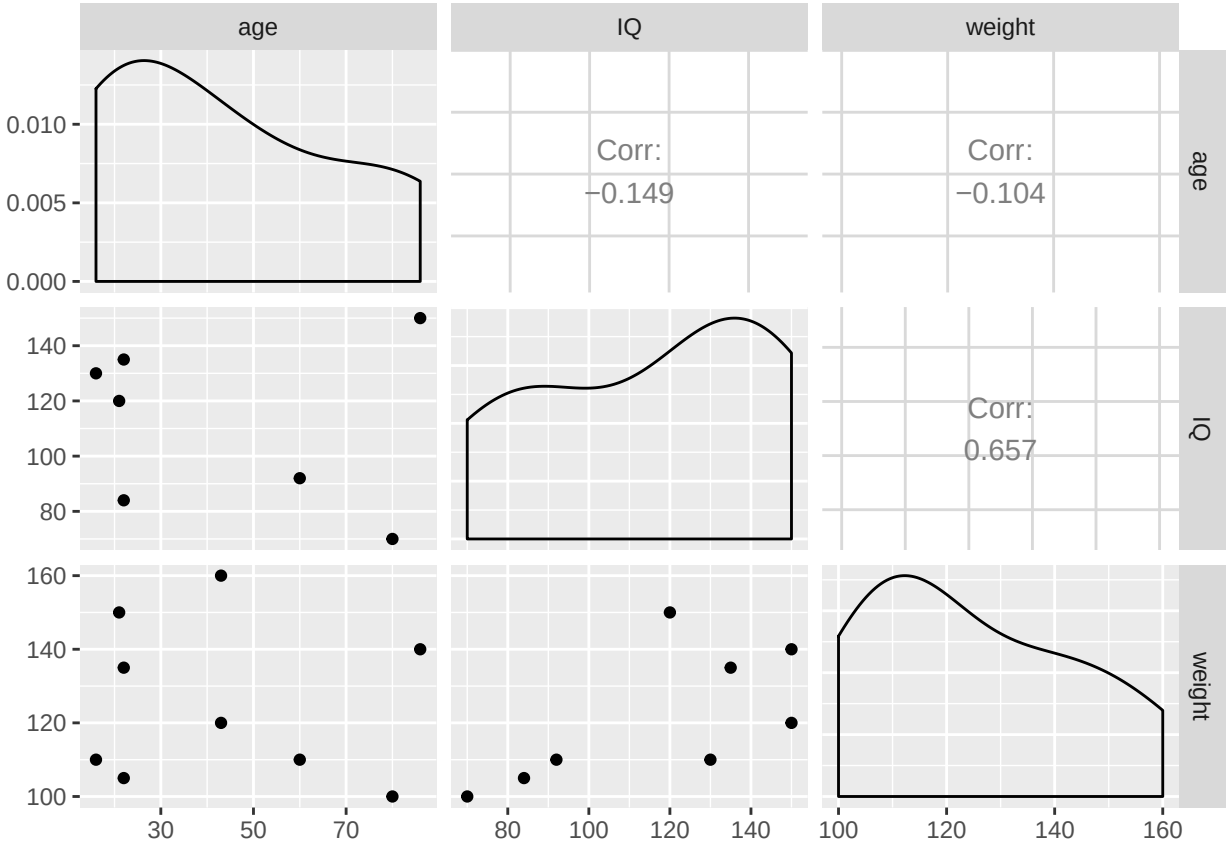
Homework 01

Cody Frisby

1/10/2017

1.1

Here I display the correlation matrix is a little different way.



And here I display the covariance matrix.

	age	IQ	weight
age	699.75000	-135.7381	-59.16667
IQ	-135.73810	942.8393	368.75000
weight	-59.16667	368.7500	411.11111

1.2

Covariance matrix:

	age	IQ	weight
age	622.00000	-91.2037	-52.59259
IQ	-91.20370	733.3194	286.80556

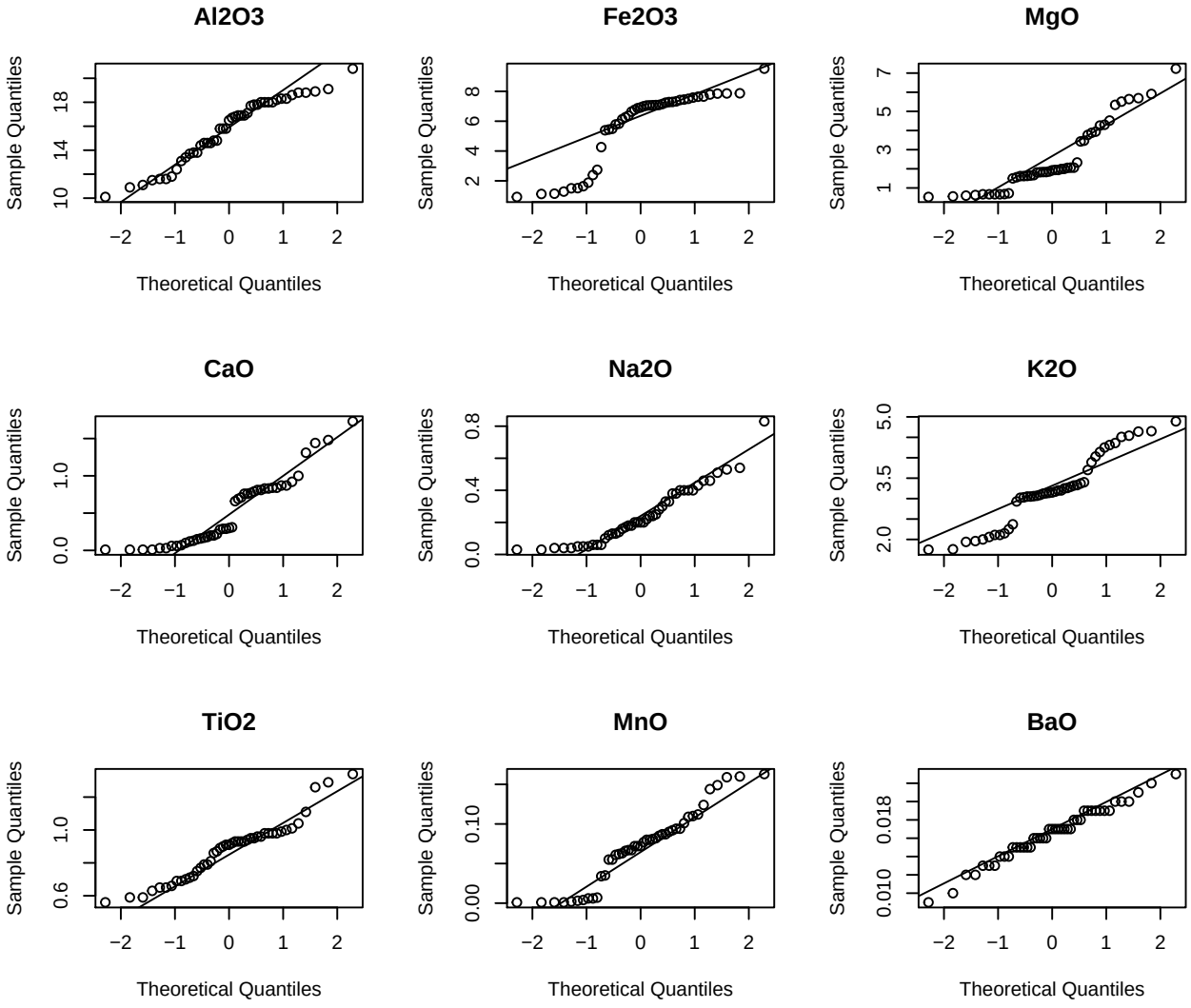
	age	IQ	weight
weight	-52.59259	286.8056	411.11111

Correlation matrix:

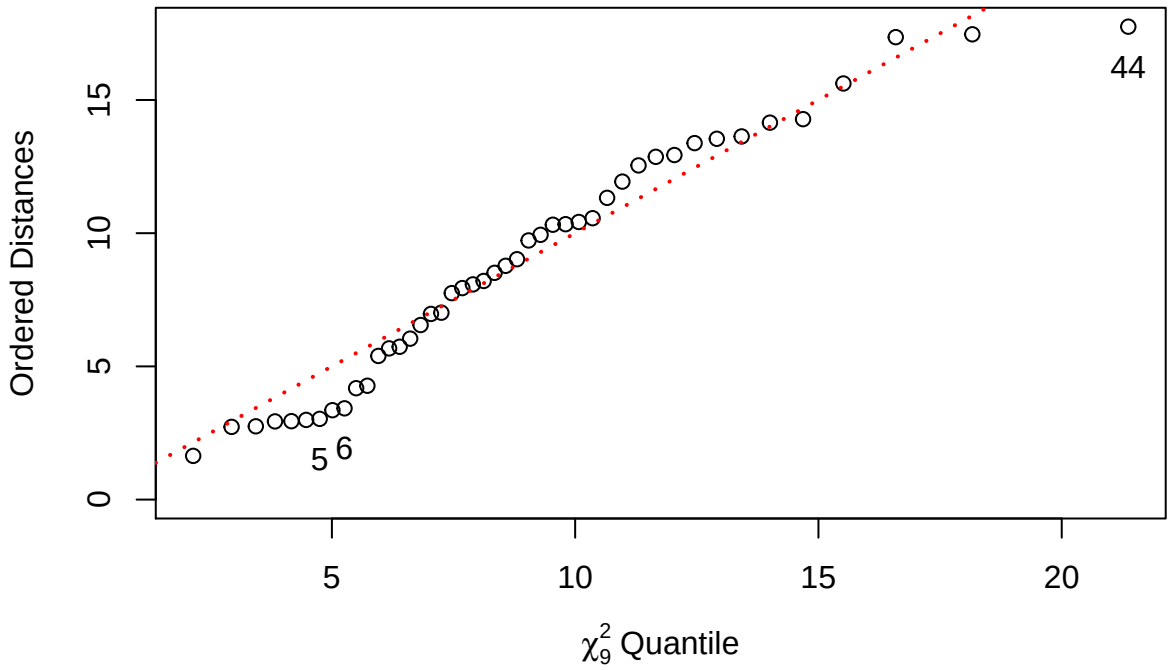
	age	IQ	weight
age	1.0000000	-0.1350426	-0.1040039
IQ	-0.1350426	1.0000000	0.5223497
weight	-0.1040039	0.5223497	1.0000000

1.3

Displaying the individual normality plots here:



And the χ^2 plot here:



Where

$$d_i^2 = (\mathbf{x_i} - \bar{x})^T \mathbf{S}^{-1}((\mathbf{x_i} - \bar{x}))$$

1.4

The correlation matrix is displayed here:

1.0000000	0.9801619	0.9810921	0.9899048
0.9801619	1.0000000	0.9552780	0.9807159
0.9810921	0.9552780	1.0000000	0.9670131
0.9899048	0.9807159	0.9670131	1.0000000

1.5

Using the R function

The *Euclidean distances* for the data are

	row	col	value
2	2	1	1.8963439
3	3	1	1.9877945
4	4	1	2.1075437
5	5	1	1.0307641
6	6	1	2.0817793
7	7	1	2.8365651
8	8	1	2.1474191
9	9	1	2.5759062
10	10	1	2.4590441
13	3	2	1.4807660
14	4	2	1.5738961
15	5	2	2.1404899

	row	col	value
16	6	2	1.9890028
17	7	2	1.9702631
18	8	2	1.0324001
19	9	2	1.9172372
20	10	2	2.1546544
24	4	3	1.4390458
25	5	3	2.0338251
26	6	3	0.9012348
27	7	3	1.9260890
28	8	3	1.2908882
29	9	3	1.1449371
30	10	3	1.9405586
35	5	4	2.1050455
36	6	4	1.1263614
37	7	4	2.1433675
38	8	4	1.6802059
39	9	4	1.5929515
40	10	4	2.0384740
46	6	5	1.9085728
47	7	5	2.2585764
48	8	5	1.9402005
49	9	5	2.4433786
50	10	5	1.7968371
57	7	6	2.0065174
58	8	6	1.7406923
59	9	6	1.0990388
60	10	6	1.7479204
68	8	7	1.1844039
69	9	7	1.7735962
70	10	7	0.9701138
79	9	8	1.7891913
80	10	8	1.6525151
90	10	9	1.5477036

and the city block, or `manhattan`, distances for the data are

	row	col	value
2	2	1	3.635607
3	3	1	3.128311
4	4	1	4.173991
5	5	1	1.971925
6	6	1	3.720347
7	7	1	5.715635
8	8	1	4.167495
9	9	1	5.389316
10	10	1	4.820562
13	3	2	2.728914
14	4	2	2.788201
15	5	2	4.544658
16	6	2	3.906491
17	7	2	4.029574

	row	col	value
18	8	2	1.988237
19	9	2	3.818840
20	10	2	3.760045
24	4	3	2.789964
25	5	3	3.066830
26	6	3	1.803120
27	7	3	3.682823
28	8	3	2.627880
29	9	3	2.261006
30	10	3	4.038837
35	5	4	4.128373
36	6	4	2.065581
37	7	4	3.937906
38	8	4	3.328572
39	9	4	3.156388
40	10	4	3.668377
46	6	5	3.273875
47	7	5	3.743710
48	8	5	3.181964
49	9	5	4.357304
50	10	5	3.788550
57	7	6	2.935200
58	8	6	3.328122
59	9	6	2.146304
60	10	6	3.291214
68	8	7	2.041337
69	9	7	3.324883
70	10	7	1.881466
79	9	8	3.240472
80	10	8	3.337264
90	10	9	3.187700

1.6

a)

There are 4 observations, or units, in the data set, and there are 4 variables - *reciept ID, number of books, dollars of sales*, and whether the customer was a *student* or not. There are categorical and continuous types of variables in the data set. Of the categorical set, *students*, there are two levels: **yes** and **no**.

b)

recieptID	books	sales	student
101	4	142	yes
521	5	252	yes
746	4	148	no
857	3	158	yes

c)

Here I display rows 2 and 4 of the data frame above.

$$X_2 = 521, 5, 252, yes$$
$$X_4 = 857, 3, 158, yes$$

1.7

a)

For reference: $x_1 = petroleum,\ x_2 = natural\ gas,\ x_3 = hydroelectric\ power,\ x_4 = nuclear\ electric\ power.$

The mean of a state's total energy consumption is simply the sum of the means of the 4 energy sources, $\sum \bar{\mathbf{x}}_\mathbf{i} = 1.873$. The variance of a state would be the sum of the variances plus 2 times the covariance terms.

$$\sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \sigma_4^2 + 2cov_{12} + 2cov_{13} + ... + 2cov_{34} = 3.912$$

or using matrix notation

$$\mathbf{x}^T\mathbf{S}\mathbf{x}$$

where x is a unit vector of length 4.

b)

We are asked to find the mean of $x_1 - x_2$ as well as the cov_{12}.

$$x_1 - x_2 = 0.766 - 0.508 = 0.258$$

We can find the covariance of these two variables by taking the 2nd row 1st column term from the supplied matrix, S, $cov_{12} = 0.635$.

c) $Y_1 = X_1 + X_2$ **and** $Y_2 = X_1 - X_2$

The variance is $0.856 + 0.568 + 2 \times 0.635 = 2.694 = \sigma^2_{Y_1}$. And $\sigma^2_{Y_2} = 0.856 + 0.568 - 2 \times 0.635 = 0.154$.

And the variance-covariance matrix would be

2.694	0.288
0.288	0.154

R Code:

```
### 1.1
df <- read.csv("~/Documents/STAT4400/dat/hypo.csv")
df <- df[-1]
n <- sapply(df, class) == "integer" | sapply(df, class) == "numeric"
cov.mat <- cov(df[n], use = "pairwise.complete.obs")
```

```r
cor.mat <- cor(df[n], use = "pairwise.complete.obs")
# or even better :)
library(ggplot2); library(GGally)
ggpairs(df[, c("age", "IQ", "weight")])
### 1.2
hypo.update <- df
hypo.update$age <- ifelse(is.na(df$age),
                          mean(df$age, na.rm = TRUE), df$age)
hypo.update$IQ <- ifelse(is.na(df$IQ),
                          mean(df$IQ, na.rm = TRUE), df$IQ)
cov.mat <- cov(hypo.update[n])
cor.mat <- cor(hypo.update[n])
### 1.3
df <- read.csv("~/Documents/STAT4400/dat/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))
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"
### 1.4
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 <- sqrt(diag(x))
y <- outer(d, d, "*")
cor.mat <- x/y
### 1.5
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(scale(x, center = FALSE))
library(reshape2)
df <- melt(as.matrix(d), varnames = c("row", "col"))
d <- dist(scale(x, center = FALSE), method = "manhattan")
df <- melt(as.matrix(d), varnames = c("row", "col"))
### 1.6
df <- data.frame(recieptID = c(101,521,746,857), books = c(4,5,4,3),
                 sales = c(142,252,148,158),
                 student = c("yes", "yes", "no", "yes"))
### 1.7
xbar <- c(.766, .508, .438, .161)
S <- matrix(c(.856,.635,.173,.096,.635,.568,.127,.067,.173,.128,.171,.039,
```

```
              .096,.067,.039,.043), ncol = 4)
v <- sum(diag(S)) + 2*sum(S[lower.tri(S)])
m.y1 <- xbar[1] + xbar[2]
m.y2 <- xbar[1] - xbar[2]
s.y1 <- S[1,1] + S[2,2] + 2 * S[2,1]
s.y2 <- S[1,1] + S[2,2] - 2 * S[2,1]
cv <- S[1,1] - S[2,2]
m <- matrix(c(s.y1,cv ,cv ,s.y2), ncol = 2)
```