实验三：利用 R 软件单变量和多变量正态检验和置信区域等

沈雨萱 3180104691

目录

1 实验概况

一．实验目的与要求：通过本试验项目，能够理解并掌握如下内容 (1) 单变量和多变量正态检验；(2) 多变量均值向量显著性检验；(3) 置信域和置信区间计算，画置信椭球等

二．实验内容 (I) 采用实验二 sample 样本。附表中的数据 sample.xls 进行分析。记 X1=BMI, X2=FPG, X3=SBP, X4=DBP, X5=TG, X6=HDL-C，并构成一个向量。X=(X1, X2, X3, X4, X5, X6), 详细分析患代谢综合症的群体与没有患代谢综合症群的差异分析。（任选一项）1. 分析患代谢综合症的年龄差异 2. 分析患代谢综合症的性别差异 3. 分析是否吸烟对患代谢综合症的影响 4. 分析是否喝酒对患代谢综合症的影响提示分析内容：(a) 数据预处理 (b) 检验相关数据正态性，相关性 (c) 分析人群患代谢综合症的比例 (d) 计算患代谢综合症的群体与没有患代谢综合症群体各类指标（体重指数、血压、血指、血糖等等指标的均值和置信区间分析差异。

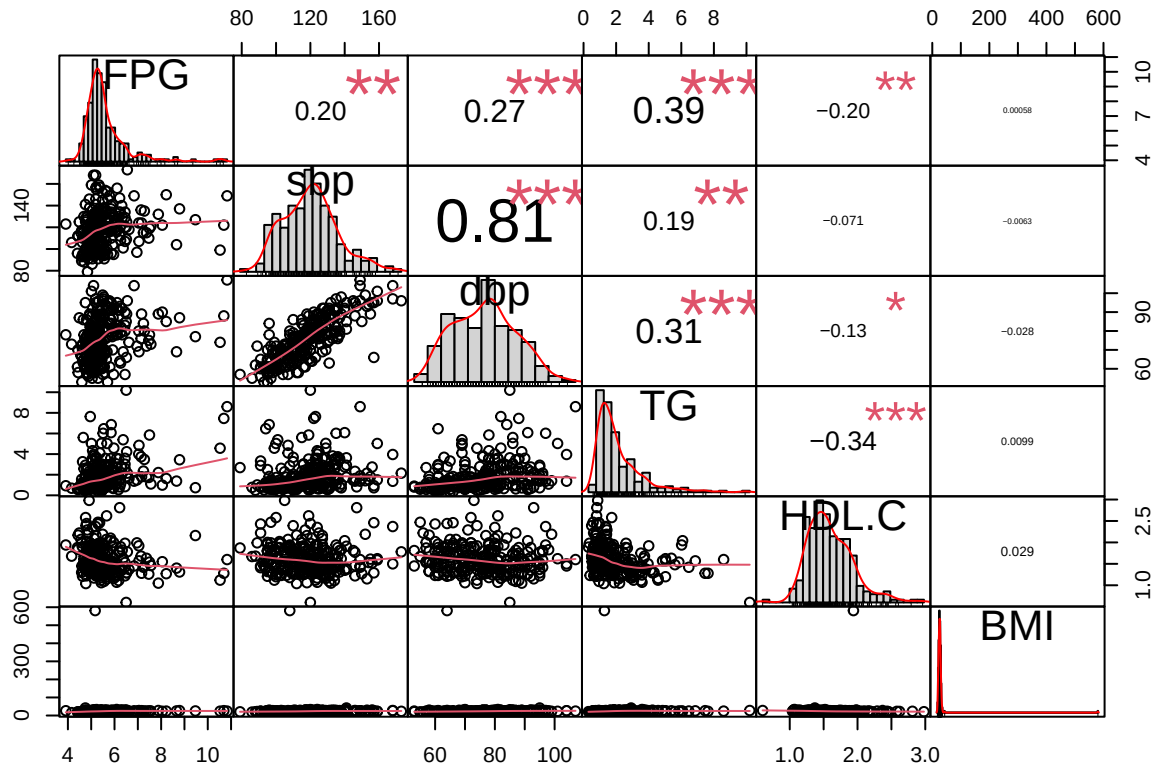
(II) 数据 ex2.1：给出了 27, 名糖尿病人血清总胆固醇 (x1), 甘油（x2）, 空腹胰岛素 (x3), 糖化血红蛋白 (x4), 空腹血糖 (y) 的测量值。（1）试建立血糖 (y) 与其他指标的线性回归方程，并进行分析；（2）（x1, x2, x3, x4）是否服从多元正态？（x1,x2）与（x3,x4）是否相互独立？

2 实验结果

2.1 (I)

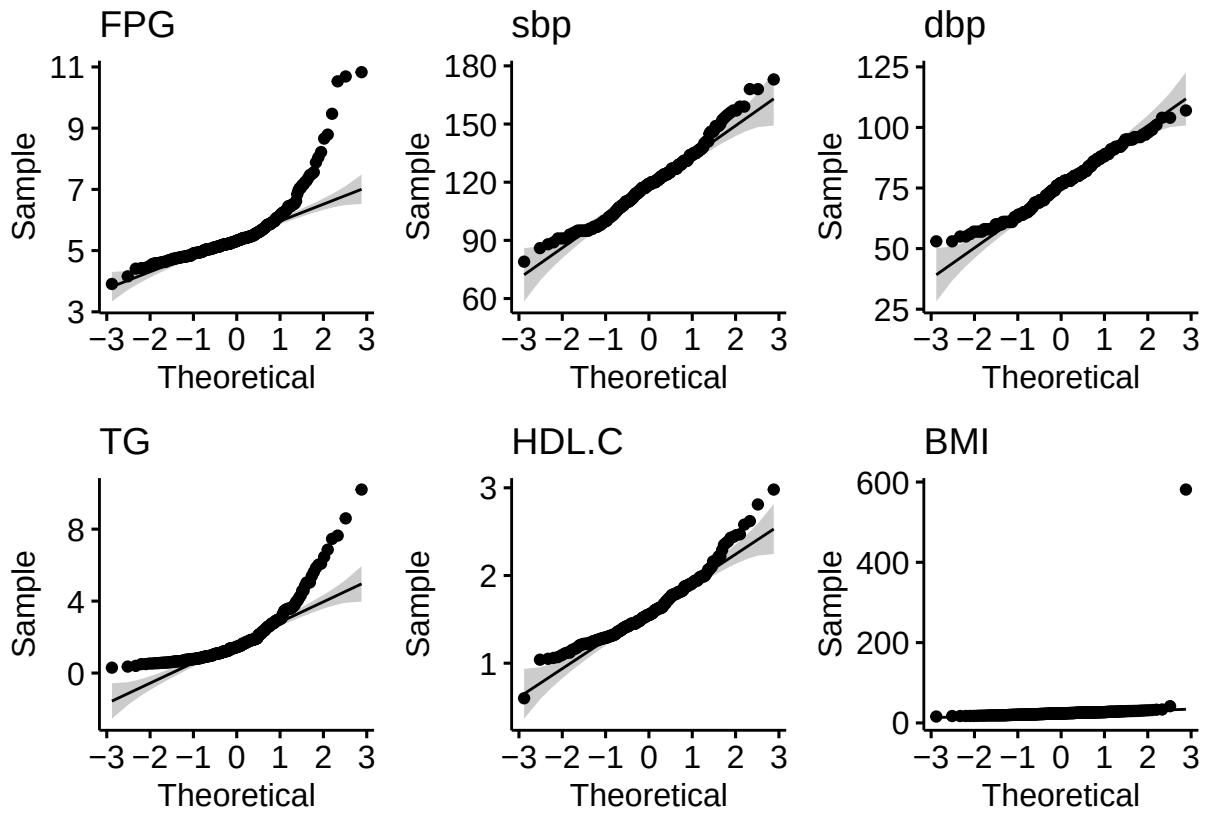
a. 患代谢综合症的群体与没有患代谢综合症群的差异分析。

```
data <- read.csv("sample.csv",encoding = "UTF-8",na.strings=c("","  ","NA"))
X <- data[,c('weight','height','FPG','sbp','dbp','TG','HDL.C')]
# 去除缺失值
X <- na.omit(X)
# 计算 BMI 并删去 weight 和 height
X$'BMI'=X$weight/(X$height*X$height)*10000
X <- X[ , !names(X) %in% c("weight", "height")]
# 计算相关性
r <- rcorr(as.matrix(X))
chart.Correlation(X, histogram=TRUE, pch=19)
```




```
# 正态性检验
c1 <- ggqqplot(X$FPG,main='FPG')
c2 <- ggqqplot(X$sbp,main='sbp')
c3 <- ggqqplot(X$dbp,main='dbp')
c4 <- ggqqplot(X$TG,main='TG')
c5 <- ggqqplot(X$HDL.C,main='HDL.C')
c6 <- ggqqplot(X$BMI,main='BMI')
```

```
c1+c2+c3+c4+c5+c6
```



可以看到 FPG 和 TG 不太符合正态分布，而 sbp,dbp,HDL.C,BMI 较符合正态分布

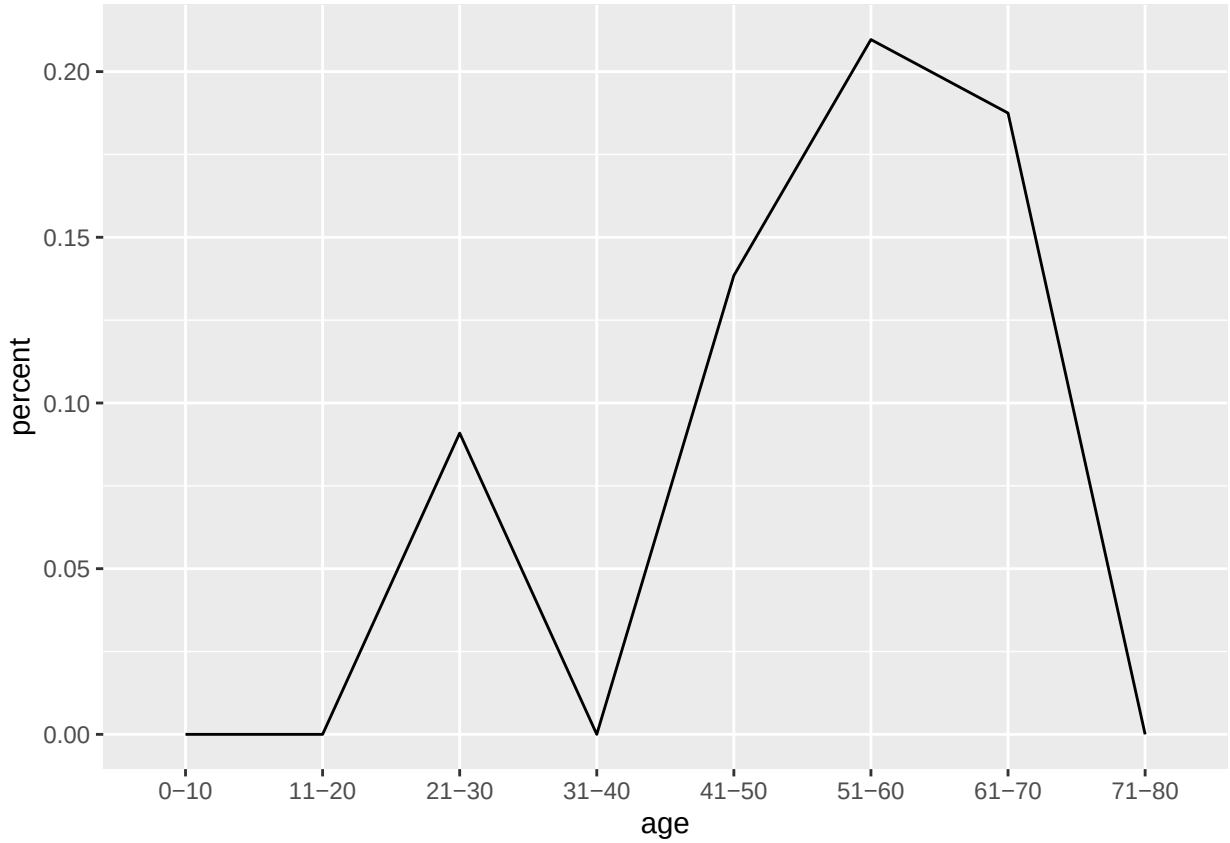
```
data <- read.csv("sample.csv",encoding = "UTF-8",na.strings=c("","  ","NA"))
# 筛选代谢综合征的病人
data <- na.omit(data)
data$'BMI'=data$weight/(data$height*data$height)*10000
data$'disease'=0
sick <- ((data[, 'BMI'] >= 25) + (data[, 'FPG'] >=6.1)+((data[, 'sbp'] >= 140)|(data[, 'dbp'] >= 9
data[which(sick == TRUE),'disease']=1
# 调整异常值
unique(data$smoke)
```

```
## [1] "是"        "否"        "已戒烟"    "戒烟2个月" "戒烟3年"
```

```
unique(data$drunk)
```

```
## [1] "是" "无" "否"
```

```
data$smoke <- gsub( " 已戒烟", " 否",data$smoke)
data$smoke <- gsub( " 戒烟 2 个月", " 是",data$smoke)
data$smoke <- gsub( " 戒烟 3 年", " 否",data$smoke)
data$drunk <- gsub( " 无", " 否",data$drunk)
# 计算不同年龄组人群中患代谢综合征的比例
age_range <- c('0-10','11-20','21-30','31-40','41-50','51-60','61-70','71-80')
p <- c()
for (i in 1:8){
  nd=nrow(data %>% filter(age>=(i-1)*10+1 & age<=i*10 & disease==1))
  na=nrow(data %>% filter(age>=(i-1)*10+1 & age<=i*10 ))
  p <- c(p,nd/na)
}
df <- data.frame(x = age_range, y = p)
ggplot(df, aes(x =x, y = y, group = 1)) + geom_line() + xlab('age')+ylab('percent')
```




```
# 显著性检验
age_d <- filter(data,disease==1)$age
age_h <- filter(data,disease==0)$age
```

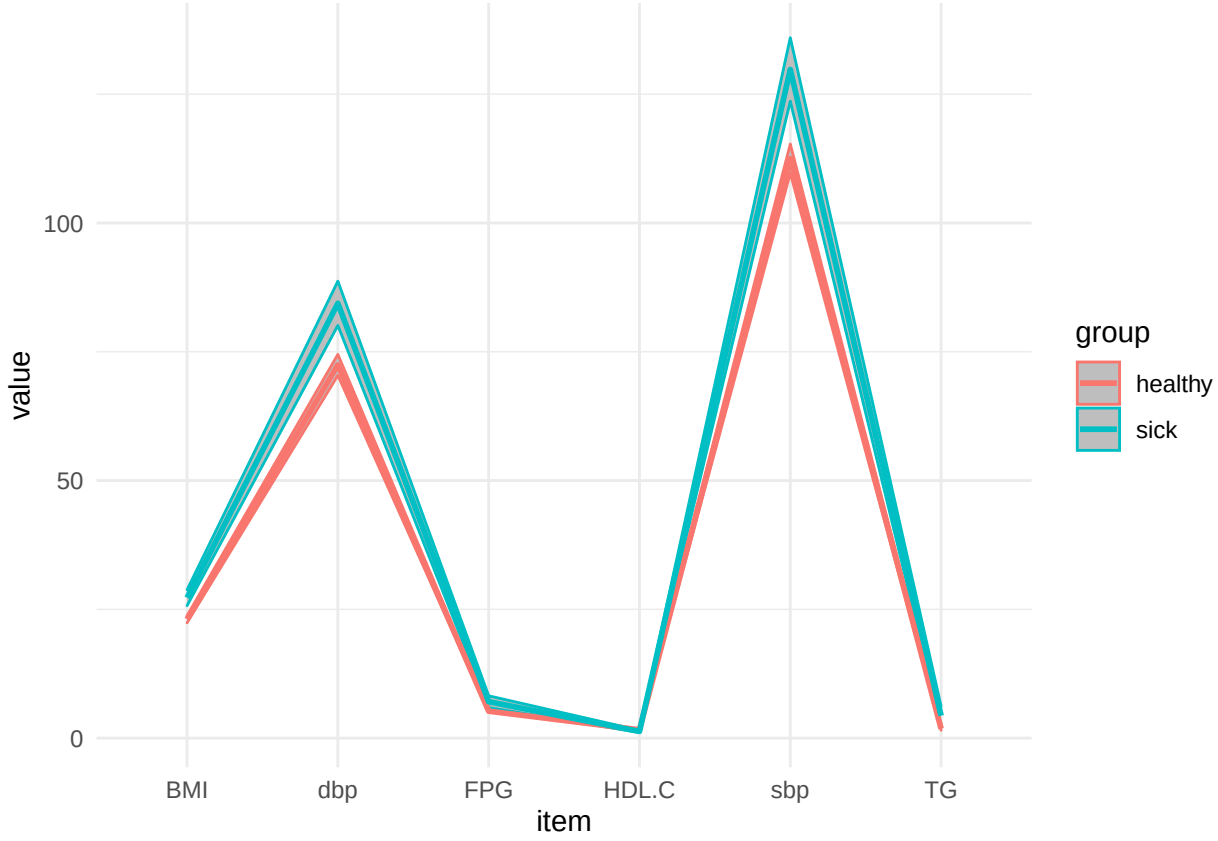
```
t.test(age_d,age_h, alternative = c("two.sided", "less", "greater"),mu = 0,paired = FALSE, var.equ
```

```
##
##  Welch Two Sample t-test
##
## data:  age_d and age_h
## t = 2.5592, df = 40.539, p-value = 0.01432
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.019159 8.660016
## sample estimates:
## mean of x mean of y
##  51.46154  46.62195
```

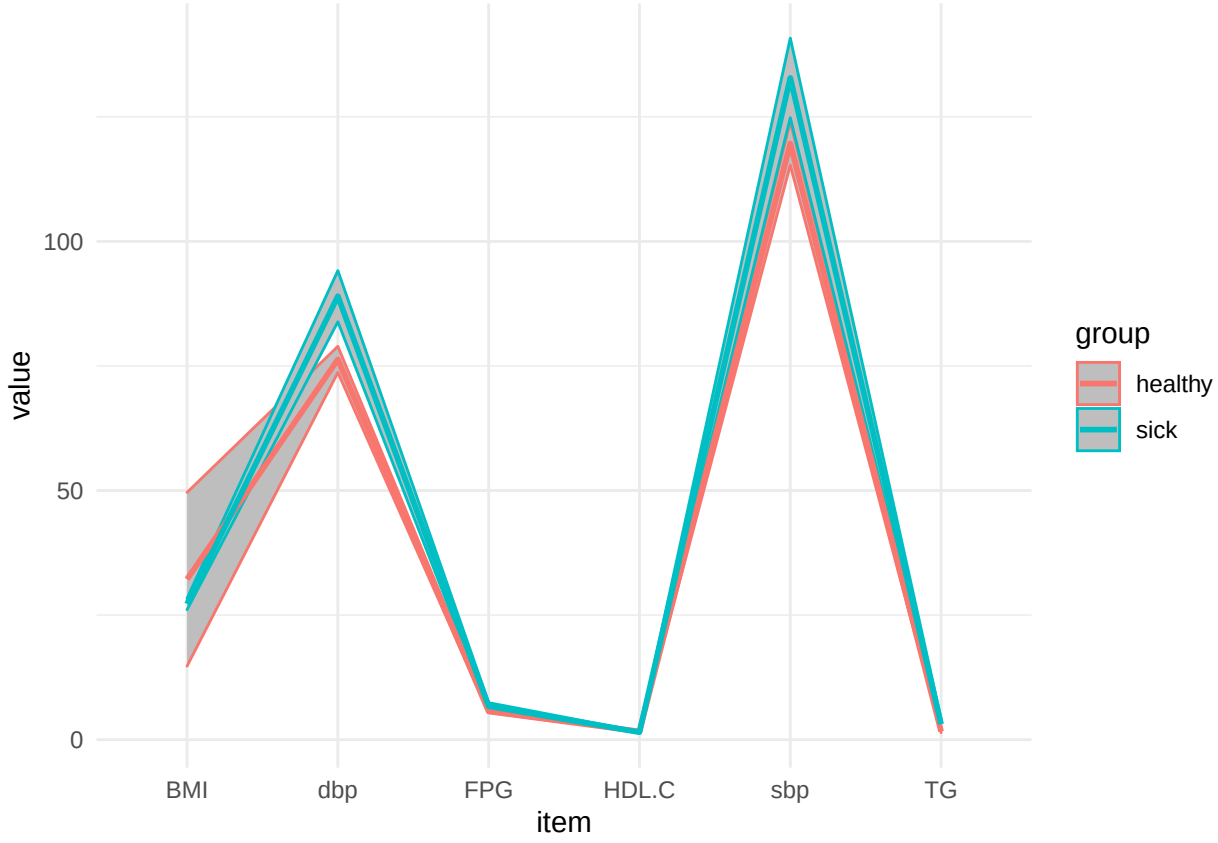
可以看到 41-70 岁人群中得代谢病得比例较高。在显著性水平 =0.05 的情况下得病与不得病人群得年龄存在显著性差异

```
# 根据之前结果，把人群分为 0-50，51-80 两个年龄组
# 计算两个组是否患代谢综合症群体各类指标的均值估计和置信区间或区域
a <- c()
items <- c('BMI','FPG','sbp','dbp','TG','HDL.C')
sick <- list(mean=a,low=a,high=a)
healthy <-list(mean=a,low=a,high=a)
#0-50 组
nd=data %>% filter(age<=50 & disease==1)
nh=data %>% filter(age<=50 & disease==0)
count=1
for (i in items){
  sick$mean[count] <- colMeans(nd[i])
  sick$low[count] <- t.test(nd[i],conf.level = 0.95)$conf.int[1]
  sick$high[count] <- t.test(nd[i],conf.level = 0.95)$conf.int[2]
  healthy$low[count] <- t.test(nh[i],conf.level = 0.95)$conf.int[1]
  healthy$high[count] <- t.test(nh[i],conf.level = 0.95)$conf.int[2]
  healthy$mean[count] <- colMeans(nh[i])
  count=count+1
}
r <- data.frame(item = items, value=c(sick$mean,healthy$mean),
ci_lower = c(sick$low,healthy$low),
ci_upper = c(sick$high,healthy$high),
```

```
group=c(rep('sick',times=6),rep('healthy',times=6)))
ggplot(r,aes(item, value,group = group,color=group)) +
  geom_ribbon(aes(ymin = r$ci_lower,
                  ymax = r$ci_upper),
              fill = "grey") +
  geom_line(size = 1)+
  theme_minimal()
```



可以看到对于 0-50 岁得病群体和健康群体，其 BMI,dbp 和 sbp 项的均值估计差异较大。下面针对 51-80 群体作图



可以看到对于 51-80 岁得病群体和健康群体，sbp 项与 BMI 项的均值估计差异没有 0-51 岁那么显著，推测可能高龄群体得代谢病会受更多因素影响。

2.2 (II)

数据 ex2.1：给出了 27, 名糖尿病人血清总胆固醇 (x1), 甘油（x2）, 空腹胰岛素 (x3), 糖化血红蛋白 (x4), 空腹血糖 (y) 的测量值。（1）试建立血糖 (y) 与其他指标的线性回归方程，并进行分析；

```
data2 <- read.csv("ex2.1.csv",encoding = "UTF-8",na.strings=c("","  ","NA"))
data2 <- na.omit(data2)
fit <- lm(y ~ x1+x2+x3+x4,data=data2)
summary(fit)
```

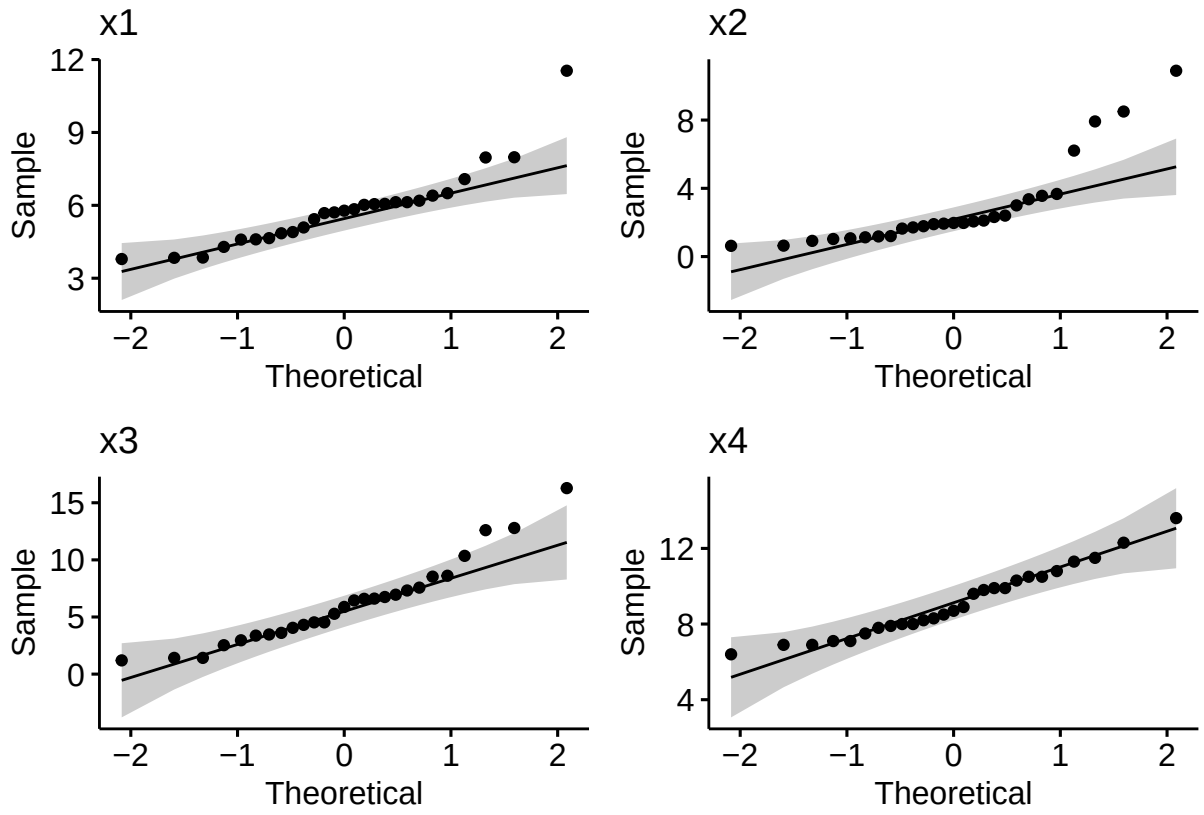
```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -3.6268 -1.2004 -0.2276  1.5389  4.4467
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.9433     2.8286   2.101   0.0473 *
## x1            0.1424     0.3657   0.390   0.7006
## x2            0.3515     0.2042   1.721   0.0993 .
## x3           -0.2706     0.1214  -2.229   0.0363 *
## x4            0.6382     0.2433   2.623   0.0155 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.01 on 22 degrees of freedom
## Multiple R-squared:  0.6008, Adjusted R-squared:  0.5282
## F-statistic: 8.278 on 4 and 22 DF,  p-value: 0.0003121
```

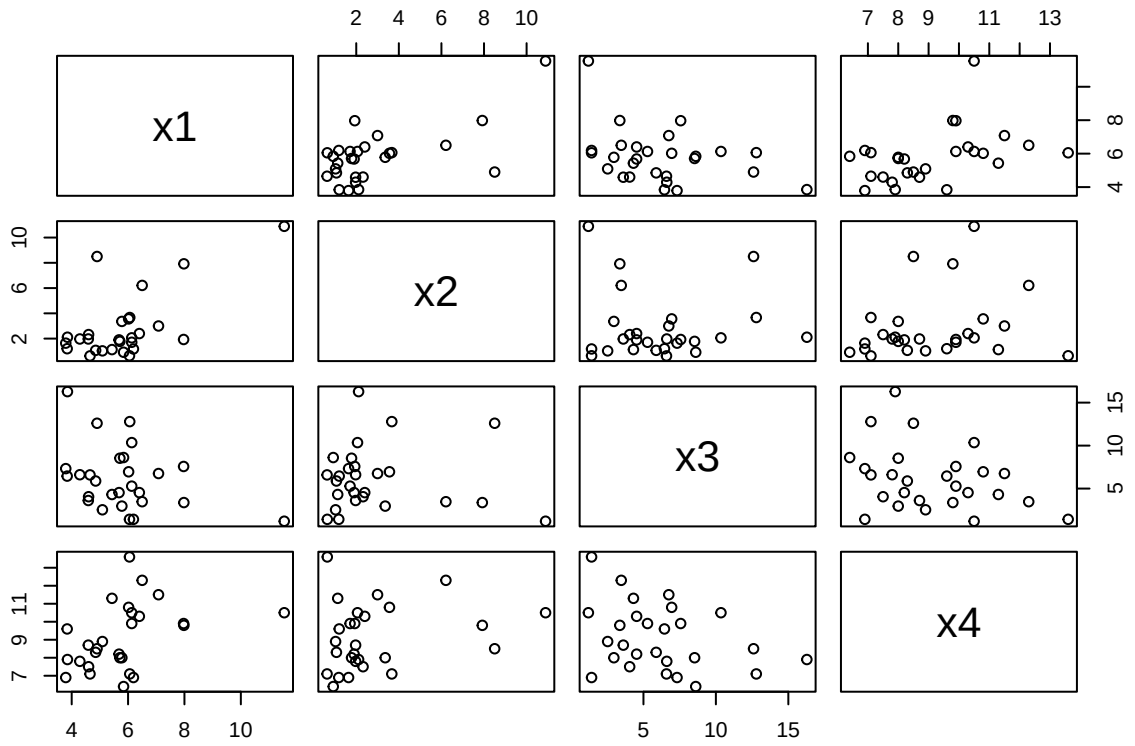
根据回归结果得到得线性回归方程为 y=0.1424x1+0.3515x2-0.2706x3+0.6382x4 根据 p-value, 该线性回归方程拟合效果较差。

（2）（x1, x2, x3, x4）是否服从多元正态？（x1,x2）与（x3,x4）是否相互独立？

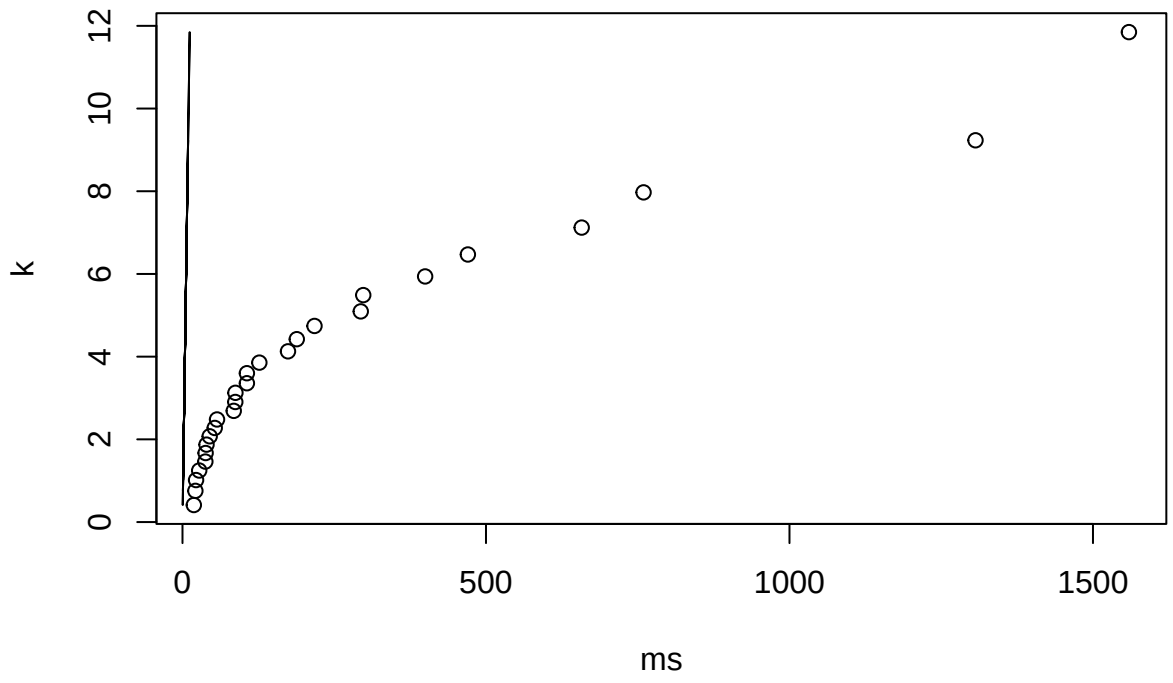
```
# 首先进行一元正态的评估
c1 <- ggqqplot(data2$x1,main='x1')
c2 <- ggqqplot(data2$x2,main='x2')
c3 <- ggqqplot(data2$x3,main='x3')
c4 <- ggqqplot(data2$x4,main='x4')
c1+c2+c3+c4
```




```
# 线性关系检验
pairs(data2[,2:5])
```




```
# 多元正态卡方统计量的 Q-Q 图检验法
mx<-t(as.matrix(data2[,2:5]))# 将数据框转化成矩阵，给矩阵转置，变成一列有 P 个变量的矩阵，一列就是一个对
m<-matrix(colMeans(data2[,2:5]))# 计算均值向量
d<-apply(mx,2,function(x){x-m})# 每个观测值减去均值向量
ms<-apply(d,2,function(x){t(x)%*%cov(data2[,2:5])%*%x})# 每个观测值的马氏距离
# 绘制 QQ 图
p<- (rank(ms)-0.5)/length(ms)# 算累积概率
k<-qchisq(p,df=4)# 根据卡方分布的概率求出 27 个分位点，一个累积概率求出一个点
plot(ms,k)
lines(k,k)
```



根据卡方统计量 Q-Q 检验法可以看出 (X1,X2,X3,X4) 不符合多元正态分布

```
# 进行独立性检验
md <- as.matrix(data2[,2:5])
mm <- diag(1,27) - matrix(1,27,27)/27
# 样本离散阵
A <- t(md)%*%mm%*%md
p=4
p1=p2=2
n=27
b = n-1.5-(p^3-sum(p1^3+p2^3))/(3*(p^2-sum(p1^2+p2^2)))
f = 0.5*(p*(p+1)-sum(p1^2+p1+p2^2+p2))
V=det(A)/(det(A[1:2,1:2])*det(A[3:4,3:4]))
ep = -b*log(V)
ep
```

```
## [1] 7.602315
```

```
qchisq(0.95,f)
```

```
## [1] 9.487729
```

因为 7.602315 < 9.487729 因此接受假设，(X1,x2) 与 (X3,X4) 独立