第 11 章　主成分分析与因子分析

1. 请结合 R 内置 stackloss 数据集进行主成分分析。

【参考答案】主成分分析的主要过程包括：数据预处理、提取主成分、确定主成分个数、计算主成分得分等步骤。

(1)数据预处理。

查看 stackloss 数据集基本信息。

```
> str(stackloss)
'data.frame':   21 obs. of  4 variables:
 $ Air.Flow  : num  80 80 75 62 62 62 62 62 58 58 ...
 $ Water.Temp: num  27 27 25 24 22 23 24 24 23 18 ...
 $ Acid.Conc.: num  89 88 90 87 87 87 93 93 87 80 ...
 $ stack.loss: num  42 37 37 28 18 18 19 20 15 14 ...
> head(stackloss)
  Air.Flow Water.Temp Acid.Conc. stack.loss
1       80         27         89         42
2       80         27         88         37
3       75         25         90         37
4       62         24         87         28
5       62         22         87         18
6       62         23         87         18
```

(2)提取主成分。

使用 R 语言自带的 princomp( )函数进行主成分分析。代码如下：

```
> stackloss.pcr <- princomp(stackloss, cor = T)
> summary(stackloss.pcr,loadings = T)
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4
Standard deviation     1.7313208 0.8567640 0.46401660 0.23059137
Proportion of Variance 0.7493679 0.1835111 0.05382785 0.01329309
Cumulative Proportion  0.7493679 0.9328791 0.98670691 1.00000000

Loadings:
           Comp.1 Comp.2 Comp.3 Comp.4
Air.Flow    0.547         0.620  0.557
Water.Temp  0.524  0.257 -0.758  0.291
Acid.Conc.  0.350 -0.927 -0.106
stack.loss  0.551  0.262  0.173 -0.773
```

参数 cor=T 表示从相关系数出发进行分析。上述结果表明第一主成分 Comp. 1 可以解释整体方差的 74.9%，前两个主成分可以累积解释整体方差的 93.29%。

(3)确定主成分个数。

前两个主成分的累积贡献率高达 93.28%，因此可以考虑舍去剩下的两个主成分，以达到降维的目的。

还可以画出主成分的碎石图(见图 11-1)，以便更为直观地观察主成分特征值的趋势和走向。代码如下：

```
> screeplot(stackloss.pcr, type = 'lines')
> abline(1,0,col='red')
```

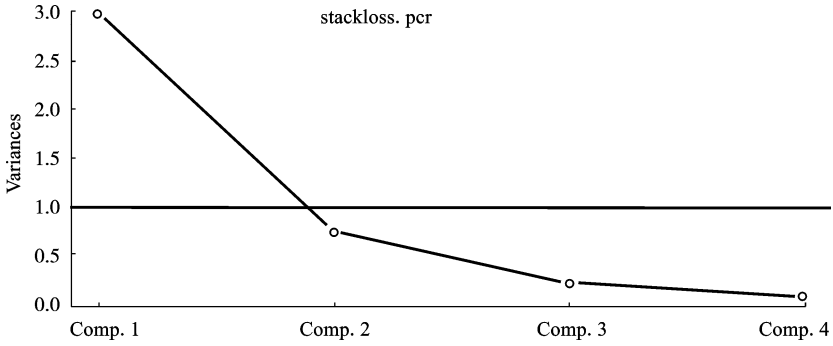

图 11-1　碎石图

(4)计算主成分得分。

利用 predict( )函数计算原数据集中每个观察对象(样本)在主成分上的得分。

```
> predict(stackloss.pcr)
           Comp.1      Comp.2       Comp.3      Comp.4
 [1,]  3.74034512  0.83712372  0.276562083 -0.17622946
 [2,]  3.39568717  0.88244179  0.209764347  0.22994657
 [3,]  2.88416459  0.31534195  0.314269884 -0.30369902
 [4,]  1.21934215  0.40596365 -0.436992247 -0.45558486
 [5,]  0.32398126 -0.02419082 -0.119549636  0.13415481
 [6,]  0.49390097  0.05900065 -0.365268391  0.22864745
 [7,]  1.12075617 -0.89465378 -0.714785938  0.14438629
 [8,]  1.17630831 -0.86827663 -0.697386448  0.06651379
 [9,]  0.08286246 -0.05717819 -0.694692897  0.21343838
[10,] -1.29056881  0.74091420  0.657899384 -0.06345746
[11,] -0.68849380 -0.85392038  0.476101962 -0.21477944
[12,] -0.98086288 -0.78628516  0.724620941 -0.21458603
[13,] -1.32343080  0.30737505  0.565301487  0.13653292
[14,] -0.36208949 -1.53229859  0.114785372 -0.03179602
[15,] -1.51057083 -1.08627805 -0.182747050 -0.24519758
[16,] -1.76681465 -0.58104368 -0.139547400 -0.11688443
[17,] -2.47790392  2.00937874 -0.085070675  0.13512546
[18,] -2.00962335  0.76895184 -0.226468670  0.01743059
[19,] -1.71725427  0.70131662 -0.474987649  0.01723718
[20,] -0.88357383  0.56074292  0.004848918 -0.11038506
[21,]  0.57383844 -0.90442584  0.793342623  0.60918592
```

注：也可以参考本章 11.1.3 节使用第三方工具包 psych 实现主成分分析。

2. 请结合第 8 章数据集 g56. csv 进行因子分析。

【参考答案】因子分析主要过程包含：数据预处理、确定因子数目、因子分析及结果解读。

(1)数据预处理，目的是从原始数据集中得到协方差矩阵或者相关矩阵。

数据集 g56. csv 共有 11 个属性，这里我们仅保留身高(厘米)、体重(千克)、50 米跑(秒)、坐位体前屈(厘米)、一分钟跳绳(次)、一分钟仰卧起坐(次)、肺活量(毫升)、50 米×8 往返跑(秒)用于因子分析。代码如下：

```
> my.g56 <- g56[,4:11]
> str(my.g56)
'data.frame':   120 obs. of  8 variables:
 $ height: int  171 145 149 155 140 148 153 145 152 171 ...
 $ weight: num  72 29 45 55 35 41 34 40 43 54 ...
 $ m50   : num  8.9 8.6 9.4 8.3 8.3 8.7 8.4 8.2 8.5 7.6 ...
 $ sitFB : int  7 -4 1 4 9 -3 6 2 8 4 ...
 $ jp    : int  53 77 73 50 70 26 109 56 137 37 ...
 $ sitUPs: int  28 24 27 25 45 28 25 36 31 48 ...
 $ lungCp: int  4380 1458 2660 1468 2581 2601 2274 1775 2710 3030 ...
 $ m50x8 : num  2.07 2.12 2.05 2.02 2.15 2.08 1.88 1.77 1.85 1.48 ...
```

(2)确定因子数目。

使用 fa. parallel( )函数，分析并判断因子数目的选择。代码及执行结果如下，执行结果如图 11-2 所示。

```
> library(psych)
> fa.parallel(my.g56,fa='fa',n.iter = 100, show.legend = F)
Parallel analysis suggests that the number of factors =  3  and the number of
 components =  NA
```

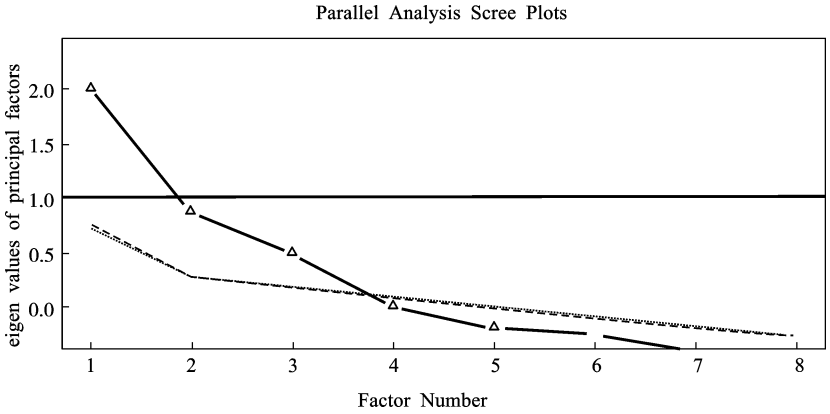

图 11-2　碎石图分析

上述分析结果建议的因子数为 3。

(3)因子分析。

R 语言提供有 factanal( )函数用于因子分析。它可以从样本数据、样本的协方差矩阵或者相关矩阵进行最大似然的因子分析，并可直接给出方差最大的载荷因子矩阵。代码及分析结果如下(通过参数 covmat 指明从协方差矩阵展开分析)：

```
> factanal(factors = 3, covmat = cov(my.g56))

Call:
factanal(factors = 3, covmat = cov(my.g56))

Uniquenesses:
height weight    m50  sitFB     jp sitUPs lungCp  m50x8
 0.368  0.057  0.407  0.415  0.678  0.734  0.435  0.604
```

```
Loadings:
       Factor1 Factor2 Factor3
height  0.783  -0.138
weight  0.952           0.167
m50             0.195   0.745
sitFB           0.756   0.112
jp              0.556  -0.110
sitUPs  0.117   0.128  -0.485
lungCp  0.693   0.134  -0.259
m50x8   0.142  -0.407   0.459

               Factor1 Factor2 Factor3
SS loadings      2.033   1.146   1.121
Proportion Var   0.254   0.143   0.140
Cumulative Var   0.254   0.397   0.538
```

(4)结果解读。

从上述的因子载荷矩阵可以看出，身高、体重和肺活量在第一个因子 Factor1 上载荷较高，表明该因子衡量的体型；坐位体前屈(sitFB)、一分钟跳绳(jp)在因子 Factor2 上载荷较大(分别为 0.756 和 0.556)，且与 50 米×8 往返跑成绩有较为明显的负相关，表明存在一个身体灵活或者柔韧性相关的因子①；50 米跑(m50)在因子 Factor3 上载荷较大(0.745)，表明存在爆发力(短跑)相关因子，该因子与 50 米×8 往返跑有较为明显的正相关，并与一分钟仰卧起坐(sitUPs)有较为明显的负相关。

需要注意的是，上述分析结果中，Cumulative Var 描述三个因子的累积方差贡献率为 53.8%，表明依据该份调查数据得到的分析模型结果不是很理想，但也在一定程度上表明当前体质健康指标体系的科学性和专业性。

3. 请根据《国家学生体质健康标准》，将数据集 g56.csv 中数据转换为标准分，然后进行主成分分析，比较并解释转换前后得到的分析结果的异同。

【参考答案】

(1)参照《国家学生体质健康标准(2014 年修订)》将 g56.csv 体质健康测量数据换算为标准分(数据文件名为 g56score.csv)。

(2)按照 11.3 节的方法，依次读取数据集(数据文件名为 g56score.csv)，挑选参与分析的 7 个属性：bmi、m50、sitFB、jp、sitUPs、lungCp、m50x8，其中，bmi 指数是国际通用衡量人体胖瘦程度以及是否健康的一个标准，计算公式：体重/身高的平方(国际单位：kg/m^2)。

(3)调用碎石分析，确定主成分个数。

(4)使用 principal()函数进行主成分分析，并尝试解读结果。

(5)参照上述步骤(1)到(4)，对数据集 g56.csv 中的体质健康测量数据进行主成分分析，并尝试解释结果。

4. 请结合 boston 数据集，使用主成分分析降维结果，建立房价预测回归模型。

【参考答案】

(1)读取 boston 数据集，查看基本结构。

```
> boston <- read.csv('boston.csv')
```

① 此结论不代表专业角度解释，仅仅是常识性直观理解。

```
> str(boston)
'data.frame':   506 obs. of  14 variables:
 $ CRIM   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
 $ ZN     : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
 $ INDUS  : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
 $ CHAS   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ NOX    : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524
...
 $ RM     : num  6.58 6.42 7.18 7 7.15 ...
 $ AGE    : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
 $ DIS    : num  4.09 4.97 4.97 6.06 6.06 ...
 $ RAD    : int  1 2 2 3 3 3 5 5 5 5 ...
 $ TAX    : int  296 242 242 222 222 222 311 311 311 311 ...
 $ PTRATIO: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
 $ B      : num  397 397 393 395 397 ...
 $ LSTAT  : num  4.98 9.14 4.03 2.94 5.33 ...
 $ MEDV   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

(2)按照 11.3 节的方法进行主成分分析，计算并查看主成分得分。

```
> head(my.boston.pc$scores)
             RC2        RC1         RC3
[1,] -0.22674609 -0.6848711 -0.84106398
[2,]  0.01523583 -0.9872704  0.05636746
[3,] -0.29067219 -0.6955108 -0.63692076
[4,] -0.74687100 -0.7073619 -0.24669540
[5,] -0.64735625 -0.7010806 -0.29143786
[6,] -0.54628707 -0.9052494  0.21059123
```

(3)构建用于回归预测的数据集。

```
> my.boston2 <- data.frame(my.boston.pc$scores,MEDV = boston$MEDV)
> str(my.boston2)
'data.frame':   506 obs. of  4 variables:
 $ RC2 : num  -0.2267 0.0152 -0.2907 -0.7469 -0.6474 ...
 $ RC1 : num  -0.685 -0.987 -0.696 -0.707 -0.701 ...
 $ RC3 : num  -0.8411 0.0564 -0.6369 -0.2467 -0.2914 ...
 $ MEDV: num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

(4)建立多元回归模型。

```
> my.boston2.lm <- lm(MEDV~., data = my.boston2)
> summary(my.boston2.lm)

Call:
lm(formula = MEDV ~ ., data = my.boston2)

Residuals:
    Min      1Q  Median      3Q     Max 
-16.774  -3.082  -0.651   2.091  35.000 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  22.5328     0.2474   91.06   <2e-16 ***
RC2          -2.9388     0.2477  -11.87   <2e-16 ***
RC1          -2.6028     0.2477  -10.51   <2e-16 ***
RC3          -6.1951     0.2477  -25.01   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.566 on 502 degrees of freedom
Multiple R-squared:  0.6359,	Adjusted R-squared:  0.6337 
F-statistic: 292.3 on 3 and 502 DF,  p-value: < 2.2e-16
```

(5)结果解读。

F 检验的 p-value：<2.2e-16，远远小于 0.05，说明模型通过了 F 检验；再看截距项(Intercept)和三个系数，P 值均远小于 0.05，都通过了 t 检验。

得到回归方程：MEDV = 22.5328 - 2.6028 * RC1 - 2.9388 * RC2 - 6.1951 * RC3。

(6)读者还可以尝试直接使用原始数据集进行回归建模，并与 PCA 降维后的回归模型比较。