HW1

Getong Zhong

2023-02-03

1.1

Correlation

```
hypo <-
  structure(list(individual = 1:10, sex = structure(c(2L, 2L, 2L,
    2L, 2L, 1L, 1L, 1L, 1L, 1L), .Label = c("Female", "Male"), class = "factor"),
    age = c(21L, 43L, 22L, 86L, 60L, 16L, NA, 43L, 22L, 80L),
    IQ = c(120L, NA, 135L, 150L, 92L, 130L, 150L, NA, 84L, 70L
    ), depression = structure(c(2L, 1L, 1L, 1L, 2L, 2L, 2L, 2L,
    1L, 1L), .Label = c("No", "Yes"), class = "factor"), health = structure(c(3L,
    3L, 1L, 4L, 2L, 2L, 3L, 1L, 1L, 2L), .Label = c("Average",
    "Good", "Very good", "Very poor"), class = "factor"), weight = c(150L,
    160L, 135L, 140L, 110L, 110L, 120L, 120L, 105L, 100L)), .Names = c("individual",
    "sex", "age", "IQ", "depression", "health", "weight"),
    class = "data.frame", row.names = c(NA, -10L))
hypo[is.na(hypo)] = 0
cor(hypo[,c(3,4,7)])
```

```
##                age         IQ      weight
## age     1.00000000 -0.2190872 -0.04901837
## IQ     -0.21908720  1.0000000 -0.08897870
## weight -0.04901837 -0.0889787  1.00000000
```

Covariance

```
cov(hypo[,c(3,4,7)])
```

```
##               age        IQ     weight
## age     812.67778 -350.0333  -28.33333
## IQ     -350.03333 3140.9889 -101.11111
## weight  -28.33333 -101.1111  411.11111
```

1.2

Correlation

```r
hypo <-
  structure(list(individual = 1:10, sex = structure(c(2L, 2L, 2L,
    2L, 2L, 1L, 1L, 1L, 1L, 1L), .Label = c("Female", "Male"), class = "factor"),
    age = c(21L, 43L, 22L, 86L, 60L, 16L, NA, 43L, 22L, 80L),
    IQ = c(120L, NA, 135L, 150L, 92L, 130L, 150L, NA, 84L, 70L
    ), depression = structure(c(2L, 1L, 1L, 1L, 2L, 2L, 2L, 2L,
    1L, 1L), .Label = c("No", "Yes"), class = "factor"), health = structure(c(3L,
    3L, 1L, 4L, 2L, 2L, 3L, 1L, 1L, 2L), .Label = c("Average",
    "Good", "Very good", "Very poor"), class = "factor"), weight = c(150L,
    160L, 135L, 140L, 110L, 110L, 120L, 120L, 105L, 100L)), .Names = c("individual",
    "sex", "age", "IQ", "depression", "health", "weight"),
    class = "data.frame", row.names = c(NA, -10L))
hypo
```

```
##    individual    sex age  IQ depression    health weight
## 1           1   Male  21 120        Yes Very good    150
## 2           2   Male  43  NA         No Very good    160
## 3           3   Male  22 135         No   Average    135
## 4           4   Male  86 150         No Very poor    140
## 5           5   Male  60  92        Yes      Good    110
## 6           6 Female  16 130        Yes      Good    110
## 7           7 Female  NA 150        Yes Very good    120
## 8           8 Female  43  NA        Yes   Average    120
## 9           9 Female  22  84         No   Average    105
## 10         10 Female  80  70         No      Good    100
```

```r
hypo[,'age'][is.na(hypo[,'age'])] = mean(hypo[,'age'],na.rm=TRUE)
hypo[,'IQ'][is.na(hypo[,'IQ'])] = mean(hypo[,'IQ'],na.rm=TRUE)
hypo[,'weight'][is.na(hypo[,'weight'])] = mean(hypo[,'weight'],na.rm=TRUE)
cor(hypo[,c(3,4,7)])
```

```
##               age         IQ     weight
## age     1.0000000 -0.1350426 -0.1040039
## IQ     -0.1350426  1.0000000  0.5223497
## weight -0.1040039  0.5223497  1.0000000
```

Covariance

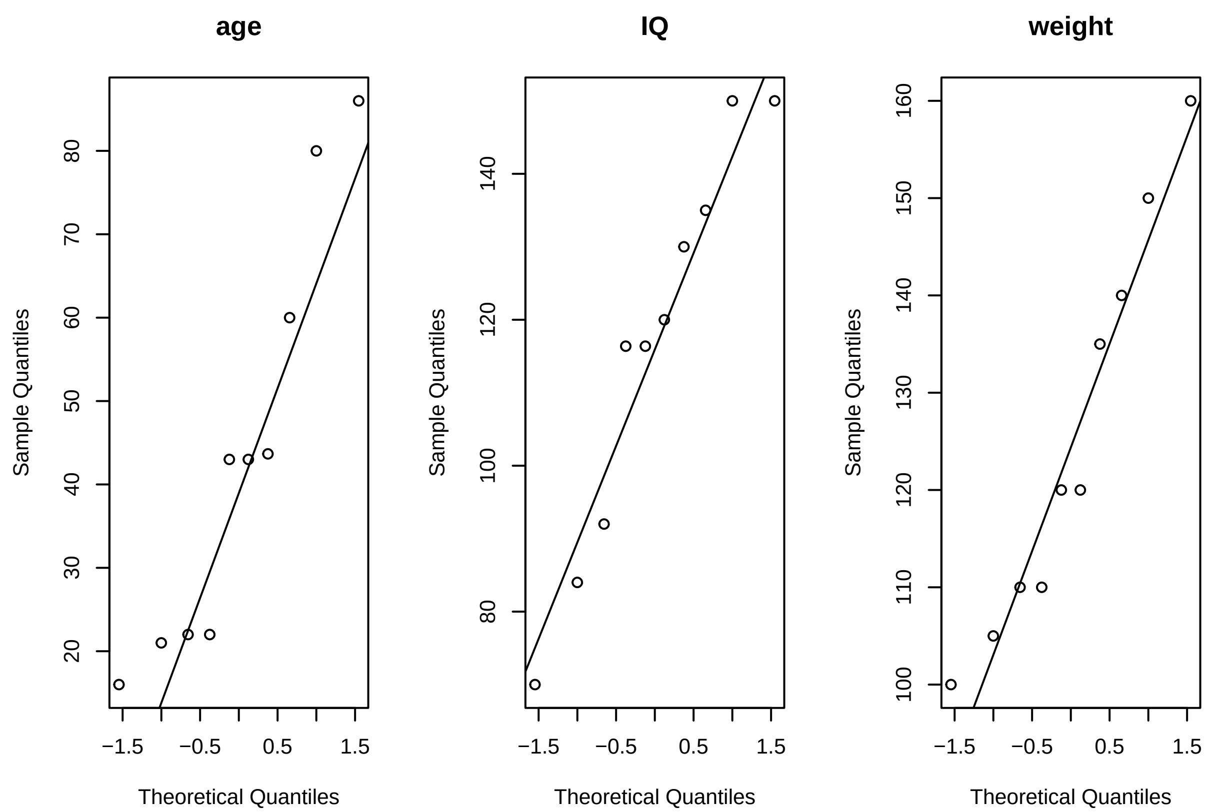
```r
cov(hypo[,c(3,4,7)])
```

```
##              age       IQ    weight
## age    622.00000 -91.2037 -52.59259
## IQ     -91.20370 733.3194 286.80556
## weight -52.59259 286.8056 411.11111
```

1.3

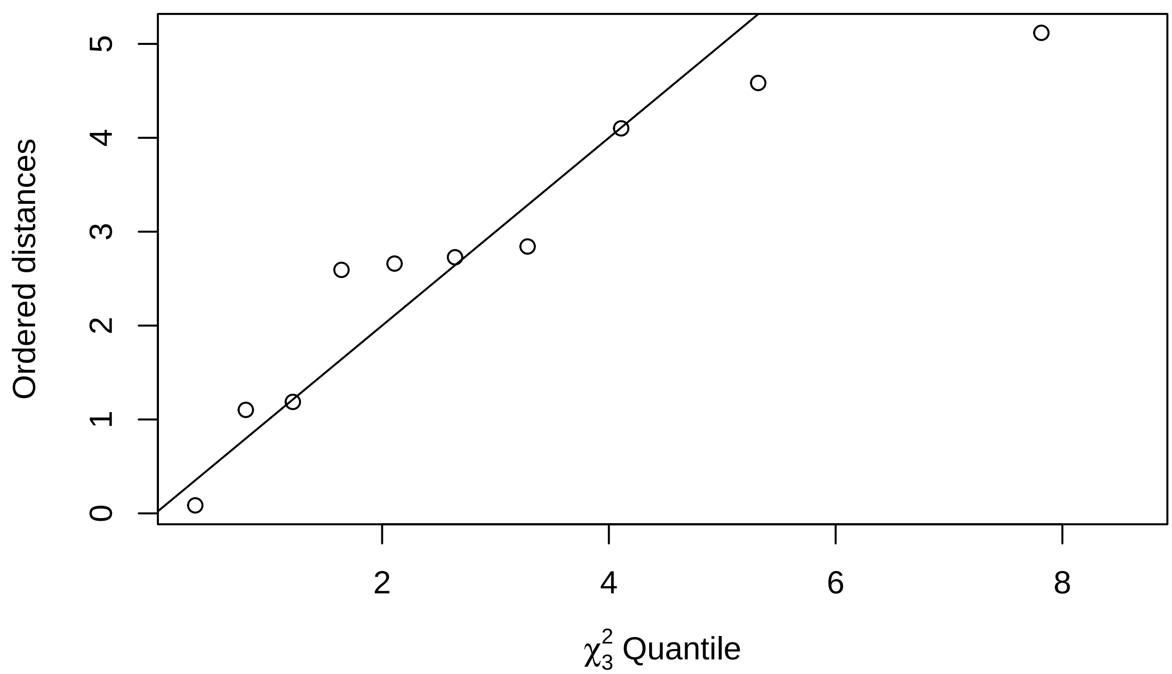
Normal Probablity plots

```
par(mfrow=c(1,3))
qqnorm(hypo[,'age'], main = "age"); qqline(hypo[,'age'])
qqnorm(hypo[,'IQ'], main = "IQ"); qqline(hypo[,'IQ'])
qqnorm(hypo[,'weight'], main = "weight"); qqline(hypo[,'weight'])
```



Chi-square plot

```
x<-hypo[,c(3,4,7)]
S <- cov(x)
cm <- colMeans(x)
d <- apply(x, MARGIN = 1, function(x) t(x - cm) %*% solve(S) %*% (x - cm))
plot(qc <- qchisq((1:nrow(hypo) - 1/2) / nrow(hypo), df = 3),
        sd <- sort(d),
        xlab = expression(paste(chi[3]^2, " Quantile")),
        ylab = "Ordered distances", xlim = range(qc) * c(1, 1.1))
abline(a = 0, b = 1)
```



Both the normal probability plots and the chi-square plot don't show much evidence of any departures from linearity. limited to the small size of the data, we cannot tell the exist of departures from linearity, but from the chi-square plot, there are certain sort of outlier exist.

1.4

```
V <- 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),
       4, 4, byrow = TRUE)

cov2cor(V)
```

```
##           [,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
```

1.5

Euclidean Distance Matrix

```r
M <- matrix(c(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),10,5,byrow = TRUE)
dist(M)
```

```
##            1         2         3         4         5         6         7
## 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
## 10  9.273618  8.774964  7.141428  7.874008  6.480741  6.164414  3.605551
##            8         9
## 2
## 3
## 4
## 5
## 6
## 7
## 8
## 9   7.416198
## 10  6.557439  5.830952
```

City Block Distance Matrix

```r
Y<-matrix(rowSums(M))
fn <- function(x, y) abs(x - y)
proxy::dist(Y, method = fn)
```

```
##     1  2  3  4  5  6  7  8  9
## 2   3
## 3   7 10
## 4   4  7  3
## 5   6  9  1  2
## 6  10 13  3  6  4
## 7   9 12  2  5  3  1
## 8   5  8  2  1  1  5  4
## 9   6  9  1  2  0  4  3  1
## 10  8 11  1  4  2  2  1  3  2
```