**浙江大学数学学院统计实验二**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **课程名称：** | **多元统计分析** | **软件平台** | | R软件 | |
| **实验项目名称:** | 实验一：利用R软件进行参数估计和假设检验 | | | | |
| **学生姓名 ：** | **袁泽清** | **专业 ：** | **计算机科学与技术（图灵班）** | **学号：** | **3200105867** |

|  |
| --- |
| 1. **实验目的与要求：**   通过本试验项目，（1）使学生理解并掌握数理统计中一些单变量参数估计和假设检验问题中在R软件包中的实现；（2）多维数据的均值、方差、协方差矩阵等等；（3）多维正态密度函数、分布函数，随机数等的产生；（4）二维正态密度曲线作图。 |
| **实验内容 ：**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),   1. 分析X各变量之间的相关性？ 2. 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？ 3. 分年龄(小于等于30， 30~50, 50~70, 70以上)，分析X中的各个指标是否有年龄上的差异？ 4. 计算X样本均值、样本离差阵、样本协方差和样本相关阵。 5. 检验各分量是否服从正态分布。   **注意：由于是原始数据，数据中有可能有缺失，有数据读入的错误等，在分析之前，要进行数据预处理，如查异常值，分析数据分布形态，检查数据缺失情况，并在实验报告中说明你是如何处理缺失数据的……。** |

**实验数据说明：**

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.

## 数据预处理



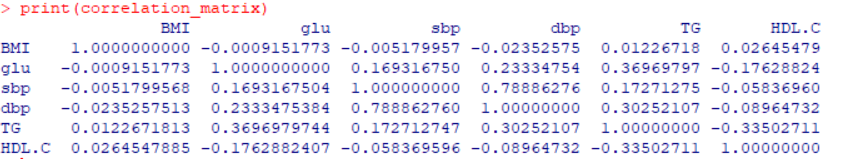
在数据预处理中，针对除了HDL-C以外的所有列的缺失值，采用fill函数用下一行可用值进行近似填补。

## 分析X各变量之间的相关性

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),



得到结果如下：



由此可见：

BMI 与其他变量的相关性非常弱，表明BMI和血糖、血压、甘油三酯以及高密度脂蛋白胆固醇之间没有明显的线性关系。

血糖（glu） 与收缩压（sbp）、舒张压（dbp）和甘油三酯（TG）有轻微到中等的正相关性，这可能表明血糖水平较高的个体可能会伴有较高的血压和甘油三酯水平。

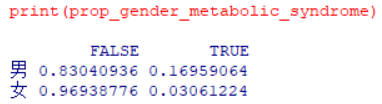
收缩压（sbp） 与舒张压（dbp）之间的相关性很高，这是符合预期的，因为这两者都是血压的组成部分。

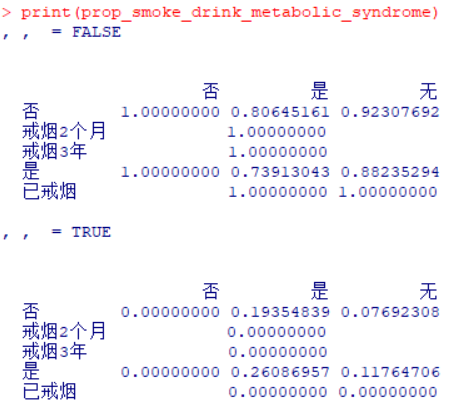
甘油三酯（TG） 与高密度脂蛋白胆固醇（HDL.C）之间存在中等程度的负相关性，暗示着甘油三酯水平升高可能与较低的HDL胆固醇水平相关。

## 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？



运行得到结果如下



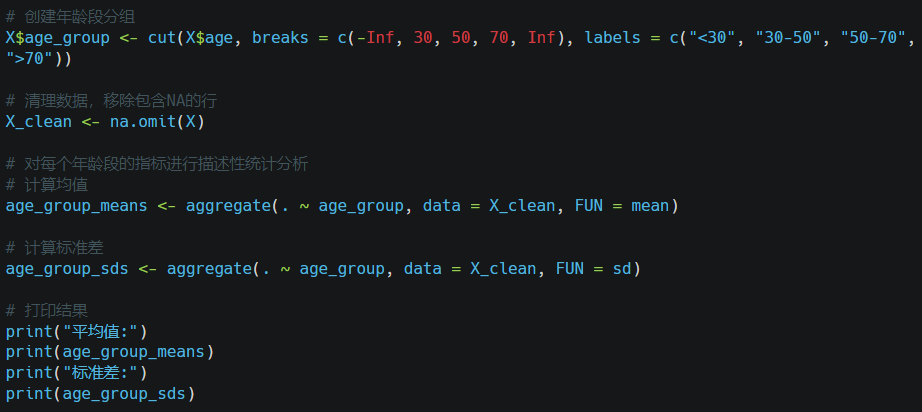


由此可见：

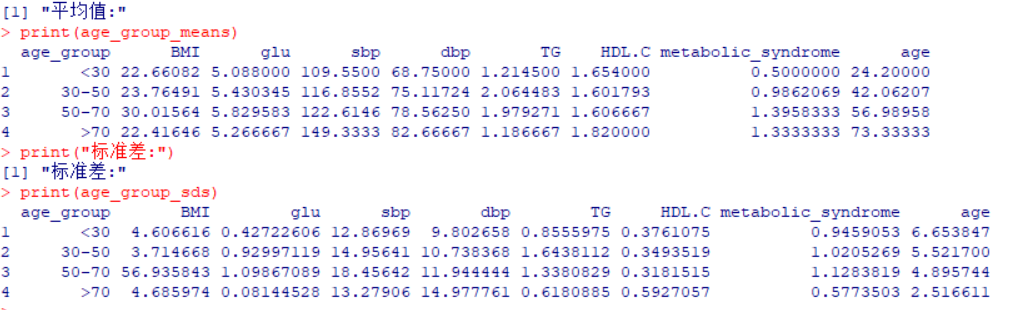
男性患有代谢综合症的比例明显高于女性，显示出性别在代谢综合症的患病率上的显著差异。

吸烟和饮酒可能与代谢综合症的患病率有一定关联，尤其是在同时吸烟和饮酒的人群中，同时吸烟和饮酒的人群中代谢综合症的患病率似乎更高。。

## 分年龄(小于等于30， 30~50, 50~70, 70以上)，分析X中的各个指标是否有年龄上的差异？



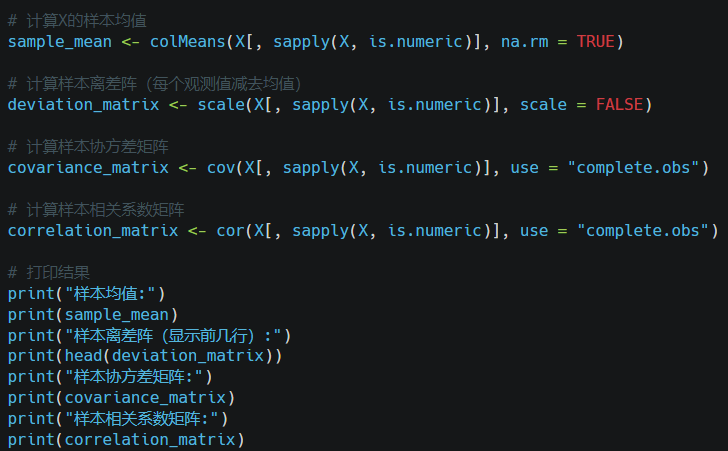
运行结果如下：



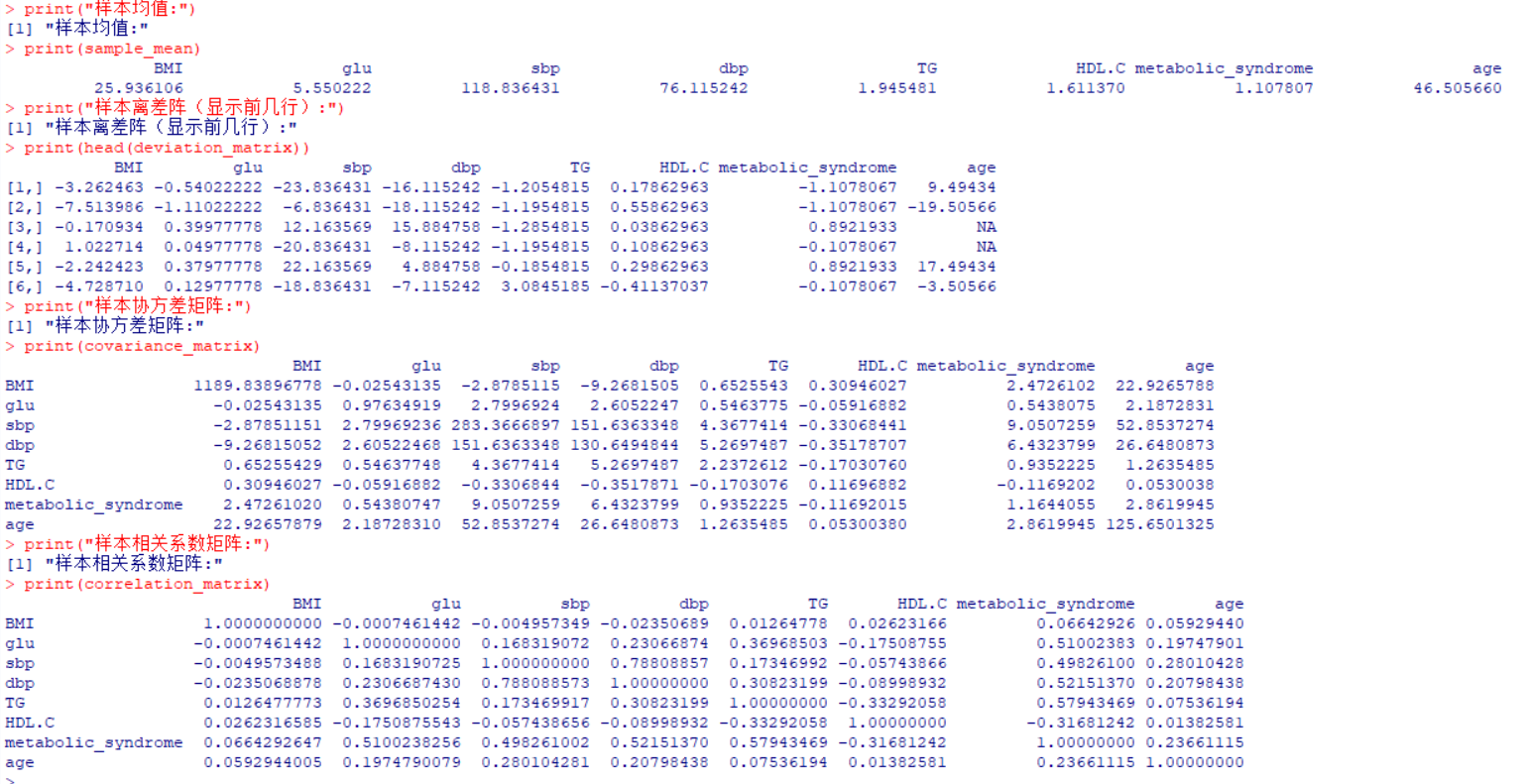
由此可见：

BMI、sbp、dbp、TG在年龄上有比较显著的差异

## 计算X样本均值、样本离差阵、样本协方差和样本相关阵。

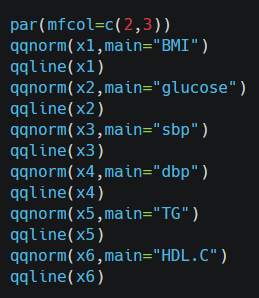


运行结果如下：

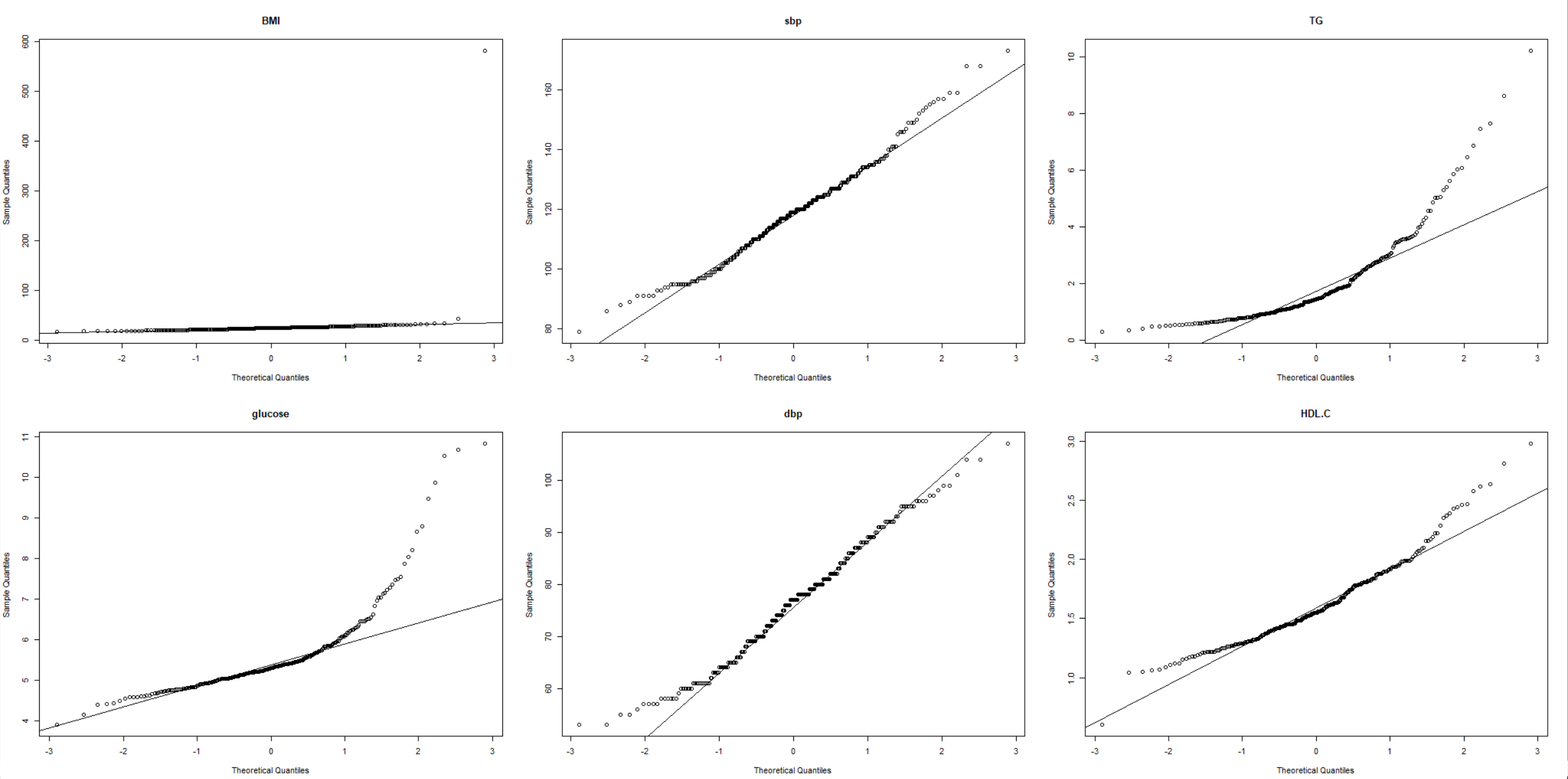


## 检验各分量是否服从正态分布。

用QQplot 的方式检验各个变量的正态性：



可得如下结果：



由此可见：

除了甘油三酯和葡萄糖的含量不服从正态分布，其他的指标均服从正态分布。