

# Homework 7

张子乐,3230104237

## 第一题

```
df <- as.data.frame(read_excel("ex6.7.xls"))

## New names:
## * `` -> `...1`


rownames(df) <- df[[1]]
df <- df[-1]
KMO(df)          # KMO 测度检验

## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = df)
## Overall MSA =  0.83
## MSA for each item =
##      食品      衣着      居住      医疗      交通通讯      教育      家庭服务
##      0.80      0.71      0.84      0.70      0.81      0.92      0.88
## 耐用消费品
##      0.88

cortest.bartlett(df)  # Bartlett 球形检验

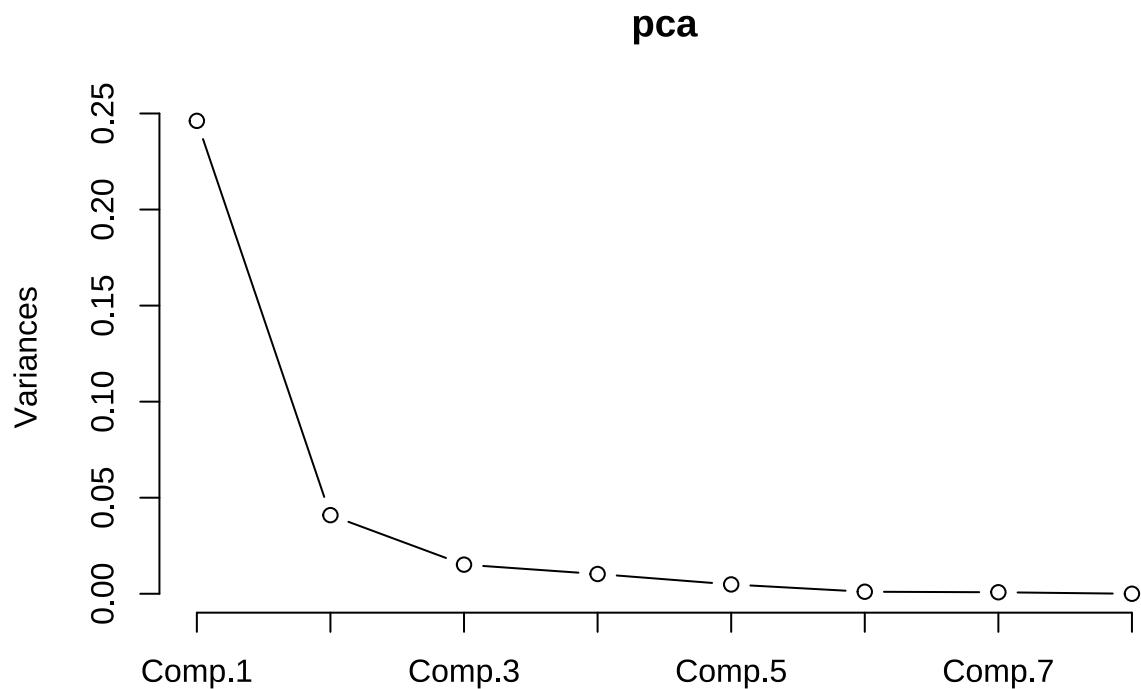
## R was not square, finding R from data

## $chisq
## [1] 227.2652
##
## $p.value
## [1] 4.260516e-33
##
## $df
## [1] 28
```

KMO 检验考察变量间的相关性和偏相关性，检验结果表明，总  $KMO = 0.83$ ，各个变量的  $KMO >= 0.7$ ，适合做因子分析。

Bartlett's 球形检验的  $H_0$  是各变量之间没有相关关系，即不能将多个变量简化为少数的成分，没有进行主成分提取的必要。检验结果表明， $p < 0.05$ ，拒绝原假设，各变量之间有足够的相关性，可以进行因子分析。

```
pca <- princomp(cor(df))
screeplot(pca, type = "lines")
abline(h = 1, col = "red", lty = 2) # 碎石图
```



根据碎石图和后面的因子分析，选择 2 个因子。

```
fre<-factanal(df, 2, scores="Bartlett",rotation = "varimax")
print(fre)
```

```
##
## Call:
## factanal(x = df, factors = 2, scores = "Bartlett", rotation = "varimax")
##
## Uniquenesses:
```

```

##      食品    衣着    居住    医疗    交通通讯    教育    家庭服务
##  0.123    0.544    0.281    0.191    0.043    0.256    0.106
## 耐用消费品
##  0.272
##
## Loadings:
##          Factor1 Factor2
## 食品      0.925   0.147
## 衣着      0.232   0.634
## 居住      0.659   0.533
## 医疗      0.126   0.891
## 交通通讯  0.941   0.268
## 教育      0.699   0.505
## 家庭服务  0.888   0.324
## 耐用消费品 0.515   0.680
##
##          Factor1 Factor2
## SS loadings   3.788   2.396
## Proportion Var  0.473   0.300
## Cumulative Var  0.473   0.773
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 14.64 on 13 degrees of freedom.
## The p-value is 0.331

```

卡方检验原假设  $H_0$ : 数据协方差结构可以由 2 个公共因子完全解释。 $\chi^2 = 14.64, p = 0.331 > 0.05$ , 不拒绝原假设, 因此该因子模型是充分且合理的。

其中, 因子 1 解释了 47.3% 的总方差, 因子 2 解释了 30% 的总方差, 累计解释方差 77.3%, 说明有效捕捉了原始变量的信息。

```

communality <- 1 - fre$uniquenesses
df_communality <- data.frame(
  变量 = names(communality),
  公因子方差 = round(communality, 3)
)
print(df_communality)

```

```

##      变量 公因子方差
## 食品    食品      0.877
## 衣着    衣着      0.456

```

```

## 居住      居住      0.719
## 医疗      医疗      0.809
## 交通通讯  交通通讯  0.957
## 教育      教育      0.744
## 家庭服务  家庭服务  0.894
## 耐用消费品 耐用消费品 0.728

```

多数变量的共同度  $> 0.7$ , “交通通讯”、“家庭服务”、“食品”等接近 0.9, 说明它们的信息被两个因子较好地捕捉; “衣着”的共同度仅为 0.456, 是所有变量中最低的, 意味着其不能被这两个因子很好的解释。

```

loadings_rotated <- loadings(fre)
df_loadings <- data.frame(
  变量 = rownames(loadings_rotated),
  Factor1 = round(loadings_rotated[, "Factor1"], 3),
  Factor2 = round(loadings_rotated[, "Factor2"], 3)
)
print(df_loadings)

```

	变量	Factor1	Factor2
## 食品	食品	0.925	0.147
## 衣着	衣着	0.232	0.634
## 居住	居住	0.659	0.533
## 医疗	医疗	0.126	0.891
## 交通通讯	交通通讯	0.941	0.268
## 教育	教育	0.699	0.505
## 家庭服务	家庭服务	0.888	0.324
## 耐用消费品	耐用消费品	0.515	0.680

根据旋转后的因子载荷矩阵, 因子 1 中食品 (0.925)、居住 (0.659)、交通通讯 (0.941)、教育 (0.699)、家庭服务 (0.888) 因子载荷较高, 主要反映居民在生存必需、子女教育、居住条件改善和日常便利服务上的综合投入水平, 可命名为基础生活因子; 因子 2 中衣着 (0.634)、医疗 (0.891)、耐用消费品 (0.680) 因子载荷较高, 教育和居住也有较高的交叉载荷, 体现个体对身体健康维护、外在形象管理、耐用资产积累的关注程度, 可命名为品质生活因子。

## 第二题

```

# 先清洗数据, 将字符串转为数值类型, 删除缺失值
df2 <- as.data.frame(read_excel(" 体验数据.xls"))
df2 <- df2[-1, ]

```

```

Age <- as.numeric(as.character(df2[["Age"]]))
selected_vars <- c("Sbp", "Dbp", "Sphygmus", "Weight", "Height",
                  "TC", "TG", "ALT", "AST", "T-BIL", "IB",
                  "ALP", "TP", "Alb", "GLB")
df_temp <- df2[, c(selected_vars, "Age"), drop = FALSE]
for (col in names(df_temp)) {
  df_temp[[col]] <- as.numeric(as.character(df_temp[[col]]))
}

## Warning: NAs introduced by coercion
## Warning: NAs introduced by coercion

df_clean <- na.omit(df_temp)
rownames(df_clean) <- NULL

df2 <- df_clean[, selected_vars, drop = FALSE]
Age <- df_clean[["Age"]]

# 主成分分析
pca <- princomp(df2, cor = TRUE)
print(summary(pca)) # 特征根的平方根和方差贡献率

## Importance of components:
##                               Comp.1    Comp.2    Comp.3    Comp.4    Comp.5
## Standard deviation      1.8852608 1.4820144 1.3810009 1.17354888 1.11047946
## Proportion of Variance 0.2369472 0.1464244 0.1271442 0.09181447 0.08221098
## Cumulative Proportion  0.2369472 0.3833717 0.5105159 0.60233036 0.68454133
##                               Comp.6    Comp.7    Comp.8    Comp.9    Comp.10
## Standard deviation      1.0288761 0.96309333 0.89071769 0.86399831 0.77506394
## Proportion of Variance 0.0705724 0.06183658 0.05289187 0.04976621 0.04004827
## Cumulative Proportion  0.7551137 0.81695032 0.86984219 0.91960839 0.95965667
##                               Comp.11   Comp.12   Comp.13   Comp.14  Comp.15
## Standard deviation      0.51872944 0.46045152 0.320359305 0.14636954      0
## Proportion of Variance 0.01793868 0.01413437 0.006842006 0.00142827      0
## Cumulative Proportion  0.97759535 0.99172972 0.998571730 1.00000000      1

print(pca$loadings) # 因子载荷

```

##

```

## Loadings:
##          Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10
## Sbp      0.335     0.109  0.428  0.345           0.189      0.221
## Dbp      0.361     0.120  0.420  0.279           0.112  0.190  0.115
## Sphygmus      -0.179 -0.103  0.180  0.474 -0.317 -0.426 -0.104 -0.625 -0.116
## Weight    0.365  0.189  0.201  0.110 -0.313           0.235 -0.185 -0.136
## Height    0.268  0.259           0.128 -0.479 -0.276           0.152 -0.247 -0.240
## TC        0.206 -0.159           0.539 -0.470 -0.259  0.197 -0.547
## TG        0.259 -0.197       -0.272  0.345 -0.259 -0.106 -0.260  0.732
## ALT       0.358     0.195 -0.500  0.154 -0.157
## AST       0.330     0.171 -0.537  0.229           0.112
## T-BIL     0.176  0.394 -0.490 -0.114  0.124  0.189
## IB        0.184  0.383 -0.489 -0.105  0.112  0.224
## ALP       0.225           0.102  0.617 -0.626 -0.344 -0.146
## TP        0.192 -0.465 -0.408       -0.157 -0.136  0.101  0.107
## Alb       0.214     -0.351       -0.202 -0.473 -0.158 -0.434  0.373  0.107
## GLB       -0.508 -0.251       0.179  0.242  0.453 -0.161 -0.148
##          Comp.11 Comp.12 Comp.13 Comp.14 Comp.15
## Sbp        0.687
## Dbp     -0.198 -0.700
## Sphygmus
## Weight    0.721     -0.202
## Height   -0.608  0.129
## TC
## TG
## ALT        0.711
## AST     -0.204     -0.660
## T-BIL           0.708
## IB        -0.704
## ALP
## TP        0.709
## Alb       -0.418
## GLB       -0.568
##
##          Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
## SS loadings  1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000
## Proportion Var 0.067  0.067  0.067  0.067  0.067  0.067  0.067  0.067  0.067
## Cumulative Var 0.067  0.133  0.200  0.267  0.333  0.400  0.467  0.533  0.600
##          Comp.10 Comp.11 Comp.12 Comp.13 Comp.14 Comp.15

```

```

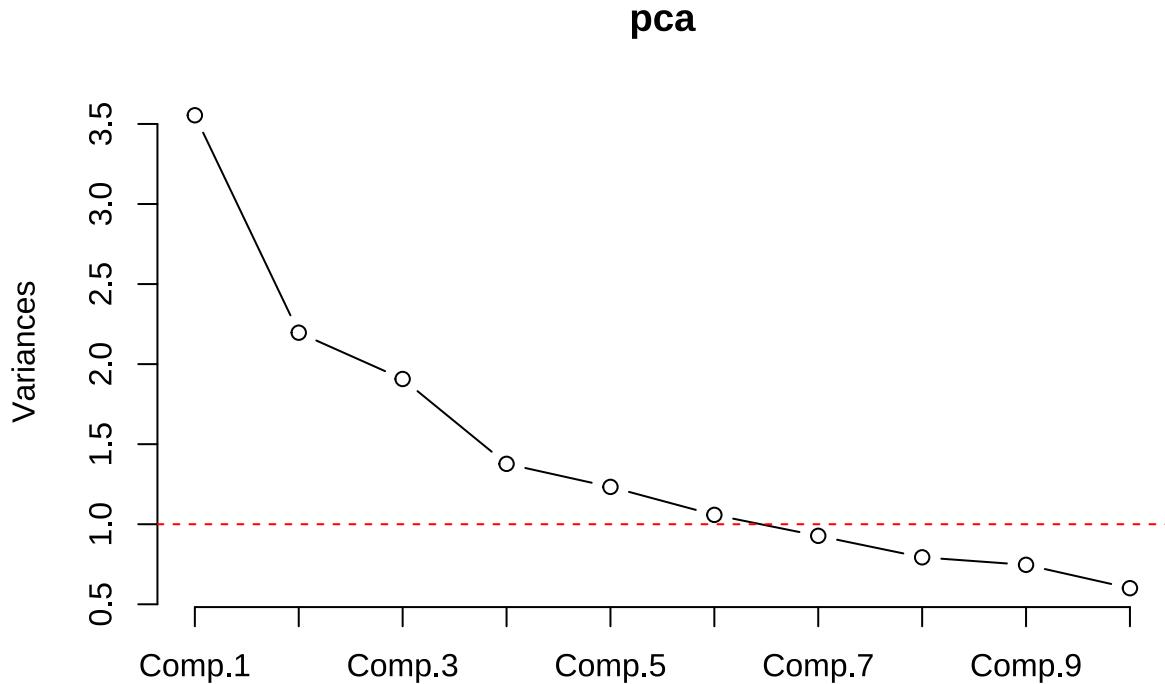
## SS loadings    1.000  1.000  1.000  1.000  1.000  1.000
## Proportion Var 0.067   0.067   0.067   0.067   0.067   0.067
## Cumulative Var 0.667   0.733   0.800   0.867   0.933   1.000

```

```

screeplot(pca, type="lines") # 碎石图
abline(h = 1, col = "red", lty = 2)

```



由碎石图和累计方差贡献率，选择前六个主成分，累计方差贡献率达到 0.75，其中第一主成分的方差贡献率为 0.24，第二主成分的方差贡献率为 0.15，第三主成分的方差贡献率为 0.13。将样本根据主成分得分用 k-means 聚类方法分为 3 类，在前两个主成分上的散点图如下图所示。

```

scores <- pca$scores[, 1:6]
# 使用 K-means 对样本进行聚类
set.seed(123)
k <- 3
km <- kmeans(scores, centers = k, iter.max = 100, nstart = 25)

df_cluster <- data.frame(df2, PC1 = scores[,1], PC2 = scores[,2], cluster = km$cluster)

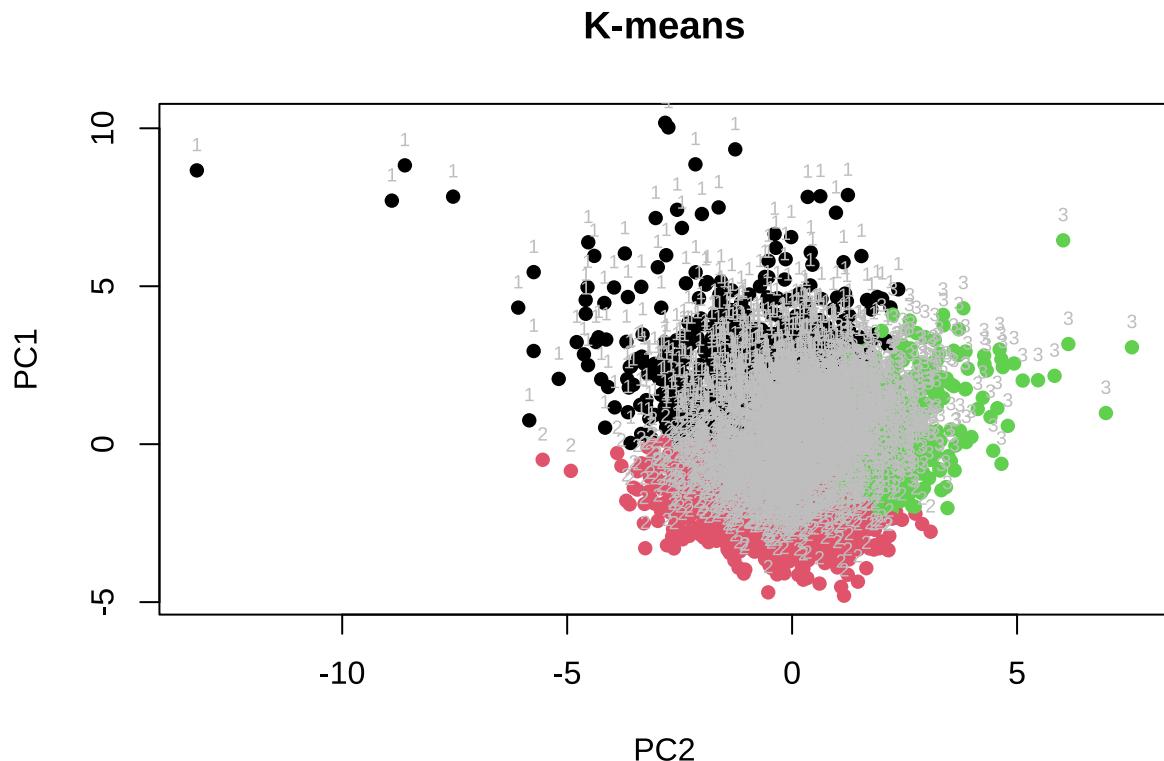
plot(PC1 ~ PC2, data = df_cluster, col = km$cluster, pch = 16,

```

```

    main = "K-means")
text(df_cluster$PC2, df_cluster$PC1, labels = km$cluster, pos = 3, cex = 0.6, col = "gray")

```



```
# 用主成分分析旋转后的因子载荷矩阵做聚类
```

```

L <- pca$loadings[, 1:6]
L_rot <- varimax(L)$loadings
class(L_rot) <- "loadings"
print(L_rot, cutoff = 0.4)

##
## Loadings:
##          Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
## Sbp      0.649
## Dbp      0.622
## Sphygmus 0.406
## Weight           -0.537
## Height          -0.676
## TC                  0.594
## TG                  0.529

```

```

## ALT           -0.679
## AST           -0.699
## T-BIL         -0.700
## IB            -0.702
## ALP
## TP            -0.675
## Alb           -0.525
## GLB           -0.457
##
##             Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
## SS loadings   1.000  1.000  1.000  1.000  1.000  1.000
## Proportion Var 0.067  0.067  0.067  0.067  0.067  0.067
## Cumulative Var 0.067  0.133  0.200  0.267  0.333  0.400

```

用主成分分析旋转后的因子载荷矩阵做聚类，Sbp、Dbp 和 Sphygmus 为一类，TP、Alb 和 GLB 为一类，T-BIL 和 IB 为一类，ALT 和 AST 为一类，Weight 和 Height 为一类，TC 和 TG 为一类。

```

error <- df2$TP - (df2$Alb + df2$GLB) # 发现 TP = Alb + GLB
max_abs_error <- max(abs(error))
print(max_abs_error)

```

```
## [1] 1.421085e-14
```

```
df2 <- df2[, !names(df2) %in% c("TP")] # 删除 TP
```

```
KMO(df2) # KMO 测度检验
```

```

## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = df2)
## Overall MSA = 0.62
## MSA for each item =
##      Sbp      Dbp Sphygmus    Weight    Height      TC      TG      ALT
## 0.65     0.69     0.57     0.68     0.63     0.76     0.77     0.59
##      AST     T-BIL      IB      ALP      Alb      GLB
## 0.58     0.52     0.53     0.89     0.70     0.61

```

```
cortest.bartlett(df2) # Bartlett 球形检验
```

```
## R was not square, finding R from data
```

```

## $chisq
## [1] 30362.47
##
## $p.value
## [1] 0
##
## $df
## [1] 91

```

KMO 检验考察变量间的相关性和偏相关性, 检验结果表明, 总  $KMO = 0.62$ , 大多数变量的  $KMO >= 0.6$ , 适合做因子分析。

Bartlett's 球形检验的  $H_0$  是各变量之间没有相关关系, 即不能将多个变量简化为少数的成分, 没有进行主成分提取的必要。检验结果表明,  $p < 0.05$ , 拒绝原假设, 各变量之间有足够的相关性, 可以进行因子分析。

```

fre2<-factanal(df2,5, scores="Bartlett",rotation = "varimax") # 由前面的主成分分析, 选择 5 个因子
print(fre2)

```

```

##
## Call:
## factanal(x = df2, factors = 5, scores = "Bartlett", rotation = "varimax")
##
## Uniquenesses:
##      Sbp      Dbp Sphygmus   Weight   Height      TC      TG      ALT
## 0.034    0.341   0.945    0.173    0.394    0.740    0.504   0.005
##      AST     T-BIL      IB      ALP      Alb      GLB
## 0.209    0.005   0.038    0.872    0.921    0.818
##
## Loadings:
##          Factor1 Factor2 Factor3 Factor4 Factor5
## Sbp            0.969        0.106
## Dbp            0.137     0.761   0.150   0.173
## Sphygmus       0.124    -0.191
## Weight         0.248     0.326   0.785   0.192
## Height         0.135     0.148   0.748
## TC             0.100     0.138
## TG             0.127     0.157   0.668
## ALT            0.970     0.117   0.130   0.139
## AST            0.866     0.126
## T-BIL          0.990
## IB             0.976

```

```

## ALP           0.201   0.205           0.201
## Alb          0.204                0.104
## GLB           0.225   -0.225          0.358
##
##             Factor1 Factor2 Factor3 Factor4 Factor5
## SS loadings   2.028   1.867   1.771   1.350   0.986
## Proportion Var 0.145   0.133   0.127   0.096   0.070
## Cumulative Var 0.145   0.278   0.405   0.501   0.572
##
## Test of the hypothesis that 5 factors are sufficient.
## The chi square statistic is 575.97 on 31 degrees of freedom.
## The p-value is 1.23e-101

```

其中，因子 1 解释了 14.5% 的总方差，因子 2 解释了 13.3% 的总方差，因子 3 解释了 12.7% 的总方差，5 个因子累计解释方差 57.2%，说明有效捕捉了原始变量的信息。

```

communality2 <- 1 - fre2$uniquenesses
df_communality2 <- data.frame(
  变量 = names(communality2),
  公因子方差 = round(communality2, 3)
)
print(df_communality2)

```

	变量	公因子方差
## Sbp	Sbp	0.966
## Dbp	Dbp	0.659
## Sphygmus	Sphygmus	0.055
## Weight	Weight	0.827
## Height	Height	0.606
## TC	TC	0.260
## TG	TG	0.496
## ALT	ALT	0.995
## AST	AST	0.791
## T-BIL	T-BIL	0.995
## IB	IB	0.962
## ALP	ALP	0.128
## Alb	Alb	0.079
## GLB	GLB	0.182

多数变量的共同度  $> 0.7$ ，“Sbp”、“ALT”、“T-BIL”、“IB”等超过 0.9，说明它们的信息被六个因子较好地

捕捉; “Sphygmus” 的共同度仅为 0.055, 是所有变量中最低的, “Alb”、“ALP”、“GLB”、“TC” 的共同度也低于 0.3, 它们不能被这五个因子很好的解释。

```
loadings_rotated2 <- loadings(fre2)
df_loadings2 <- data.frame(
  变量 = rownames(loadings_rotated2),
  Factor1 = round(loadings_rotated2[, "Factor1"], 3),
  Factor2 = round(loadings_rotated2[, "Factor2"], 3),
  Factor3 = round(loadings_rotated2[, "Factor3"], 3),
  Factor4 = round(loadings_rotated2[, "Factor4"], 3),
  Factor5 = round(loadings_rotated2[, "Factor5"], 3)
)
print(df_loadings2)

##          变量 Factor1 Factor2 Factor3 Factor4 Factor5
## Sbp        Sbp   0.079   0.097   0.969   0.012   0.106
## Dbp        Dbp   0.097   0.137   0.761   0.150   0.173
## Sphygmus  Sphygmus -0.017   0.042   0.124  -0.191   0.040
## Weight     Weight  0.082   0.248   0.326   0.785   0.192
## Height     Height  0.135   0.082   0.148   0.748   0.021
## TC          TC    0.025   0.100   0.138   0.028   0.480
## TG          TG    0.003   0.127   0.092   0.157   0.668
## ALT         ALT   0.058   0.970   0.117   0.130   0.139
## AST         AST   0.072   0.866   0.126  -0.004   0.142
## T-BIL       T-BIL  0.990   0.022   0.031   0.072  -0.090
## IB          IB    0.976   0.017   0.037   0.078  -0.037
## ALP         ALP   0.021   0.201   0.205   0.064   0.201
## Alb         Alb   0.204   0.098   0.094   0.090   0.104
## GLB         GLB  -0.051   0.011   0.006  -0.225   0.358
```

根据旋转后的因子载荷矩阵, 因子 1 中总胆红素 T-BIL(0.990)、间接胆红素 IB(0.976) 因子载荷较高, 反映胆红素生成与肝前/肝内处理能力; 因子 2 中谷丙转氨酶 ALT(0.970)、谷草转氨酶 AST(0.866) 因子载荷较高, 它们是经典的肝细胞损伤敏感指标; 因子 3 中舒张压 Sbp(0.969)、收缩压 Dbp(0.761) 因子载荷较高, 它们共同构成动脉血压核心指标, 反映心脏泵血与外周血管阻力的综合状态; 因子 4 中体重 Weight(0.724)、身高 Height(0.816) 因子载荷较高, 它们是基础体格指标, 共同决定 BMI; 因子 5 中总胆固醇 TC(0.480)、甘油三酯 TG(0.668) 因子载荷较高, 该因子以内源性甘油三酯代谢为主导。

```
scores <- as.data.frame(fre2$scores)
names(scores) <- paste0("Factor", 1:5)
```

```

factor_names <- c(
  "胆红素代谢因子",    # Factor1: T-BIL/IB
  "肝细胞损伤因子",    # Factor2: ALT/AST
  "动脉血压因子",      # Factor3: Sbp/Dbp
  "体型规模因子",      # Factor4: Weight/Height
  "甘油三酯代谢因子"  # Factor5: TG/TC
)

cor_res <- data.frame(
  因子 = factor_names,
  r = sapply(scores, function(x) cor(x, Age, use = "complete.obs")),
  p = sapply(scores, function(x) cor.test(x, Age)$p.value)
)

print(cor_res, digits = 3, row.names = FALSE)

##          因子      r        p
## 胆红素代谢因子 0.00909 5.63e-01
## 肝细胞损伤因子 -0.03436 2.86e-02
## 动脉血压因子   0.32456 3.33e-100
## 体型规模因子   -0.07162 4.95e-06
## 甘油三酯代谢因子  0.13616 2.96e-18

```

计算各个因子得分与年龄的皮尔逊相关性，结果如上表所示。在 95% 的置信水平下，除了因子 1 与年龄的相关性不显著外，其余 4 个因子均与年龄具有显著的相关性。其中，因子 2 和因子 4 与年龄呈负相关关系，相关系数为-0.03 和-0.07；因子 3 和因子 5 与年龄呈正相关，相关系数为 0.32 和 0.14。