is no enough evidence to reject the Null hypothsis.

# two population groups: female and male
# compare these two groups mean score difference
# assumption: 1. the group covariance are equal
according to the Box's M-test result
# asummption: 2 both population are multivairate
normal
# H0: mu1-mu2=0     page 285
# using F-test to see whether there exists significant
(ch6 page 285)

# since the sample size is large,n1-p= , n2-p= , we can
assume (Ch6 ppt 14) T^2 follow chi-square
distribution
twins_pair_female = twins_pair[twins_pair$sex==2,] #
976
twins_pair_male = twins_pair[twins_pair$sex==1,]
#702
x1bar = colMeans(twins_pair_female[,17:21])
x2bar = colMeans(twins_pair_male[,17:21])
S1 = var(twins_pair_female[,17:21])
S2 = var(twins_pair_male[,17:21])
n1 = nrow(twins_pair_female)
n2 = nrow(twins_pair_male)
twins_pair[1:5,]
p=5
Spool=((n1-1)*S1+(n2-1)*S2)/(n1+n2-2)
T2=t(x1bar-
x2bar)%*%solve((1/n1+1/n2)*Spool)%*%(x1bar-
```

```
x2bar)
T2
# [1,] 313.1292
c2 = (n1+n2-2)*p*qf(df1=p,df2=n1+n2-p-
1,0.95)/(n1+n2-p-1)
c2
# [1] 11.17768
#############
#Since T2>c2, the conclusion could be drawn that
mean score exists significant
#difference in between different sexe groups.
#############

#Bonferroni 95% CI for mu_1-mu_2  (female-male)
crit = qt(1-0.05/(2*p),n1+n2-2)
se = sqrt(diag(Spool)*(1/n1+1/n2))
BCI = cbind((x1bar - x2bar) - crit*se,
            (x1bar - x2bar) + crit*se)
colnames(BCI)=c("LB","UB")
rownames(BCI)=c("English","math","socsci","natsci","v
ocab" )
BCI
# confidence interval show that:
# 1. only vocabulary mean score contains 0.
# 2. Female will get slightly higher score in English.
Male would performs better in math, social science,
natual science.

# Dotplot: Grouped Sorted and Colored
# Sort by score, group and color by sex
x <-
twins_pair[order(twins_pair$sex,twins_pair$mean_ma
th),] # sort by mpg
x$sex <- factor(x$sex) # it must be a factor
x$color[x$sex==1] <- "red"
x$color[x$sex==2] <- "blue"
dotchart(x$mean_math,labels=row.names(x),cex=.7,gr
oups= x$sex,
            main="Math score\ngrouped by sex",
  xlab="Math score", ylab="1-male,2-
female",gcolor="black", color=x$color)
# form the plot, we can see a clear difference between
male and female in mean math score. Male performs
better than female.
```

```{r}
#zygosity
############
# mean score differences
##########
#zygosity: different type of twins (identical, fraternal)
can cause mean score difference?

# we can check the covariance matrix equality CH 6
textbook page310
# H0: different type of twins have same covariance
matrix.
boxM(twins_pair[,17:21],twins_pair[,3])
# result: Chi-Sq (approx.) = 30.4749, df = 15, p-value =
0.01032
# The coviance matrix in different zygosity are
different

# two population groups: identical(1) and fraternal(2)
# compare these two groups mean score difference
# assumption: 1. covariance matrix are different
# assumption: 2. both are multi mornal
# H0: mu1-mu2=0     page 285
# using F-test to see whether there exists significant
(ch6 page 285)

twins_pair_identical =
twins_pair[twins_pair$zygosity==1,] # 976
twins_pair_fraternal =
twins_pair[twins_pair$zygosity==2,] #702
x1bar = colMeans(twins_pair_identical[,17:21])
x2bar = colMeans(twins_pair_fraternal[,17:21])
S1 = var(twins_pair_identical[,17:21])
S2 = var(twins_pair_fraternal[,17:21])
n1 = nrow(twins_pair_identical)
n2 = nrow(twins_pair_fraternal)
p=5
T2=t(x1bar-
x2bar)%*%solve((1/n1*S1+1/n2*S2))%*%(x1bar-
x2bar)
T2
# [1,] 8.783256
c2 = (n1+n2-2)*p*qf(df1=p,df2=n1+n2-p-
1,0.95)/(n1+n2-p-1)
c2
# [1] 11.17768
#############
#Since T2<c2, the conclusion could be drawn that
there is no enough statistical evience to reject the Null.
No significant difference between different type of
twins.

#############
#Bonferroni 95% CI for mu_2-mu_1
crit = qt(1-0.05/(2*p),n1+n2-2)
se = sqrt(diag(1/n1*S1+1/n2*S2))
BCI = cbind((x1bar - x2bar) - crit*se,
            (x1bar - x2bar) + crit*se)
colnames(BCI)=c("LB","UB")
rownames(BCI)=c("English","math","socsci","natsci","v
ocab" )
BCI
# 0 are include in all the confidence interval.
```

```
######################
##the difference in absolute difference between identical twins and fraternal twins##
##############################
twin[1:10,]
twin2[1:10,]
x=twin2[,7:11]
nrow(x)
y=twin2[,12:16]
nrow(y)
z=abs(x-y)#the absolute difference of scores between twins
z
nrow(z)
colMeans(z)
df=data.frame(twin2,z)#dataset we will use
df[1:10,]
names(df)
z1=df[c(df$zygosity==1),17:21]#data who are identical twins
head(z1)
nrow(z1)##509 ∂‘Õ¨¬—À´∞°Ã•
z2=df[c(df$zygosity==2),17:21]#data who are not identical twins
nrow(z2) ##330 ∂‘“Ï¬—À´∞°Ã•
head(z2)

nrow(z1)
nrow(z2)
##two absokute difference comparison test
x1bar=colMeans(z1)
x1bar
x2bar=colMeans(z2)
x2bar
n1=nrow(z1)
n1
n2=nrow(z2)
n2
p=5
#assuming that the variance is equal within two groups
# ch 6 ppt 4
S1=var(z1)
S2=var(z2)
Spool=((n1-1)*S1+(n2-1)*S2)/(n1+n2-2)
T2=t(x1bar-x2bar)%*%solve((1/n1+1/n2)*Spool)%*%(x1bar-x2bar)
T2
T2a=(n1+n2-2)*p*qf(df1=p,df2=n1+n2-p-1,0.96)/(n1+n2-p-1)
T2a
#T2>T2a we can reject H0 that two population mean vectors are equivalen
#t-Test to compare the means of two groups under the assumption that both samples are random, independent, and come from normally distributed population with unknow but equal variances

#To solve this problem we must use to a Student’s t-test with two samples, assuming that the two samples are taken from populations that follow a Gaussian distribution (if we cannot assume that, we must solve this problem using the non-parametric test called Wilcoxon-Mann-Whitney test; we will see this test in a future post). Before proceeding with the t-test, it is necessary to evaluate the sample variances of the two groups, using a Fisher’s F-test to verify the homoskedasticity (homogeneity of variances).
var.test(as.vector(z1[1,1:10]),as.vector(z2[1,1:10]))
#http://www.r-bloggers.com/two-sample-students-t-test-1/
t.test(z1[1],z2[1])
t.test(z1[2],z2[2])
t.test(z1[3],z2[3])
t.test(z1[4],z2[4])
t.test(z1[5],z2[5])
t.test(z1,z2)
nrow(z)
```

```{r}
#motheredu: different mother education can cause score difference?
table(twins_pair$moed)
#  1   2   3   4   5   6
#  55 108 310 203 126  37

boxM(twins_pair[,17:21],twins_pair[,4])
#          Box's M-test for Homogeneity of Covariance Matrices
#   Chi-Sq (approx.) = 91.733, df = 75, p-value = 0.09182
#########
## We don't have enough evidence to reject that the six group mean score covariance matrix are the same
#########

# six population groups: 1~6
# compare these six groups mean score difference
# H0:    there is no mother education effect page 301-302
# 1: X_li is distributed as N(mu+tao_l,sigma) page(297)
# 2: covariance matrix is the same for all populations page(301)
# 3: error_li are independent Np(0,sigma) variables
# Anova vs. Manova
# Why not multiple Anovas?
```

Anovas run separately cannot take into account the pattern of covariation among the dependent measures:
It may be possible that multiple Anovas may show no differences while the Manova brings them out
MANOVA is sensitive not only to mean differences but also to the direction and size of correlations among the dependents

# One-way MANOVA
score <- as.matrix(twins_pair[,17:21])
mothereducation <- as.factor(twins_pair[,4])
manova(score~mothereducation)
summary(manova(score~mothereducation))
# Get a p-value for the effect of mother education on the 5 score measurements
# Mother education has effect on mean score.

# if we ignore the pattern of covariation among the dependent measures, we try 5 times one way anova:
summary(aov(score~mothereducation))
# P-value tells us that Mother education has effect on each mean score.

Simultanious confidence interval
```

```{r}
#fatheredu
table(twins_pair$faed)
#  1   2   3   4   5   6
# 93 101 219 179 125 122
#fatheredu: different father education can cause score difference?
boxM(twins_pair[,17:21],twins_pair[,5])
# Box's M-test for Homogeneity of Covariance Matrices
# results: Chi-Sq (approx.) = 92.1389, df = 75, p-value = 0.08714
# same covariance matrix

score <- as.matrix(twins_pair[,17:21])
fathereducation <- as.factor(twins_pair[,5])
manova(score~fathereducation)
summary(manova(score~fathereducation))
# Get a p-value for the effect of father education on the 5 score measurements
# father education has effect on mean score.
#CI
```

```{r}
#family income
table(twins_pair$faminc)
#  1   2   3   4   5   6   7
# 99 213 177 199  83  25  43
#fatheredu: different father education can cause score difference?
boxM(twins_pair[,17:21],twins_pair[,6])
# Box's M-test for Homogeneity of Covariance Matrices
# Chi-Sq (approx.) = 95.569, df = 90, p-value = 0.3241
familyincome <- as.factor(twins_pair[,6])
manova(score~familyincome)
summary(manova(score~familyincome))
# Get a p-value for the effect of family income on the 5 score measurements
# family income has effect on mean score.
#CI
```

```{r}
# 6 moed levels divide into 2 group
group1 <- twins_pair[which(twins_pair$moed==1|twins_pair$moed==2|twins_pair$moed==3),17:21]
group2 <- twins_pair[which(twins_pair$moed==4|twins_pair$moed==5|twins_pair$moed==6),17:21]
x1bar = colMeans(group1)
x2bar = colMeans(group2)
S1 = var(group1)
S2 = var(group2)
n1 = nrow(group1)
n2 = nrow(group2)
p=5
Spool=((n1-1)*S1+(n2-1)*S2)/(n1+n2-2)
T2=t(x1bar-x2bar)%*%solve((1/n1+1/n2)*Spool)%*%(x1bar-x2bar)
T2
# [1,] 8.5515
c2 = (n1+n2-2)*p*qf(df1=p,df2=n1+n2-p-1,0.95)/(n1+n2-p-1)
c2
# [1] 11.17768
#Bonferroni 95% CI for mu_1-mu_2  (female-male)
crit = qt(1-0.05/(2*p),n1+n2-2)
se = sqrt(diag(Spool)*(1/n1+1/n2))
BCI = cbind((x1bar - x2bar) - crit*se,
  (x1bar - x2bar) + crit*se)
colnames(BCI)=c("LB","UB")
rownames(BCI)=c("English","math","socsci","natsci","vocab" )
BCI
# 6 faed levels divide into 2 group
group1 <- twins_pair[which(twins_pair$faed==1|twins_pair$faed==2|twins_pair$faed==3),17:21]
group2 <- twins_pair[which(twins_pair$faed==4|twins_pair$faed==5|twins_pair$faed==6),17:21]
x1bar = colMeans(group1)
x2bar = colMeans(group2)
```
```

```
S1 = var(group1)
S2 = var(group2)
n1 = nrow(group1)
n2 = nrow(group2)
p=5
Spool=((n1-1)*S1+(n2-1)*S2)/(n1+n2-2)
T2=t(x1bar-
x2bar)%*%solve((1/n1+1/n2)*Spool)%*%(x1bar-
x2bar)
T2
# [1,] 8.5515
c2 = (n1+n2-2)*p*qf(df1=p,df2=n1+n2-p-
1,0.95)/(n1+n2-p-1)
c2
# [1] 11.17768
#Bonferroni 95% CI for mu_1-mu_2  (female-male)
crit = qt(1-0.05/(2*p),n1+n2-2)
se = sqrt(diag(Spool)*(1/n1+1/n2))
BCI = cbind((x1bar - x2bar) - crit*se,
  (x1bar - x2bar) + crit*se)
colnames(BCI)=c("LB","UB")
rownames(BCI)=c("English","math","socsci","natsci","v
ocab" )
BCI
```

```{r}
# interaction: using MANOVA
sex=as.factor(twins_pair$sex)
zygosity=as.factor(twins_pair$zygosity)
moedu=as.factor(twins_pair$moed)
faedu=as.factor(twins_pair$faed)
faminc=as.factor(twins_pair$faminc)
score <- as.matrix(twins_pair[,17:21])

# interaction: sex*zygosity
fit=manova(score~sex*zygosity)
summary(fit)
# no zygosity effect or intercation

# intercation: sex*family income
fit=manova(score~sex*faminc)
summary(fit)
# no intercation effect

# intercation: sex*moedu
fit=manova(score~sex*moedu)
summary(fit)
# no intercation effect

# intercation: sex*faedu
fit=manova(score~sex*faedu)
summary(fit)
# no intercation effect

# intercation: zygosity*moedu
fit=manova(score~moedu*zygosity)
summary(fit)
# no intercation effect or zygosity

# intercation: zygosity*moedu
fit=manova(score~zygosity*faedu)
summary(fit)
# no intercation effect or zygosity

# intercation: zygosity*faminc
fit=manova(score~zygosity*faminc)
summary(fit)
# no intercation effect or zygosity

# intercation: moedu*faedu
fit=manova(score~moedu*faedu)
summary(fit)
# no intercation effect

# intercation: moedu*faminc
fit=manova(score~moedu*faminc)
summary(fit)
# no intercation effect

# intercation: faedu*faminc
fit=manova(score~faedu*faminc)
summary(fit)
# no intercation effect

# summary: no intercation effect across sex, zygosity,
mather education, father education, family income.
```

```{r}
#score
# now we want to see how to use these factors to
explain the score: regression
#ch5 ppt 15 assume all score are from same normal
distribution N(mu,sigma)
twins_mean <- colMeans(twins[,7:11])
source("http://www.stat.wmich.edu/wang/561/code
s/R/ci.R")
confidence(n=1678,xbar=mean,S=S,
conf.region=T,alpha=.05)

sex=as.factor(twins_pair$sex)
zygosity=as.factor(twins_pair$zygosity)
moedu=as.factor(twins_pair$moed)
faedu=as.factor(twins_pair$faed)
faminc=as.factor(twins_pair$faminc)
score <- as.matrix(twins_pair[,17:21])

#factor <- data.frame(sex,zygosity,moedu,faedu,faminc)
fit <- lm(score~sex+zygosity+moedu+faedu+faminc)
```

```
summary(fit)
# For english score :
## coefficients in sex, father education, family income are significant
## But the Adjust R-squre is low: 0.07
# For math score:
##coefficients in sex, father education, family income are significant
## But the Adjust R-squre is low: 0.19
# For socsci:
## coefficients in sex, father education, family income are significant
## But the Adjust R-squre is low: 0.08
# For natsci:
## coefficients in sex, father education, family income are significant
## But the Adjust R-squre is low: 0.11
# For vocab:
## coefficients in father education, family income  are significant
## But the Adjust R-squre is low: 0.11
## ??? why  mother education is not significant?
```

```{r}
# one way anova: twin effect:
oneway_anova_twineffect <-
aov(twins$english~twins$pairnum)
summary(oneway_anova_twineffect)
```
################################
##the difference in absolute difference between identical twins and fraternal twins############################
twin[1:10,]
twin2[1:10,]
x=twin2[,7:11]
nrow(x)
y=twin2[,12:16]
nrow(y)
z=abs(x-y)#the absolute difference of scores between twins
z
nrow(z)
colMeans(z)
df=data.frame(twin2,z)#dataset we will use
spss=df[,c(-4:-16)]
head(spss)
write.csv(spss,"spss.csv",row.names=F)
df1=df[c(df$zygosity==1),]
head(df1)
df11=df1[c(df1$sex==1),17:21]
head(df11)
df11=na.omit(df11)
colMeans(df11)#identical male abs diff mean
df12=df1[c(df1$sex==2),17:21]
df12=na.omit(df12)
colMeans(df12)#identical female abs diff mean
##english.1   math.1 socsci.1 natsci.1   vocab.1
##2.310580 3.436860 2.450512 3.573379 1.897611
df2=df[c(df$zygosity==2),]
df21=df2[c(df2$sex==1),17:21]
df21=na.omit(df21)
colMeans(df21)#fraternal male
##english.1   math.1 socsci.1 natsci.1   vocab.1
##3.207407 4.859259 4.200000 4.088889 3.251852
df22=df2[c(df2$sex==2),17:21]
df22=na.omit(df22)
colMeans(df22)#fraternal female
##english.1   math.1 socsci.1 natsci.1   vocab.1
##3.589744 5.020513 3.471795 4.184615 2.820513
#test the normality of scores
par(mfrow=c(1,2))
source("http://www.stat.wmich.edu/wang/561/codes/R/chisqplot.R")
chisqplot(df[,17:21])
#box-cox transformation
source("http://www.stat.wmich.edu/wang/561/codes/R/bcplot.R")
bcplot(df[,17:21],main="Box-Cox Transformation")

#test covariate matrix equality
boxM(df[,17:21],df[,2])
##Box's M-test for Homogeneity of Covariance Matrices

##data: df[, 17:21]
##Chi-Sq (approx.) = 30.4548, df = 15, p-value = 0.01038

##covariate matrix is not the same
##two absolute difference comparison test
##after transformation
##Box's M-test for Homogeneity of Covariance Matrices

##data: df[, 17:21]
##Chi-Sq (approx.) = 13.0983, df = 15, p-value = 0.5947
z=log(z+1)
df=data.frame(twin2,z)
df[1:10,]
z1=df[c(df$zygosity==1),17:21]#data who are identical twins
head(z1)
nrow(z1)##509 ???????????
z2=df[c(df$zygosity==2),17:21]#data who are not identical twins
nrow(z2) ##330 ???????????
head(z2)
```
```

```
nrow(z1)
nrow(z2)
x1bar=colMeans(z1)
x2bar=colMeans(z2)
n1=nrow(z1)
n2=nrow(z2)
p=5
S1=cov(z1)
S2=cov(z2)
T2a=(n1+n2-2)*p*qf(df1=p,df2=n1+n2-p-
1,0.96)/(n1+n2-p-1)
T2a
##[1] 11.76112
Spool=((n1-1)*S1+(n2-1)*S2)/(n1+n2-2)
T2=t(x1bar-
x2bar)%*%solve((1/n1+1/n2)*Spool)%*%(x1bar-
x2bar)
T2
pval = 1-pf(T2*(n1+n2-p-1)/((n1+n2-2)*p), p, n1+n2-
p-1)
pval
##[,1]
##[1,] 0

################################
#  sex and zygosity interaction  #######
################################
head(df)
Y=as.matrix(df[,17:21])
fit.zyandsex=manova(Y~df$zygosity*df$sex)
summary(fit.zyandsex)
##Df   Pillai approx F num Df den Df    Pr(>F)
##df$zygosity        1 0.109382  20.4120     5    831 <
2.2e-16 ***
##  df$sex           1 0.012521   2.1074     5    831
0.062495 .
##df$zygosity:df$sex   1 0.018281   3.0949     5    831
0.008931 **
##  Residuals        835
##---
##  Signif. codes:  0 ??***?? 0.001 ??**?? 0.01 ??*??
0.05 ??.?? 0.1 ?? ?? 1
##>

################################
#twins and nontwins effect#
################################
X1=twin2[c(twin2$zygosity==1),7:11]
X2=twin2[c(twin2$zygosity==1),12:16]
Xbar=(X1+X2)/2
X1=as.matrix(X1)
X2=as.matrix(Xbar)
Xbar=as.matrix(Xbar)
t(X1-Xbar)%*%(X1-Xbar)
W1=t(X1-Xbar)%*%(X1-Xbar)
W2=t(X2-Xbar)%*%(X2-Xbar)
W=W1+W2
W

twin[1:10,]
bar=colMeans(twin[c(twin$zygosity==1),7:11])
matrix=Xbar-bar
sum=matrix(c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0),nrow=5)
sum

for (i in 1:509){
 a=matrix[i,]%*%t(matrix[i,])
 sum=sum+a
 print(sum)
}
B=sum

Lam=det(W)/det(B+W)
g=509
p=5
n=nrow(X1)
tval = -(n-1-(p+g)/2)*log(Lam)
tval
#[1] 1544.884
qchisq(0.95,p*(g-1))
#[1] 2658.361
pval = 1-pchisq(tval, p*(g-1))
pval
#[1] 1

X1=twin2[,7:11]
X2=twin2[,12:16]
Xbar=(X1+X2)/2
X1=as.matrix(X1)
X2=as.matrix(X2)
Xbar=as.matrix(Xbar)
t(X1-Xbar)%*%(X1-Xbar)
W1=t(X1-Xbar)%*%(X1-Xbar)
W2=t(X2-Xbar)%*%(X2-Xbar)
W=W1+W2
W

twin[1:10,]
bar=colMeans(,7:11])
matrix=Xbar-bar
sum=matrix(c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0),nrow=5)
sum

for (i in 1:839){
 a=matrix[i,]%*%t(matrix[i,])
 sum=sum+a
 print(sum)
```

```
}
B=sum
B

Lam=det(W)/det(B+W)
g=839
p=5
n=nrow(X1)
tval = -(n-1-(p+g)/2)*log(Lam)
tval
#[1]2151.728
qchisq(0.95,p*(g-1))
#4341.702
pval = 1-pchisq(tval, p*(g-1))
pval
#[1] 1

twin[1:10,]
y=as.matrix(twin[,7:11])
fit=manova(y~twin$pairnum)
summary(fit,test="Wilks")

y1=as.matrix(twin[c(twin$zygosity==1),7:11])
data1=twin[c(twin$zygosity==1),]
fit1=manova(y1~data1$pairnum)
summary(fit1,test="Wilks")

y2=as.matrix(twin[c(twin$zygosity==2),7:11])
data2=twin[c(twin$zygosity==2),]
fit2=manova(y2~data2$pairnum)
summary(fit2,test="Wilks")

```{r}
# PCA
# Why PCA: plot two dimention. data visualization
library(corrplot)
corrplot(cor(twins_pair[,17:21]),method="number")
# all the score have some correlation.
# scoial science and vocabulary are high correlated.
#
fit <- princomp(twins_pair[,17:21], cor=TRUE)
summary(fit)
loadings(fit)
fit$scores[1:10,]
plot(fit,type="lines") #indicate 2 component

#reduction data: using correlation matrix
raw = as.matrix(twins_pair[,17:21])
xbar = colMeans(raw) #overall mean across all 3
species
c.raw = t(t(raw)-xbar)
S1 <- cor(twins_pair[,17:21])
S1
#plot the first two principal components
eg1 = eigen(S1)$vectors[,1:2]
PC1 = c.raw%*%eg1
twins_pair$sex_fm[which(twins_pair$sex==1)] <- "m"
twins_pair$sex_fm[which(twins_pair$sex==2)] <- "f"
plot(PC1,col=twins_pair$sex, pch=twins_pair$sex_fm,
main="PC",ylab="PC2",xlab="PC1")

#plot(PC1,col=twins_pair$zygosity,
pch=twins_pair$zygosity,
main="PC",ylab="PC2",xlab="PC1")

#plot(PC1,col=twins_pair$faedu,
pch=twins_pair$faedu,
main="PC",ylab="PC2",xlab="PC1")

#plot(PC1,col=twins_pair$faminc,
pch=twins_pair$faminc,
main="PC",ylab="PC2",xlab="PC1")
```
```