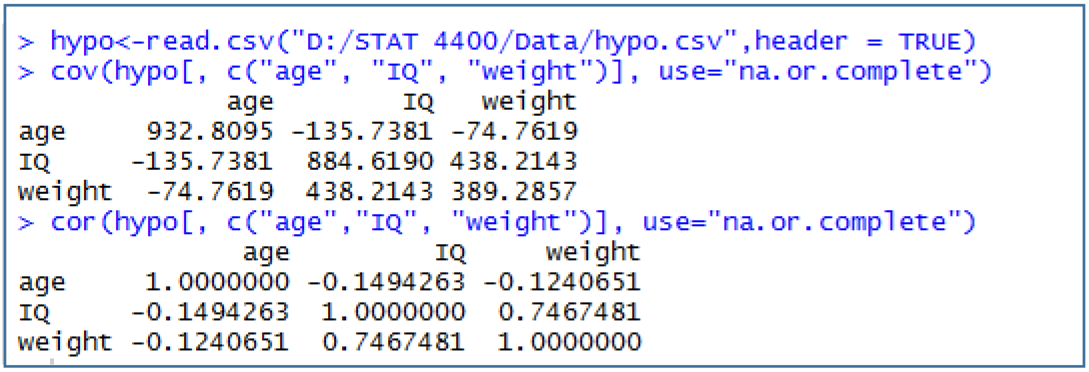
Ex. 1.1 a. Find the correlation matrix based on covariance matrix and statistical software for the data in Table 1.1.
b. Find covariance matrix of the data in Table 1.1.

Code and output:



```
> hypo<-read.csv("D:/STAT 4400/Data/hypo.csv",header = TRUE)
> cov(hypo[, c("age", "IQ", "weight")], use="na.or.complete")
             age        IQ    weight
age     932.8095 -135.7381 -74.7619
IQ     -135.7381  884.6190 438.2143
weight  -74.7619  438.2143 389.2857
> cor(hypo[, c("age","IQ", "weight")], use="na.or.complete")
              age         IQ     weight
age     1.0000000 -0.1494263 -0.1240651
IQ     -0.1494263  1.0000000  0.7467481
weight -0.1240651  0.7467481  1.0000000
```

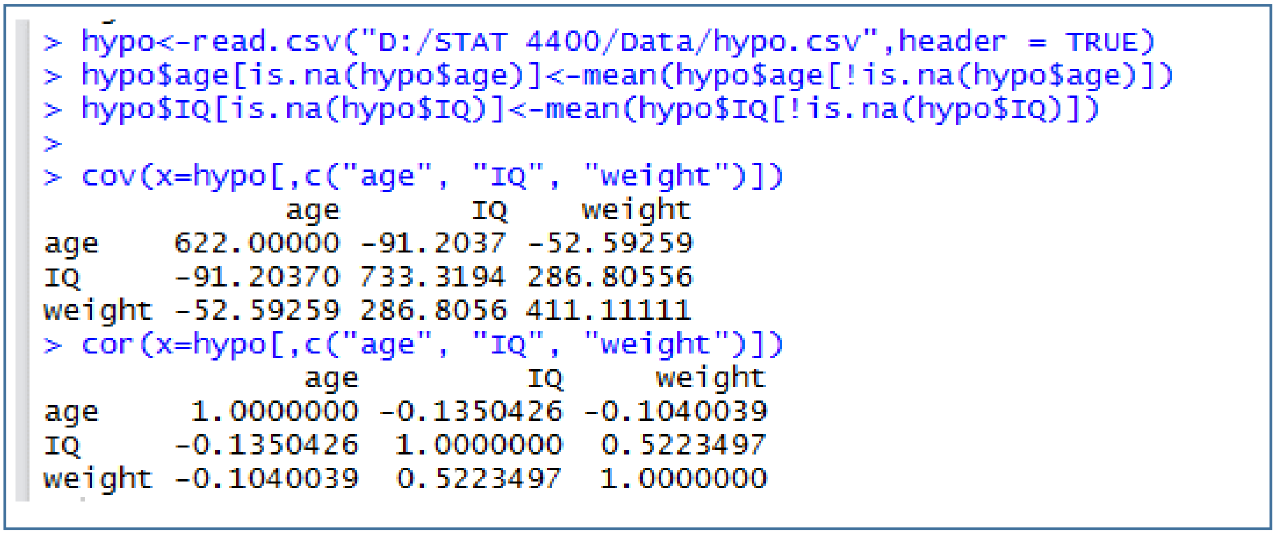
Comments:
Both covariance and correlation matrixes are available and provided the ways to determine how two variable are related. However, covariance determines whether the variables were increasing or decreasing, but it was impossible to measure the degree to which the variables moved together because covariance does not use one standard unit of measurement. To measure the degree to which variables move together, we use correlation.

The correlation matrix shows that age and IQ scores, age and weight are weakly negatively related. It is interesting to note the highly positively correlated between IQ and Weight. They have a high correlation 0.75, meaning when weight increases, IQ increasing as well. However, this data set is very small. Especially after removing the observations using complete-case analysis, the sample size decreases to 7. Therefore, the sample correlation is biased estimates of the population correlation between weight and IQ scores.

Ex. 1.2 Fill in the missing values in Table 1.1 with appropriate mean values, and recalculate the correlation matrix of the data.

a. Fill in the missing values in Table 1.1 with appropriate mean values.
b. Recalculate the correlation matrix of the data.

Code and output:

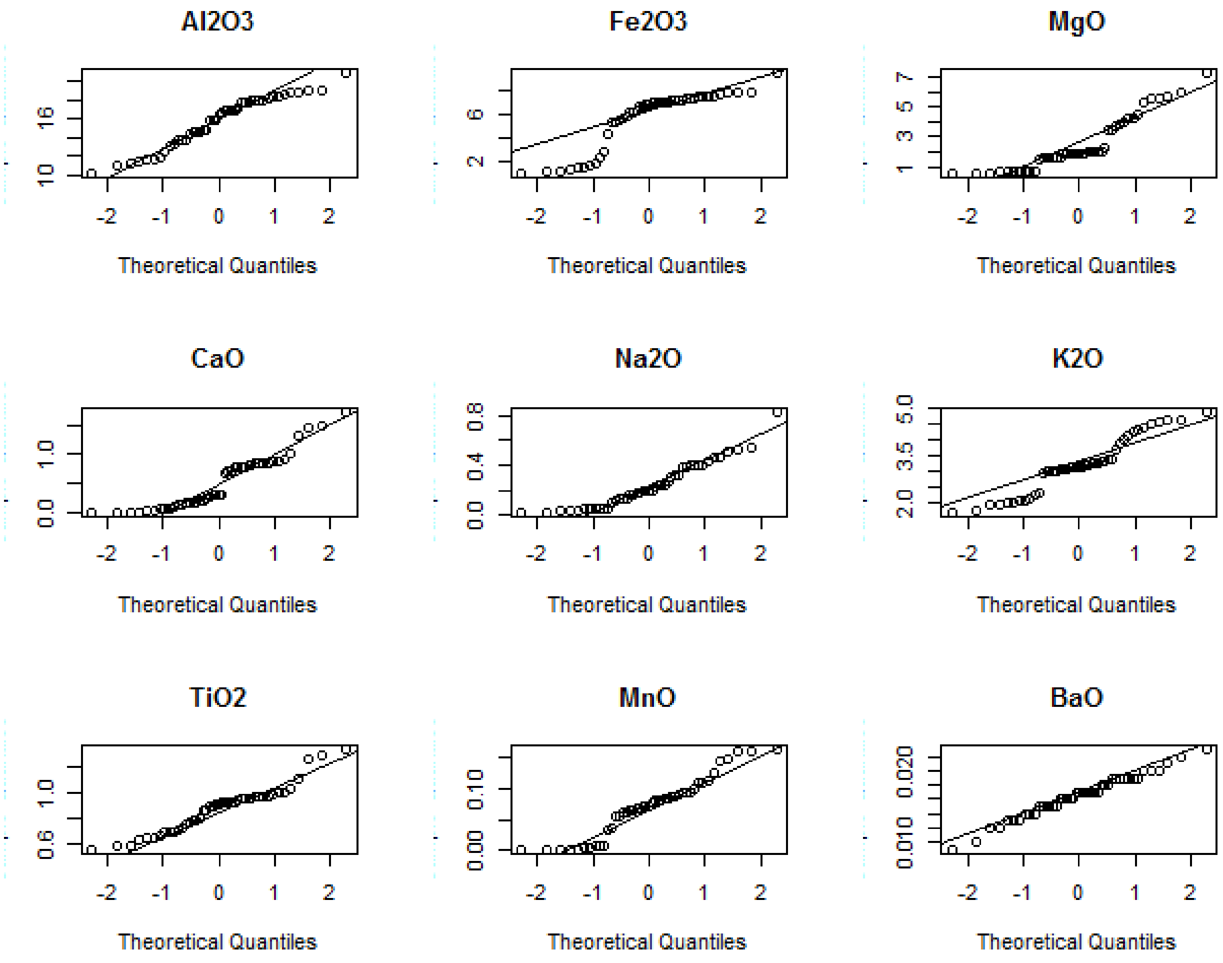


```
> hypo<-read.csv("D:/STAT 4400/Data/hypo.csv",header = TRUE)
> hypo$age[is.na(hypo$age)]<-mean(hypo$age[!is.na(hypo$age)])
> hypo$IQ[is.na(hypo$IQ)]<-mean(hypo$IQ[!is.na(hypo$IQ)])
>
> cov(x=hypo[,c("age", "IQ", "weight")])
             age       IQ    weight
age    622.00000 -91.2037 -52.59259
IQ     -91.20370 733.3194 286.80556
weight -52.59259 286.8056 411.11111
> cor(x=hypo[,c("age", "IQ", "weight")])
              age         IQ     weight
age     1.0000000 -0.1350426 -0.1040039
IQ     -0.1350426  1.0000000  0.5223497
weight -0.1040039  0.5223497  1.0000000
```

Comments: The correlation between weight and IQ score is decreasing when fill in the missing values with appropriate mean values.

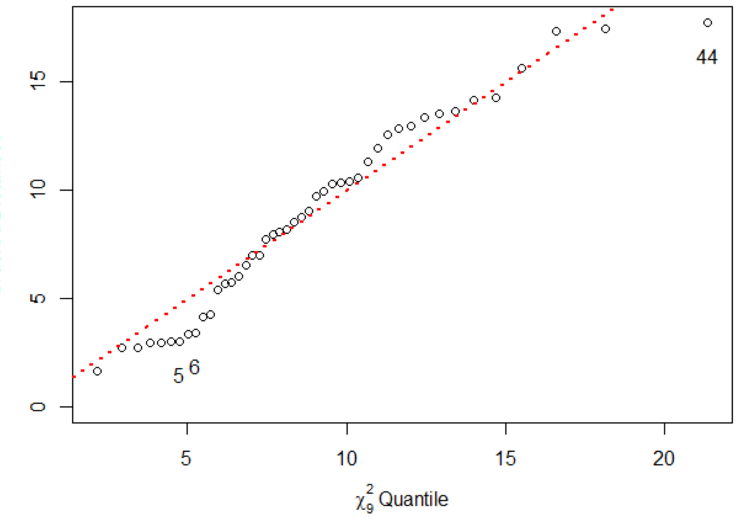
Ex. 1.3 Examine both the normal probability plots of each variable in the archaeology data in Table 1.3 and the chi-square plot of the data. Do the plots suggest anything unusual about the data?

```
df <- read.csv("D:/STAT 4400/Data/pottery.csv")
x <- df[-1]
par(mfrow = c(3,3))
for(i in 1:9){
  qqnorm(x[,i], main = names(x)[i]); qqline(x[,i])
}
row.names(x) <- df$X # to be able to display which obs are outliers on the plot
# following the example from page 19 in the textbook.
cm <- colMeans(x)
S <- cov(x)
d <- apply(x, 1, function(x) t(x-cm) %*% solve(S) %*% (x-cm))
par(mfrow=c(1,1)
plot(qc <- qchisq((1:nrow(x) - 1/2) / nrow(x), 9), sd <- sort(d),
    xlab = expression(paste(chi[9]^2, " Quantile")),
    ylab = "Ordered Distances", ylim = c(0, max(d)))
abline(0, 1, lty = 3, col = "red", lwd = 2)
outliers <- which(rank(abs(qc - sd), ties.method = "random") > nrow(x) - 3)
text(qc[outliers], sd[outliers] - 1.5, names(outliers)) # label the "outliers"
```



Comments: The qq plots show that all eight variables are not normal distributed except the variable of Bao.

qqChi-square plot



Comments: The variable $d^2 = (x-\mu)^T \Sigma^{-1}(x-\mu)$ has a chi-square distribution with 9 degrees of freedom, and for “large” samples (n=45 in the data set) the observed Mahalanobis distances have an approximate chi-square distribution. This result can be used to evaluate (subjectively) whether a data point may be an outlier and whether observed data may have a multivariate normal distribution.

The plot of the Mahalanobis distances is given above. The distances are on the vertical and the chi-square quantiles are on the horizontal. We see an upward and a downward bending from left to right. This indicates a possible violation of multivariate normality. In particular, the three points, labeled as 5, 6, and 44 might also be outliers.

Ex. 1.4 Convert the covariance matrix given below into the corresponding correlation matrix.

$$\begin{pmatrix} 3.8778 & 2.8110 & 3.1480 & 3.5062 \\ 2.8110 & 2.1210 & 2.2669 & 2.5690 \\ 3.1480 & 2.2669 & 2.6550 & 2.8341 \\ 3.5062 & 2.5690 & 2.8341 & 3.2352 \end{pmatrix}$$

Analysis:

The relationship between covariance and correlation matrix is that $R = D^{-\frac{1}{2}} S D^{\frac{1}{2}}$, where D is the diagonal matrix with elements of the four variances of the variables in its diagonal.

Code and output

```
x <- matrix(c(3.8778,2.8110,3.1480,3.5062,
+             2.8110,2.1210,2.2669,2.5690,
+             3.1480,2.2669,2.6550,2.8341,
+             3.5062,2.5690,2.8341,3.2352), byrow = T, ncol = 4)
> d<-diag(1/sqrt(diag(x)))
> x.cor<-d%*%x%*%d
> x.cor
          [,1]      [,2]      [,3]      [,4]
[1,] 1.0000000 0.9801619 0.9810921 0.9899048
[2,] 0.9801619 1.0000000 0.9552780 0.9807159
[3,] 0.9810921 0.9552780 1.0000000 0.9670131
[4,] 0.9899048 0.9807159 0.9670131 1.0000000
```

Ex. 1.5 For the small set of (10 × 5) multivariate data given below, find the (10 × 10) Euclidean distance matrix for the rows of the matrix. An alternative to Euclidean distance that might be used in some cases is what is known as city block distance (think New York). Write some R code to calculate the city block distance matrix for the data.

$$\begin{pmatrix} 3 & 6 & 4 & 0 & 7 \\ 4 & 2 & 7 & 4 & 6 \\ 4 & 0 & 3 & 1 & 5 \\ 6 & 2 & 6 & 1 & 1 \\ 1 & 6 & 2 & 1 & 4 \\ 5 & 1 & 2 & 0 & 2 \\ 1 & 1 & 2 & 6 & 1 \\ 1 & 1 & 5 & 4 & 4 \\ 7 & 0 & 1 & 3 & 3 \\ 3 & 3 & 0 & 5 & 1 \end{pmatrix}$$

- Analysis:

If the five variables are on completely different scales of measurement, then the larger values of the variables have larger inter sample differences. Those variables will dominate in the calculation of Euclidean distances and City-block distance. So the standardization must be performed before to find the distances.

- Code

```
x <- matrix(c(3,4,4,6,1,5,1,1,7,3,6,2,0,2,6,1,1,1,0,3,4,7,3,6,2,2,2,5,1,0,0,4,1,1,1,0,6,4,3,5,7,6,5,1,4,2,1,4,3,1),
ncol = 5)
d <- dist(x, method = "euclidean")
d.st<- dist(scale(x, center = FALSE))
d.mh <- dist(scale(x, method = "manhattan")
d.mhst <- dist(scale(x, center = FALSE), method = "manhattan")
```

- Output

```
> x <- matrix(c(3,4,4,6,1,5,1,1,7,3,6,2,0,2,6,1,1,1,0,3,4,7,3,6,2,2,2,5,1,0,0,4,1,1,1,0,6,4,3,5,7,6,5,1,4,2,1,4,3,1), ncol =
5)
> d <- dist(x, method ="euclidean")
> d
           1         2         3         4         5         6         7         8
2   6.557439
3   6.557439  5.477226
4   8.124038  6.244998  5.744563
5   4.242641  7.937254  6.855655  8.124038
6   7.615773  7.681146  3.605551  4.472136  6.782330
7  10.246951  8.000000  7.211103  8.185353  7.681146  7.280110
8   7.416198  4.242641  4.898979  6.708204  6.557439  6.708204  4.690416
9   9.273618  7.681146  4.582576  6.164414  8.831761  4.000000  7.141428  7.416198
10  9.273618  8.774964  7.141428  7.874008  6.480741  6.164414  3.605551  6.557439
           9
2
3
4
5
6
7
8
9
10  5.830952
```

```
> d.st<- dist(scale(x, center = FALSE))
> d.st
           1         2         3         4         5         6         7         8
2  1.8963439
3  1.9877945 1.4807660
4  2.1075437 1.5738961 1.4390458
5  1.0307641 2.1404899 2.0338251 2.1050455
6  2.0817793 1.9890028 0.9012348 1.1263614 1.9085728
7  2.8365651 1.9702631 1.9260890 2.1433675 2.2585764 2.0065174
8  2.1474191 1.0324001 1.2908882 1.6802059 1.9402005 1.7406923 1.1844039
9  2.5759062 1.9172372 1.1449371 1.5929515 2.4433786 1.0990388 1.7735962 1.7891913
10 2.4590441 2.1546544 1.9405586 2.0384740 1.7968371 1.7479204 0.9701138 1.6525151
           9
2
3
4
5
6
7
8
9
10 1.5477036

> d.mh<-dist(x, method="manhattan")
> d.mh
    1  2  3  4  5  6  7  8  9
2  13
3  11 10
4  16 11 11
5   8 17 11 16
6  14 15  7  8 12
7  21 16 14 15 13 11
8  15  8 10 13 11 13  8
9  20 15  9 12 16  8 13 13
10 18 15 15 14 14 12  7 13 12

> d.mhst <- dist(scale(x, center = FALSE), method = "manhattan")
> d.mhst
          1        2        3        4        5        6        7        8        9
2  3.635607
3  3.128311 2.728914
4  4.173991 2.788201 2.789964
5  1.971925 4.544658 3.066830 4.128373
6  3.720347 3.906491 1.803120 2.065581 3.273875
7  5.715635 4.029574 3.682823 3.937906 3.743710 2.935200
8  4.167494 1.988237 2.627880 3.328572 3.181964 3.328122 2.041337
9  5.389317 3.818840 2.261006 3.156388 4.357304 2.146304 3.324883 3.240472
10 4.820562 3.760045 4.038837 3.668377 3.788550 3.291214 1.881466 3.337263 3.187700
```

Ex. 1.6 A selection of four receipts from a university bookstore was obtained in order to investigate the nature of book sales. Each receipt provided, among other things, the number of books sold and the total amount of each sale.

a. Identify and interpret the number of units, the number of variables, the types of variables, and the levels of measurements of variables, respectively.

Comments:

There are four observations ($n = 4$).

The variable receipt is categorical variable at the level of ordinal measurement.

The variables of the amount of books and the total cost are quantitative variable at the level of ratio measurement.

The variable of student status is qualitative at the level of nominal measurement.

```
receipt<-c(101,521,746,857)
books<-c(4,5,4,3)
cost<-c(142,252,148,158)
stud.check<-c("yes","yes","no","yes")

ex1.6<-cbind.data.frame(receipt,books,cost,stud.check)
ex1.6

##   receipt books cost stud.check
## 1     101     4  142        yes
## 2     521     5  252        yes
## 3     746     4  148         no
## 4     857     3  158        yes
```

c. Read the observations x2 and x4 in a sub data matrix using statistical software.

```
ex1.6[c(2,4),]

##   receipt books cost stud.check
## 2     521     5  252        yes
## 4     857     3  158        yes
```

Ex. 1.7 Energy consumption in 2015, by state, from the major sources x_1= petroleum, x_2= natural gas, x_3= hydroelectric power, x_4= nuclear electric power is recorded in quadrillions (10^5) of BTU. The resulting mean and co-variance matrix are:

```
x.bar<-matrix(c(.766,.508,.438,.161),nrow = 4,ncol = 1,byrow = T)
s<-matrix(c(.856,.635,.173,.096,
                  .635,.568,.128,.067,
                  .173,.127,.171,.039,
                  .096,.067,.039,.043),
                nrow = 4,ncol = 4,byrow = T)


a.) Using the summary statistics, determine the sample mean and variance of a state’s tot
al energy consumption for these major sources.
```

Let Y = total energy consumption, then $Y = X_1 + X_2 + X_3 + X_4$, and $C = (1, 1, 1, 1)^T$.

$$Y = C^T X, \quad E(Y) = C^T \boldsymbol{\mu} \approx C^T \overline{\boldsymbol{X}}, \quad Var(Y) = C^T \sum C \approx C^T \boldsymbol{S} C$$

```
c<-matrix(c(1,1,1,1),nrow = 4)
sample.mean<-t(c) %*% (x.bar)
sample.var<-t(c)%*%s%*% c
sample.mean

## [1] 1.873

sample.var

##       [,1]
## [1,] 3.913
```

b. Determine the sample mean and variance of the excess of petroleum consumption over natural gas consumption. Also find the sample co-variance of this variable with the total variable in part a.

Let Y = The excess of petroleum consumption over natural gas, then $Y = X_1 - X_2 + 0X_3 + 0X_4$, and $C = (1, -1, 0, 0)^T$

$$Y = C^T \boldsymbol{X}, \quad E(Y) = C^T \boldsymbol{\mu} \approx C^T \overline{\boldsymbol{X}}, \quad Var(Y) = C^T \sum C \approx C^T \boldsymbol{S} C$$

```
c1_d<-matrix(c(1,-1,0,0),nrow = 4,ncol = 1) #X1-X2

mean.diff<- t(c1_d)%*%x.bar
var.diff<- t(c1_d)%*%S%*% c1_d


mean.diff

##       [,1]
## [1,] 0.258

var.diff<-t(c1_d) %*% ex1.7.s %*% c1_d
var.diff

##       [,1]
## [1,] 0.154
```

c. Find the variance-covariance matrix between Y1 = X1 + X2 and Y2 = X1 − X2.

Let $Y = (Y_1, Y_2)^T$, then $Y = C^T \boldsymbol{X},\ where\ C = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \end{pmatrix}^T$.

$$Var(Y) = C^T \sum C \approx C^T \boldsymbol{S} C$$

```
c<-matrix(c(1, 1, 0, 0, 1, -1, 0, 0), nrow=2)
t(c) %*% s %*% c

##       [,1]     [,2]
## [1,] 2.694    0.288
## [2,] 0.288    0.154
```

Cov($Y_1, Y_2) = 0.288$, which is very weak association.