《多元统计分析》第三次上机作业
利用R软件单变量和多变量正态检验和置信区域

3180103000 许乐乐

- 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

习题一

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。选择分析患代谢综合症的年龄差异。

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。

- 导出数据

  用`write.csv()`把数据框保存成CSV格式

```
write.csv(d.y, file="hw3_sample.csv", row.names=FALSE) #将hw2中的d.y导出
```

  在Excel中删除无关列和行，并重命名。

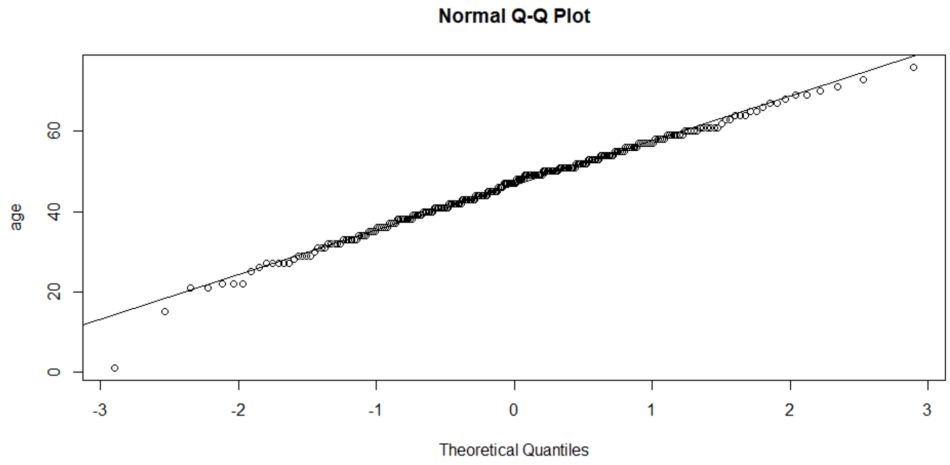
	BMI	FPG	SBP	DBP	TG	HDL-C	age	sick
1	15.03	5.01	95	60	0.74	1.79	56	0
2	10.04	4.44	112	58	0.75	2.17	27	0
3	19.30	5.95	131	92	0.66	1.65	*NA*	0
4	18.14	5.60	98	68	0.75	1.72	*NA*	0
5	14.19	5.93	141	81	1.76	1.91	64	0
6	12.62	5.68	100	69	5.03	1.20	43	0

- 单变量正态性检验

  1. QQ图

     用`qqnorm()`来绘制QQ图

```
qqnorm(d$age, ylab = 'age') # 做年龄的QQ图，将y轴命名为age
qqline(d$age) # 增加趋势直线，利于比较
```



从 Q-Q 图直观上看，数据大致沿直线分布，符合正态性。

2. Shapiro-Wilk正态检验

用`shapiro.test()`进行Shapiro-Wilk正态检验

```
shapiro.test(d$age)
```

```
    Shapiro-Wilk normality test

data:  d$age
W = 0.9917, p-value = 0.1416
```

p值为0.1416>0.05，所以接受原假设，认为年龄分布符合正态性。

- 多变量正态性检验

1. Mardia多元正态性检验

用MVN扩展包的`mvn()`进行Mardia多元正态性检验

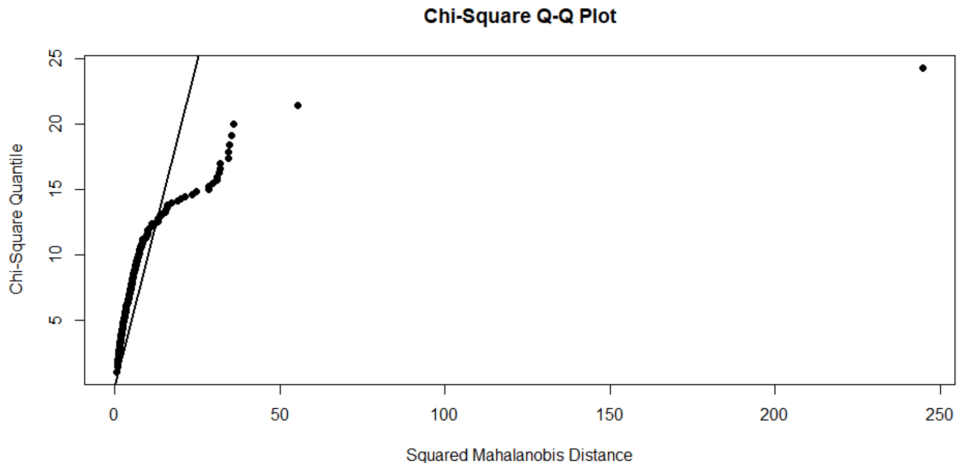
```
library(MVN)
d.na<-na.omit(d)
mvn(d.na,mvnTest=c("mardia"),multivariatePlot = c("qq"))
# data:数据集 mvnTest:检验多元变量正态性的方法 multivariatePlot:作图
```

```
$multivariateNormality
             Test         Statistic p value Result
1 Mardia Skewness 11583.6940773533       0     NO
2 Mardia Kurtosis 169.612808631536       0     NO
3             MVN              <NA>    <NA>     NO

$univariateNormality
          Test  Variable Statistic   p value Normality
1 Shapiro-Wilk    BMI       0.0771  <0.001      NO
2 Shapiro-Wilk    FPG       0.7520  <0.001      NO
3 Shapiro-Wilk    SBP       0.9801  0.0015      NO
4 Shapiro-Wilk    DBP       0.9848  0.0099      NO
5 Shapiro-Wilk    TG        0.7856  <0.001      NO
6 Shapiro-Wilk   HDL-C      0.9512  <0.001      NO
7 Shapiro-Wilk    age       0.9899  0.0837      YES
8 Shapiro-Wilk   sick       0.1819  <0.001      NO

$Descriptives
```

```
        n          Mean     Std.Dev Median   Min     Max
BMI   247  19.89898785 60.0651743  15.74  5.78 956.11
FPG   247   5.55761134  0.9740780   5.30  3.91  10.83
SBP   247 118.67611336 16.9117856 119.00 79.00 173.00
DBP   247  76.01619433 11.4781403  77.00 53.00 107.00
TG    247   1.94429150  1.5097182   1.47  0.30  10.20
HDL-C 247   1.60846154  0.3401806   1.55  0.60   2.98
age   247  46.57085020 11.0439827  47.00  1.00  76.00
sick  247   0.03643725  0.1877559   0.00  0.00   1.00
          25th    75th       Skew    Kurtosis
BMI     11.945  19.370 15.3297932 235.9505745
FPG      5.040   5.765  2.6471432   9.5028684
SBP    107.000 127.500  0.4750314   0.1959284
DBP     67.000  84.000  0.1812160  -0.6265180
TG       0.940   2.475  2.1632006   6.0083748
HDL-C    1.365   1.810  0.8861732   1.4100695
age     40.000  54.000 -0.3599318   0.7251188
sick     0.000   0.000  4.9179372  22.2763432
```

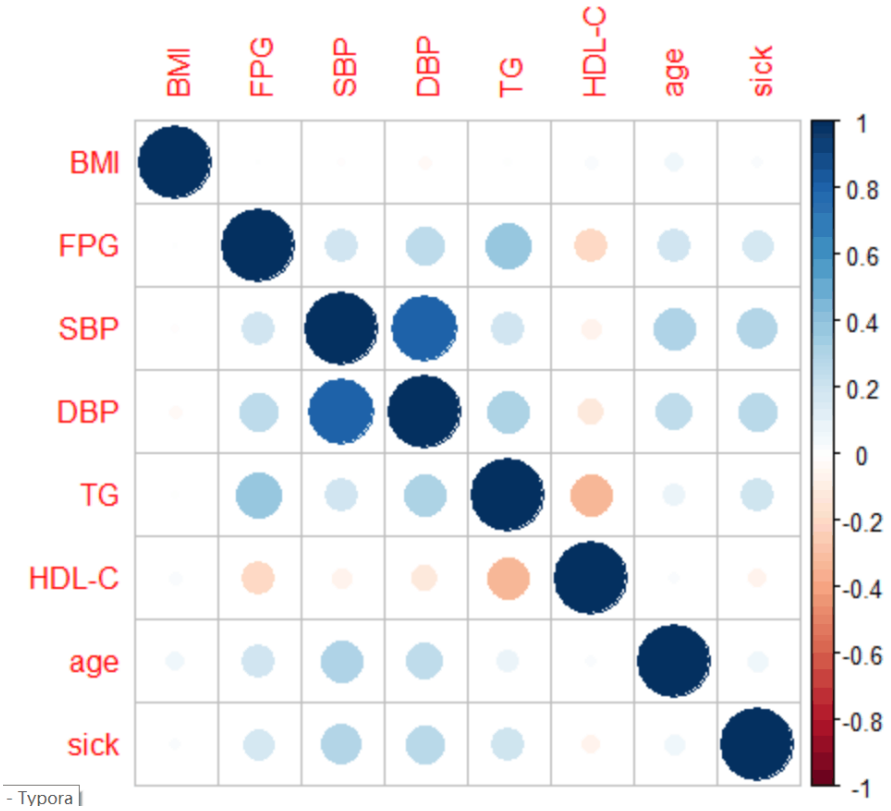


由p值和qq图都可以看出，不符合多元正态性，而且除了年龄之外的变量都不符合单变量正态性。

- 绘制相关矩阵图

  用corrplot扩展包中的`corrplot()`绘制相关矩阵图

```
d.cor<-cor(d.na) # 计算相关矩阵
round(d.cor, digits=2) # 保留两位小数
library(corrplot)
corrplot(d.cor)
```



由相关矩阵图显然可见，SBP和DBP有较强正相关性，其余变量几乎没有相关性。

- 相关系数的显著性检验

  用psych扩展包的`corr.test()`对单个的Pearson, Spearman 和Kendall相关系数进行检验，计算数据集的相关系数矩阵和显著性。

```
corr.test(d.na)
```

```
Call:corr.test(x = d.na)
Correlation matrix
        BMI   FPG   SBP   DBP    TG HDL-C  age  sick
BMI    1.00  0.00 -0.01 -0.03  0.01  0.03 0.06  0.02
FPG    0.00  1.00  0.20  0.27  0.39 -0.20 0.19  0.19
SBP   -0.01  0.20  1.00  0.81  0.20 -0.07 0.31  0.29
DBP   -0.03  0.27  0.81  1.00  0.32 -0.13 0.25  0.28
TG     0.01  0.39  0.20  0.32  1.00 -0.33 0.09  0.21
HDL-C  0.03 -0.20 -0.07 -0.13 -0.33  1.00 0.02 -0.07
age    0.06  0.19  0.31  0.25  0.09  0.02 1.00  0.07
sick   0.02  0.19  0.29  0.28  0.21 -0.07 0.07  1.00
Sample Size
[1] 247
Probability values (Entries above the diagonal are adjusted for multiple
tests.)
       BMI FPG  SBP  DBP   TG HDL-C  age sick
BMI   0.00   1 1.00 1.00 1.00  1.00 1.00 1.00
FPG   0.97   0 0.03 0.00 0.00  0.03 0.03 0.05
SBP   0.87   0 0.00 0.00 0.03  1.00 0.00 0.00
DBP   0.63   0 0.00 0.00 0.00  0.61 0.00 0.00
TG    0.89   0 0.00 0.00 0.00  0.00 1.00 0.02
HDL-C 0.68   0 0.27 0.05 0.00  0.00 1.00 1.00
age   0.34   0 0.00 0.00 0.18  0.72 0.00 1.00
sick  0.70   0 0.00 0.00 0.00  0.30 0.28 0.00
```

```
 To see confidence intervals of the correlations, print with the
short=FALSE option
```

返回的结果给出了两部分，第一部分是相关矩阵，第二部分是Probability values即显著性。

- 分析患代谢综合症的年龄差异
    - 将数据按年龄进行分组

      用dplyr扩展包的`group_by()`和`summarise()`作数据分组汇总

```
d.age<-d%>%
  select(age,sick)%>%
  drop_na()
d.age.group<-d.age%>%
  group_by(sick)%>%
  summarise(
    age0_20=sum(age<=20),
    age20_40=sum(age<=40&age>=20),
    age40_60=sum(age<=60&age>=40),
    age60_80=sum(age<=80&age>=60)
    )
```

sick	age0_20	age20_40	age40_60	age60_80
不患病	2	62	160	25
患病	0	1	6	2
患病比例	0	0.01587	0.03614	0.07407

    - 相关性的显著性检验

      用`cor.test()`进行一种相关关系的显著性检验

```
cor.test(d.age$age,d.age$sick)
```

```
    Pearson's product-moment correlation

data:  d.age$age and d.age$sick
t = 1.0723, df = 245, p-value = 0.2847
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.05696113  0.19152987
sample estimates:
       cor
0.06834427
```

      p值为0.2847>0.05，因此认为患代谢综合症的年龄差异不大。

- 计算患病与不患病的各类指标的均值和置信区间分析差异。

  用pastecs扩展包的`stat.desc()`可以计算种类繁多的描述性统计量。

  使用格式为：`stat.desc(x,basic=TRUE,desc=TRUE,norm=FALSE,p=0.95)`
  其中的x是一个数据框或时间序列。若basic=TRUE（默认值），则计算其中所有值、空值、缺失值的数量，以及最小值、最大值、值域，还有总和。若desc=TRUE（同样也是默认值），则计算中位数、平均数、平均数的标准误、平均数置信度为95%的置信区间、方差、标准差以及

变异系数。最后，若norm=TRUE（不是默认的），则返回正态分布统计量，包括偏度和峰度（以及它们的统计显著程度）和Shapiro-Wilk正态检验结果。

```
d.sick<-d%>%
  filter(sick==1)
d.nosick<-d%>%
  filter(sick==0)
library(pastecs)
stat.desc(d.sick,desc=TRUE,p=0.95)
stat.desc(d.nosick,desc=TRUE,p=0.95)
```

sick	BMI	FPG	SBP	DBP	TG	HDL-C
患病	26.00 [18.00,34.00]	6.48 [5.31,7.65]	142.70 [131.34,154.06]	92.20 [87.06,97.35]	3.54 [2.18,4.89]	1.49 [1.31,1.66]
不患病	19.59 [11.91,27.26]	5.52 [5.41,5.64]	117.76 [115.72,119.81]	75.5 [74.11,76.91]	1.87 [1.68,2.04]	1.614 [1.57,1.66]

患病人群相较于不患病人群，BMI、FPG、SBP、TG均值偏高，DBP、HDL-C均值偏低。

习题二

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?

- 读入数据

```
library(readr)
d<-read_csv("ex2.1.csv")
d<-d%>%
  na.omit()
```

- 建立线性回归方程

  用`lm()`来拟合线性回归模型

```
d.lm<-lm(y~x1+x2+x3+x4,d)
summary(d.lm)
```

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4, data = d)

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
```

故线性回归方程为：

$$y = 5.9433 + 0.1424x_1 + 0.3515x_2 - 0.2706x_3 + 0.6328x_4$$

分析得：

- 由回归系数的估计值知，血糖与胆固醇、甘油、血红蛋白正相关，与胰岛素负相关。
- 由回归系数估计t检验的p值知，血红蛋白、胰岛素对血糖影响较大，甘油对血糖影响较小，胆固醇几乎不影响血糖。
- 由拟合优度$R^2 = 0.6008$和修正的拟合优度$Adjusted\ R^2 = 0.5282$知，回归方程的拟合程度较高。
- 由F检验的p值知，p-value: 0.0003121<0.05，方程在p=0.05的水平上通过显著性检验，认为血糖变化可以由胆固醇、甘油、胰岛素、血红蛋白的变化来解释。

- 多元正态性检验

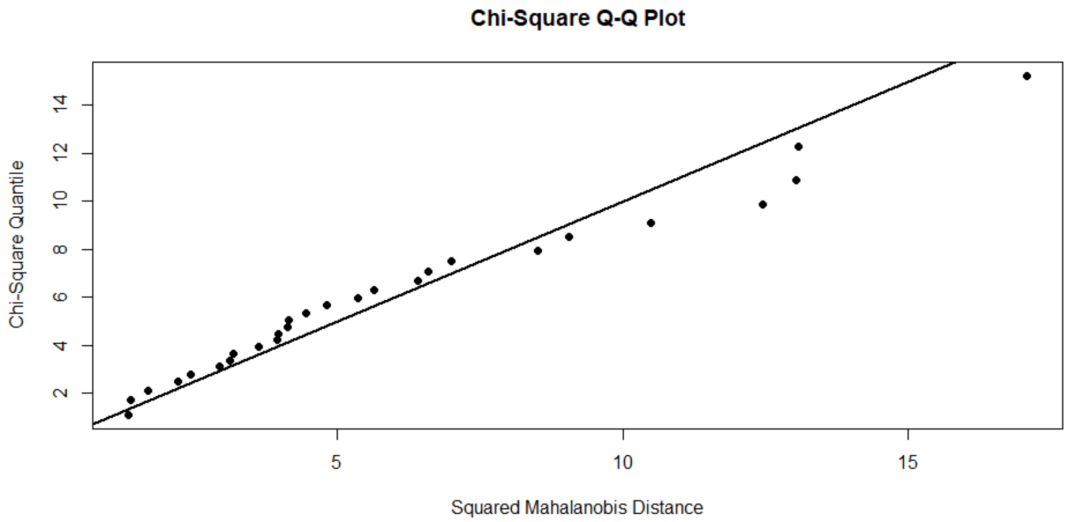
```
library(MVN)
mvn(d,mvnTest=c("royston"),multivariatePlot = c("qq"))
```

```
$multivariateNormality
     Test        H         p value MVN
1 Royston 42.78297 0.0000001220787  NO

$univariateNormality
          Test  Variable Statistic   p value
1 Shapiro-Wilk    No        0.9581  0.3335
2 Shapiro-Wilk    x1        0.8484  0.0011
3 Shapiro-Wilk    x2        0.7318  <0.001
4 Shapiro-Wilk    x3        0.9255  0.0537
5 Shapiro-Wilk    x4        0.9556  0.2916
6 Shapiro-Wilk     y        0.9032  0.0159
  Normality
1    YES
2    NO
3    NO
4    YES
5    YES
6    NO

$Descriptives
    n      Mean  Std.Dev Median  Min   Max 25th
No 27 14.000000 7.937254  14.00 1.00 27.00 7.50
x1 27  5.812593 1.593380   5.78 3.79 11.54 4.75
x2 27  2.840741 2.574765   1.97 0.63 10.89 1.19
x3 27  6.146667 3.670625   5.88 1.20 16.28 3.54
x4 27  9.118519 1.823361   8.70 6.40 13.60 7.85
y  27 11.925926 2.925694  11.10 8.40 20.00 9.85
     75th      Skew   Kurtosis
```

```
No 20.500 0.0000000 -1.3339212
x1  6.160 1.6160190  3.7968990
x2  3.180 1.7591262  2.1677125
x3  7.445 0.9116462  0.4035898
x4 10.400 0.5376065 -0.5482759
y  13.350 1.0514540  0.5120622
```



由$multivariateNormality知，在Royston Test中，p value=0.0000001220787<0.05，所以拒绝原假设，认为(x1, x2, x3, x4)不服从多元正态。