实验二：利用 R 软件实现多元数据进行可视化和参数估计

沈雨萱 3180104691

目录

1 实验概况

一. 实验目的与要求：通过本试验，实现下列目标：（1）多元数据分析图示，轮廓图、雷达图、调和曲线图和散布图矩阵；（2）能够利用 R 求解多元正态随机向量的均值、协方差、样本相关矩阵等参数的极大似然估计；（3）能够对简单时间序列参数估计的效果进行估计。

二. 实验内容 (1) 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)。a. 分析 X 各变量之间的相关性？b. 分析患代谢综合症的比例有没有性别差异，与吸烟或喝酒是否有关？c. 利用多元数据分析图给出 20~30 年龄段，X 各个指标的分布情况。d. 给出总体 X 的均值、协方差矩阵和相关矩阵的估计。

(2) 假设 Yt 服从下面的模型：Y0=0, Yt= Yt-1 + t, t=1, 2, ..., n。利用随机模拟分析下面两种情况下，对参数 估计量进行分析 (如通过均方误差 (MSE) 来说明)。（1） 的真值为 0.6；（2） 的真值为 1。

2 实验结果

2.1 附表分析

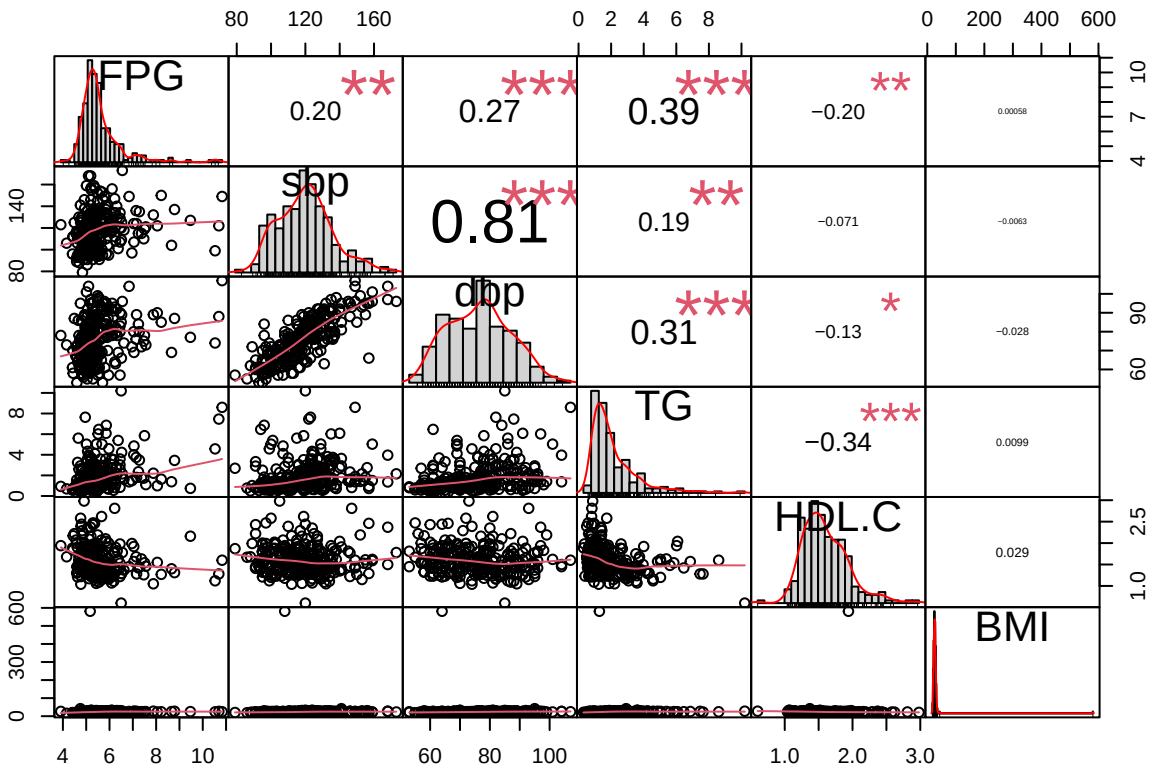
有代谢综合症的人, 其罹患心血管疾病、脑血管疾病及肾脏疾病的危险比没有代谢综合症的人高, 因此代谢综合症的预防与治疗, 是目前临床医学及基础研究关注的主题. 中华医学会糖尿病学分会（CDS）建议代谢综合征的诊断标准: 具备以下 4 项中的 3 项及以上即为代谢综合症: (1) 超重: BMI>= 25.0 Kg/M^2(体

重/身高平方) ; (2) 高血糖:FPG>= 6.1mmol/L(110mg/dl) 或 2hPG>=7.8 mmol/L(140mg/dl), 或已确诊糖尿病并治疗者; (3) 高血压: 收缩压 SBP >=e 140 mmHg 或舒张压 DBP>= 90mmHg, 或已确诊高血压并治疗者; (4) 空腹血: 甘油三脂 TG>=1.7 mmol/L(110mg/dl) 或 HDL-C <0.9 mmol/L(35 mg/dl)（男）, <1.0 mmol/L(39 mg/dl) （女）.

代谢综合征的发病机制至今为止还不甚清楚, 但可以明确直接发病的原因是胰岛素抵抗, 与不良的饮食习惯 (如经常抽烟、喝酒等) 和生活方式 (如缺乏运动) 密切相关. 为了进一步研究代谢综合征影响因素, 现收集了某个地区的体检资料。见 sample.xls.

a. 分析 X 各变量之间的相关性

```
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)
```



b. 分析患代谢综合症的比例有没有性别差异，与吸烟或喝酒是否有关？

```
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"
```

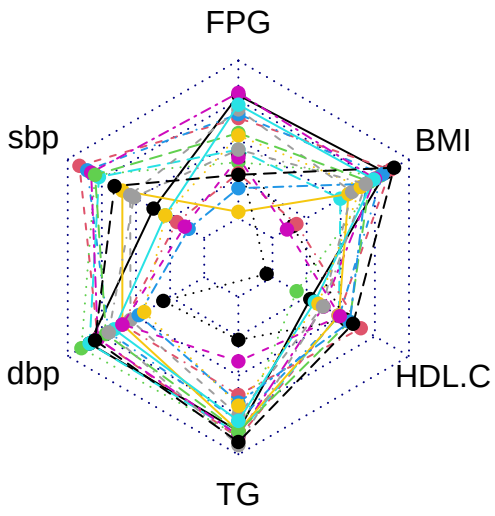
```
data$smoke <- gsub( " 已戒烟", " 否",data$smoke)
data$smoke <- gsub( " 戒烟 2 个月", " 是",data$smoke)
data$smoke <- gsub( " 戒烟 3 年", " 否",data$smoke)
data$drunk <- gsub( " 无", " 否",data$drunk)
gender <- c(sum(((data[, 'gender']=='男')&(data[, 'disease']==1))==TRUE)/sum((data[, 'gender']=='男
smoke <- c(sum(((data[, 'smoke']=='是')&(data[, 'disease']==1))==TRUE)/sum((data[, 'smoke']=='是')=
drunk <- c(sum(((data[, 'drunk']=='是')&(data[, 'disease']==1))==TRUE)/sum((data[, 'drunk']=='是')=
show <- data.frame(gender,smoke,drunk)
show
```

```
##       gender      smoke      drunk
## 1 0.20661157 0.22950820 0.24000000
## 2 0.01449275 0.09302326 0.06956522
```

可以看到，患代谢病比例与男女、吸烟与否、饮酒与否有较大关系，男性，吸烟者，饮酒者中患代谢病的人比例更高。

c. 利用多元数据分析图给出 20~30 年龄段，X 各个指标的分布情况筛选并绘制雷达图

```
data <- read.csv("sample.csv",encoding = "UTF-8",na.strings=c("","  ","NA"))
X <- data[,c('age','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
X3<- filter(X,age>=20&age<=30)
X3 <- X3[ , !names(X3) %in% c("weight", "height","age")]
X3.2<-rbind(c(4,90,50,0,0,15),c(6,125,100,5,3,35),X3)
radarchart(X3.2)
```



d. 给出总体 X 的均值、协方差矩阵和相关矩阵的估计

```
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
X3 <- X3[ , !names(X3) %in% c("weight", "height","age")]
rmean <- vector()
for (i in 1:ncol(X3))
{
  rmean = c(rmean,mean(X3[,i]))
}
# 显示平均值，协方差矩阵，相关矩阵
rbind(names(X3),rmean)
```

```
##       [,1]          [,2]          [,3]
##       "FPG"         "sbp"         "dbp"
```

```
## rmean "5.04117647058824" "108.058823529412" "66.8235294117647"
##       [,4]               [,5]               [,6]
##       "TG"               "HDL.C"            "BMI"
## rmean "1.19882352941176" "1.66705882352941" "22.6862984751759"
```

```
cov(X3)
```

```
##               FPG         sbp        dbp          TG       HDL.C        BMI
## FPG    0.18626103   1.3505515  0.8783456  0.07467022 -0.07678382  0.7727965
## sbp    1.35055147 159.5588235 80.0735294  9.86132353 -0.92919118 42.6730381
## dbp    0.87834559  80.0735294 64.5294118  5.68352941 -0.97617647 23.3303076
## TG     0.07467022   9.8613235  5.6835294  0.83918603 -0.10940368  3.7847764
## HDL.C -0.07678382  -0.9291912 -0.9761765 -0.10940368  0.14365956 -0.7551085
## BMI    0.77279646  42.6730381 23.3303076  3.78477637 -0.75510852 24.8593905
```

```
cor(X3)
```

```
##              FPG        sbp        dbp         TG      HDL.C        BMI
## FPG    1.0000000  0.2477363  0.2533527  0.1888675 -0.4693983  0.3591360
## sbp    0.2477363  1.0000000  0.7891321  0.8522081 -0.1940785  0.6775602
## dbp    0.2533527  0.7891321  1.0000000  0.7723429 -0.3206138  0.5825006
## TG     0.1888675  0.8522081  0.7723429  1.0000000 -0.3150907  0.8286399
## HDL.C -0.4693983 -0.1940785 -0.3206138 -0.3150907  1.0000000 -0.3995737
## BMI    0.3591360  0.6775602  0.5825006  0.8286399 -0.3995737  1.0000000
```

2.2 参数估计

(2) 假设 Yt 服从下面的模型：Y0=0, Yt= Yt-1 + t, t=1, 2, ..., n。利用随机模拟分析下面两种情况下，对参数 估计量进行分析 (如通过均方误差 (MSE) 来说明)。a. 的真值为 0.6；所展示为样本容量为 10 的代码，其余代码省略

```
y <- c(0)
for (i in 1:10){
  y = c(y,0.6*y[i]+rnorm(1))
}
y1 <- y[2:11]
x <- y[1:10]
# 最小二乘法求估计量
lm.sol<-lm(y1 ~ 0+x)
summary(lm.sol)
```

```
##
## Call:
## lm(formula = y1 ~ 0 + x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5256 -0.3791  0.2434  1.0896  2.4462
##
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
## x   0.8822     0.2686   3.284  0.00947 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.264 on 9 degrees of freedom
## Multiple R-squared:  0.5451, Adjusted R-squared:  0.4946
## F-statistic: 10.79 on 1 and 9 DF,  p-value: 0.009466
```

```
MSE.10 <- sum(residuals(lm.sol)^2)/9
```

不同样本容量下估计量的 MSE：

```
data.frame(MSE.10,MSE.100,MSE.1000)
```

```
##     MSE.10   MSE.100  MSE.1000
## 1 1.597136 0.9001335 0.9471368
```

b. 的真值为 1。

不同样本容量下估计量的 MSE：

```
data.frame(MSE2.10,MSE2.100,MSE2.1000)
```

```
##   MSE2.10 MSE2.100 MSE2.1000
## 1 1.13054 1.018937  1.035999
```

可见样本容量越大，估计量越接近真实值