Applied Statistic HW9

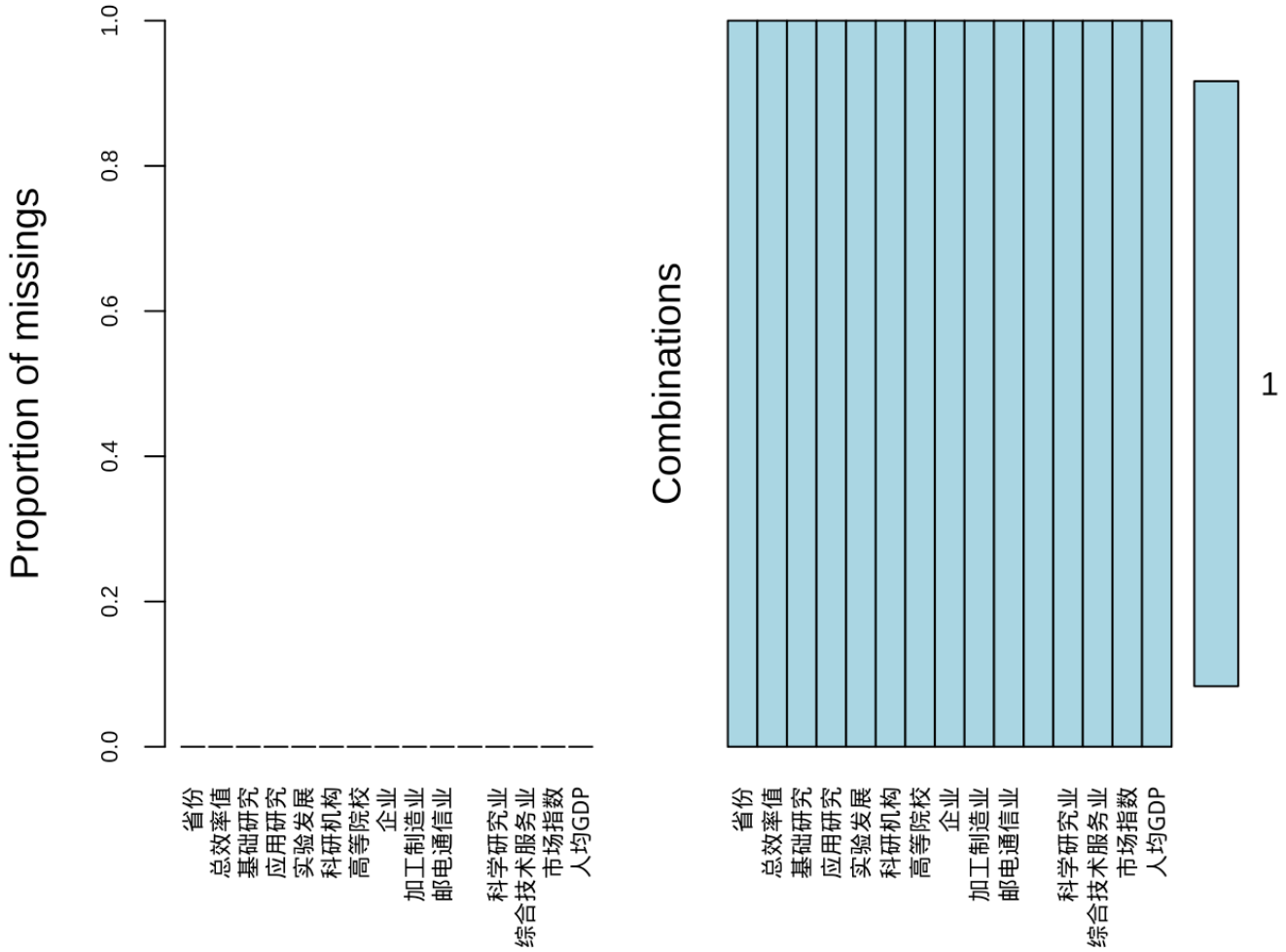
2020270026 王姿文

2020/12/03

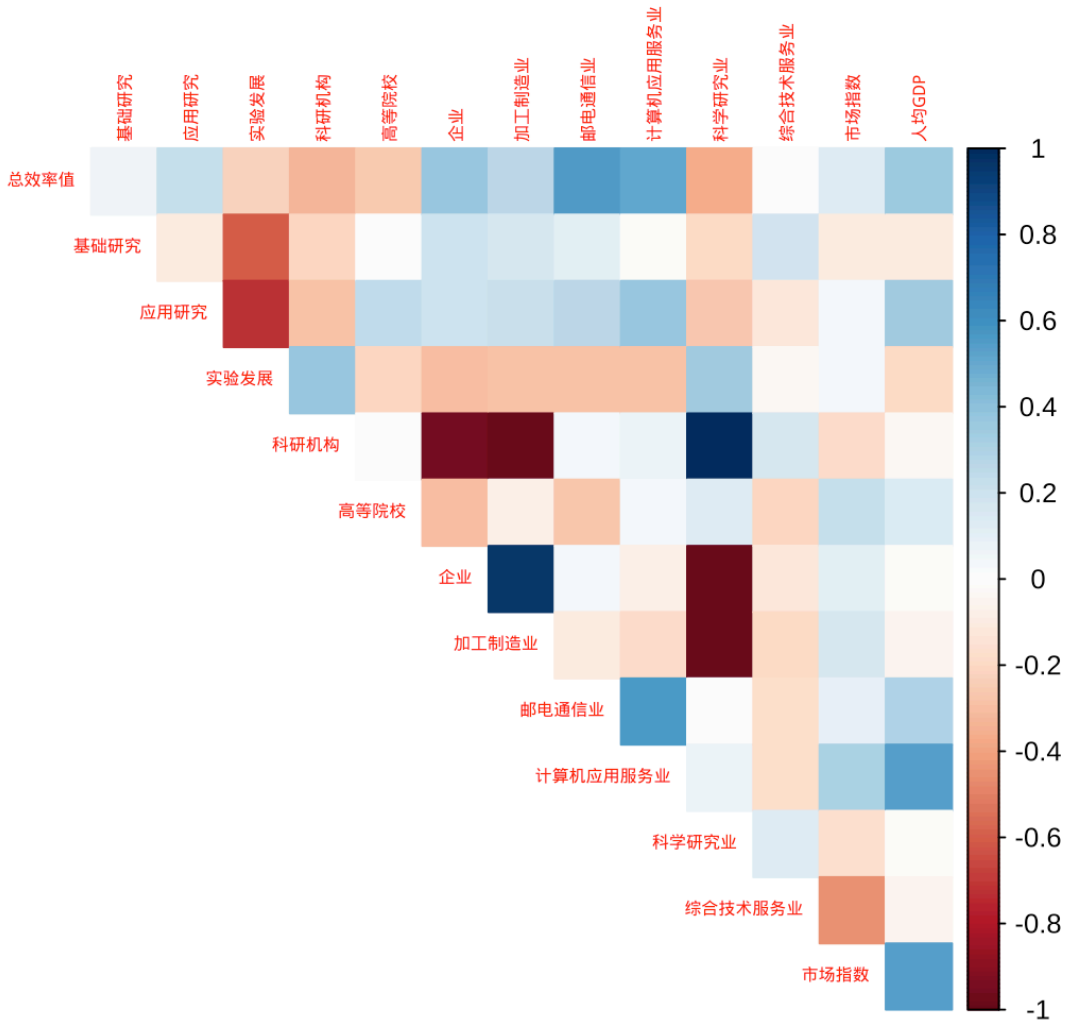
1. 根据《2000年各地区科技状况统计有关数据一览表》的数据

1.1 进行多元线性回归分析，并尝试用逐步回归法进行变量选择

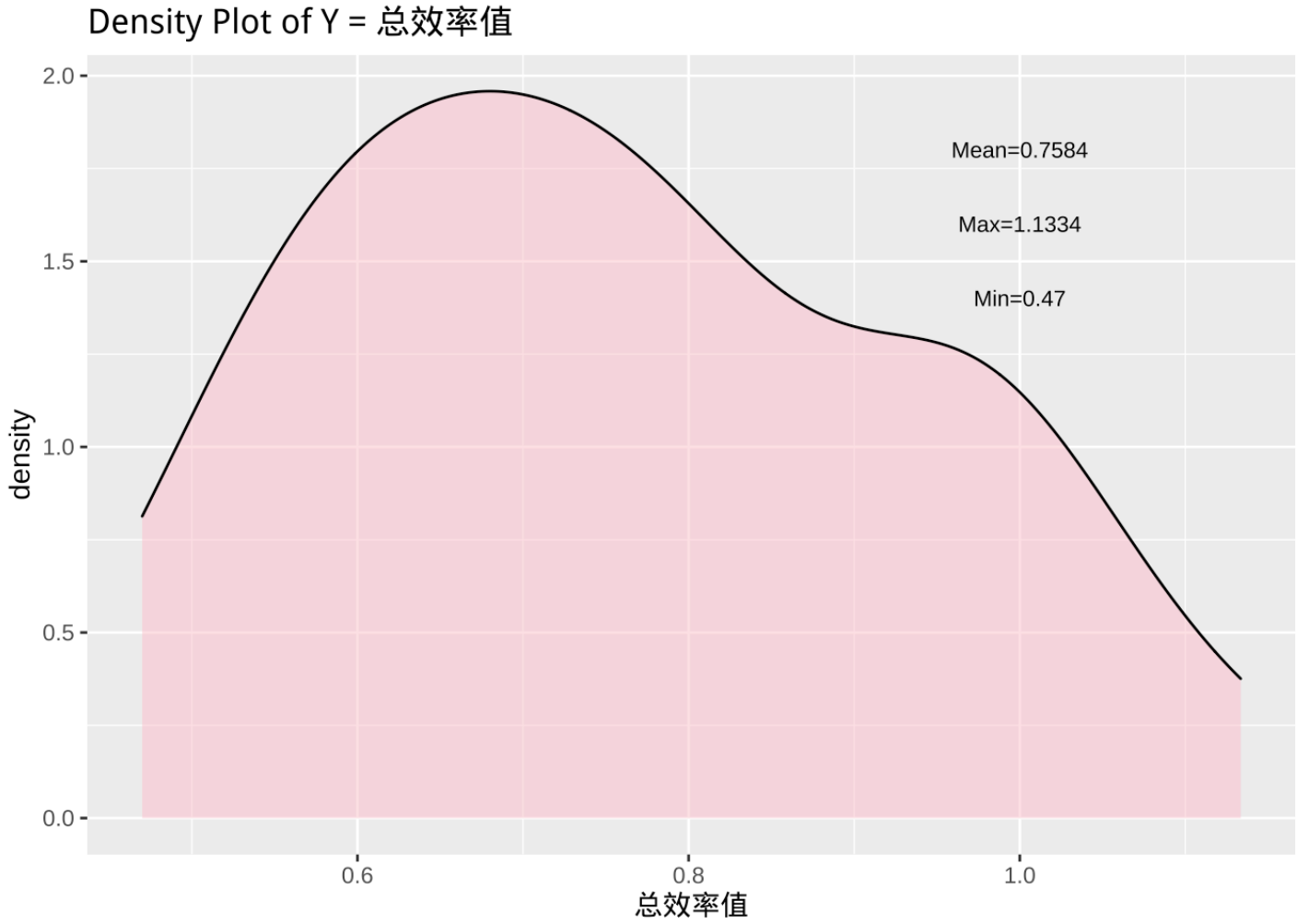
首先简单地检查是否有遗失值问题。下图中的左图为**Proportion Plot of Missing Data**，横轴为变数名称，纵轴为遗失值比例，可以看出无遗失值。



接下来检查是否有共线性，虽然单看**Correlation Plot**确实有共线性的可能，但若使用VIF（方差膨胀因子）来衡量多重共线性，可以得知没有若单变量的VIF超过2，因此没有严重的共线性问题。


基础研究
应用研究
实验发展
科研机构
高等院校
企业
加工制造业
邮电通信业
计算机应用服务业
科学研究业
综合技术服务业
市场指数
人均GDP
总效率值
基础研究
应用研究
实验发展
科研机构
高等院校
企业
加工制造业
邮电通信业
计算机应用服务业
科学研究业
综合技术服务业
市场指数
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

再由下看出Y的分布正常，不需要特别去转换数据。



下面分别使用forward、backward、both stepwise regression：

Forward

此處以AIC作為衡量線性回歸優劣的指標，若AIC越小則線性回歸的解釋程度越好。一開始創建一個只有截距項的模型，接著一個個選入每一步AIC最低的變量，到了最後一步時，由於再新增 实验发展 也不會使得模型的AIC下降，因此

$$总效率值 = 邮电通信业 + 科学研究业 + 加工制造业$$

```
## Start:  AIC=-103.1
## 总效率值 ~ 1
##
##                      Df Sum of Sq   RSS  AIC
## + 邮电通信业          1    0.2739 0.628 -112
## + 计算机应用服务业    1    0.2435 0.658 -111
## + 企业                1    0.1298 0.772 -106
## + 科学研究业          1    0.1216 0.780 -106
## + 人均GDP             1    0.1165 0.785 -105
## + 科研机构            1    0.0931 0.808 -104
## + 加工制造业          1    0.0632 0.838 -103
## <none>                           0.902 -103
## + 高等院校            1    0.0580 0.844 -103
## + 应用研究            1    0.0470 0.855 -103
## + 实验发展            1    0.0461 0.855 -103
## + 市场指数            1    0.0153 0.886 -102
```

```
## + 基础研究           1    0.0037 0.898 -101
## + 综合技术服务业     1    0.0001 0.901 -101
##
## Step:  AIC=-112
## 总效率值 ~ 邮电通信业
##
##                     Df Sum of Sq   RSS  AIC
## + 科学研究业         1    0.1180 0.510 -116
## + 企业               1    0.1153 0.512 -116
## + 科研机构           1    0.1068 0.521 -116
## + 加工制造业         1    0.0944 0.533 -115
## + 计算机应用服务业   1    0.0566 0.571 -113
## <none>                           0.628 -112
## + 人均GDP            1    0.0391 0.589 -112
## + 综合技术服务业     1    0.0102 0.617 -110
## + 高等院校           1    0.0100 0.618 -110
## + 应用研究           1    0.0066 0.621 -110
## + 市场指数           1    0.0059 0.622 -110
## + 实验发展           1    0.0044 0.623 -110
## + 基础研究           1    0.0000 0.628 -110
##
## Step:  AIC=-116.3
## 总效率值 ~ 邮电通信业 + 科学研究业
##
##                     Df Sum of Sq   RSS  AIC
## + 加工制造业         1    0.0882 0.421 -120
## + 计算机应用服务业   1    0.0745 0.435 -119
## + 人均GDP            1    0.0380 0.472 -117
## <none>                           0.510 -116
## + 综合技术服务业     1    0.0217 0.488 -116
## + 科研机构           1    0.0117 0.498 -115
## + 基础研究           1    0.0044 0.505 -114
## + 实验发展           1    0.0037 0.506 -114
## + 高等院校           1    0.0031 0.506 -114
## + 市场指数           1    0.0005 0.509 -114
## + 应用研究           1    0.0002 0.509 -114
## + 企业               1    0.0001 0.509 -114
##
## Step:  AIC=-120
## 总效率值 ~ 邮电通信业 + 科学研究业 + 加工制造业
##
##                     Df Sum of Sq   RSS  AIC
## <none>                           0.421 -120
## + 实验发展           1   0.02048 0.401 -120
## + 计算机应用服务业   1   0.01748 0.404 -119
## + 综合技术服务业     1   0.01730 0.404 -119
## + 基础研究           1   0.01307 0.408 -119
## + 人均GDP            1   0.01122 0.410 -119
## + 市场指数           1   0.00441 0.417 -118
## + 应用研究           1   0.00418 0.417 -118
## + 科研机构           1   0.00173 0.420 -118
## + 高等院校           1   0.00092 0.420 -118
## + 企业               1   0.00031 0.421 -118
```

但可以再进一步检测forward stepwise regression的结果，得出Adjusted R-squared=0.479 ，且 邮电通信业 不为显著变量，因此forward stepwise regression还是存有缺陷。

```
##
## Call:
## lm(formula = 总效率值 ~ 邮电通信业 + 科学研究业 +
##     加工制造业, data = Q1L)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.1680 -0.0791 -0.0130  0.0496  0.2989
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.58       1.20    2.99    0.006 **
## 邮电通信业       3.04       2.44    1.25    0.224
## 科学研究业      -3.29       1.25   -2.63    0.014 *
## 加工制造业      -2.84       1.22   -2.33    0.028 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.127 on 26 degrees of freedom
## Multiple R-squared:  0.533,  Adjusted R-squared:  0.479
## F-statistic: 9.88 on 3 and 26 DF,  p-value: 0.00016
```

Backward

和forward不同的是，这次需要先把所有变量放入模型再一个个剔除，一样用AIC来衡量，最后一步是可以看出若再剔除 综合技术服务业 ，模型也不会更好，所以最终结果为

$$总效率值 = 加工制造业 + 邮电通信业 + 计算机应用服务业 + 综合技术服务业$$

```
## Start:  AIC=-107.3
## 总效率值 ~ 基础研究 + 应用研究 + 实验发展 + 科研机构 +
##     高等院校 + 企业 + 加工制造业 + 邮电通信业 +
##     计算机应用服务业 + 科学研究业 + 综合技术服务业 +
##     市场指数 + 人均GDP
##
##                    Df Sum of Sq   RSS  AIC
## - 基础研究          1   0.00002 0.330 -109
## - 实验发展          1   0.00002 0.330 -109
## - 应用研究          1   0.00004 0.330 -109
## - 企业              1   0.00830 0.338 -109
## - 高等院校          1   0.00833 0.338 -109
## - 科研机构          1   0.00840 0.338 -109
## - 科学研究业        1   0.01004 0.340 -108
## - 综合技术服务业    1   0.01018 0.340 -108
## - 加工制造业        1   0.01018 0.340 -108
## - 邮电通信业        1   0.01036 0.340 -108
## - 计算机应用服务业  1   0.01046 0.340 -108
## - 市场指数          1   0.01309 0.343 -108
## - 人均GDP           1   0.01722 0.347 -108
## <none>                          0.330 -107
```

```
##
## Step:  AIC=-109.3
## 总效率值 ~ 应用研究 + 实验发展 + 科研机构 + 高等院校 +
##     企业 + 加工制造业 + 邮电通信业 + 计算机应用服务业 +
##     科学研究业 + 综合技术服务业 + 市场指数 +
##     人均GDP
##
##                      Df Sum of Sq   RSS  AIC
## - 实验发展            1   0.00007 0.330 -111
## - 企业                1   0.01004 0.340 -110
## - 高等院校            1   0.01007 0.340 -110
## - 科研机构            1   0.01015 0.340 -110
## - 科学研究业          1   0.01061 0.340 -110
## - 综合技术服务业      1   0.01076 0.340 -110
## - 加工制造业          1   0.01077 0.340 -110
## - 邮电通信业          1   0.01096 0.341 -110
## - 计算机应用服务业    1   0.01106 0.341 -110
## - 应用研究            1   0.01158 0.341 -110
## - 市场指数            1   0.01320 0.343 -110
## - 人均GDP             1   0.01721 0.347 -110
## <none>                            0.330 -109
##
## Step:  AIC=-111.3
## 总效率值 ~ 应用研究 + 科研机构 + 高等院校 + 企业 +
##     加工制造业 + 邮电通信业 + 计算机应用服务业 +
##     科学研究业 + 综合技术服务业 + 市场指数 +
##     人均GDP
##
##                      Df Sum of Sq   RSS  AIC
## - 企业                1   0.00999 0.340 -112
## - 高等院校            1   0.01003 0.340 -112
## - 科研机构            1   0.01011 0.340 -112
## - 科学研究业          1   0.01092 0.341 -112
## - 综合技术服务业      1   0.01108 0.341 -112
## - 加工制造业          1   0.01110 0.341 -112
## - 邮电通信业          1   0.01129 0.341 -112
## - 计算机应用服务业    1   0.01141 0.341 -112
## - 市场指数            1   0.01328 0.343 -112
## - 人均GDP             1   0.01840 0.348 -112
## - 应用研究            1   0.01955 0.349 -112
## <none>                            0.330 -111
##
## Step:  AIC=-112.4
## 总效率值 ~ 应用研究 + 科研机构 + 高等院校 + 加工制造业 +
##     邮电通信业 + 计算机应用服务业 + 科学研究业 +
##     综合技术服务业 + 市场指数 + 人均GDP
##
##                      Df Sum of Sq   RSS  AIC
## - 高等院校            1    0.0118 0.351 -113
## - 应用研究            1    0.0163 0.356 -113
## - 市场指数            1    0.0175 0.357 -113
## - 科学研究业          1    0.0178 0.357 -113
## - 综合技术服务业      1    0.0181 0.358 -113
## - 加工制造业          1    0.0181 0.358 -113
```

```
## - 邮电通信业          1    0.0183 0.358 -113
## - 计算机应用服务业   1    0.0184 0.358 -113
## <none>                              0.340 -112
## - 科研机构            1    0.0243 0.364 -112
## - 人均GDP             1    0.0303 0.370 -112
##
## Step:  AIC=-113.4
## 总效率值 ~ 应用研究 + 科研机构 + 加工制造业 +
##     邮电通信业 + 计算机应用服务业 + 科学研究业 +
##     综合技术服务业 + 市场指数 + 人均GDP
##
##                       Df Sum of Sq   RSS  AIC
## - 应用研究            1   0.00919 0.361 -115
## - 市场指数            1   0.00969 0.361 -115
## - 科研机构            1   0.01328 0.365 -114
## - 科学研究业          1   0.01377 0.365 -114
## - 加工制造业          1   0.01390 0.365 -114
## - 综合技术服务业      1   0.01393 0.365 -114
## - 邮电通信业          1   0.01412 0.366 -114
## - 计算机应用服务业   1   0.01421 0.366 -114
## - 人均GDP             1   0.02188 0.373 -114
## <none>                              0.351 -113
##
## Step:  AIC=-114.6
## 总效率值 ~ 科研机构 + 加工制造业 + 邮电通信业 +
##     计算机应用服务业 + 科学研究业 + 综合技术服务业 +
##     市场指数 + 人均GDP
##
##                       Df Sum of Sq   RSS  AIC
## - 市场指数            1   0.00561 0.366 -116
## - 科学研究业          1   0.01616 0.377 -115
## - 加工制造业          1   0.01631 0.377 -115
## - 综合技术服务业      1   0.01635 0.377 -115
## - 人均GDP             1   0.01651 0.377 -115
## - 邮电通信业          1   0.01653 0.377 -115
## - 计算机应用服务业   1   0.01662 0.377 -115
## - 科研机构            1   0.01868 0.379 -115
## <none>                              0.361 -115
##
## Step:  AIC=-116.2
## 总效率值 ~ 科研机构 + 加工制造业 + 邮电通信业 +
##     计算机应用服务业 + 科学研究业 + 综合技术服务业 +
##     人均GDP
##
##                       Df Sum of Sq   RSS  AIC
## - 人均GDP             1    0.0109 0.377 -117
## - 科学研究业          1    0.0148 0.381 -117
## - 加工制造业          1    0.0150 0.381 -117
## - 综合技术服务业      1    0.0150 0.381 -117
## - 邮电通信业          1    0.0152 0.381 -117
## - 计算机应用服务业   1    0.0152 0.381 -117
## - 科研机构            1    0.0158 0.382 -117
## <none>                              0.366 -116
##
```

```
## Step:  AIC=-117.3
## 总效率值 ~ 科研机构 + 加工制造业 + 邮电通信业 +
##     计算机应用服务业 + 科学研究业 + 综合技术服务业
##
##                     Df Sum of Sq   RSS  AIC
## - 科研机构           1    0.0107 0.388 -118
## - 科学研究业         1    0.0124 0.390 -118
## - 加工制造业         1    0.0126 0.390 -118
## - 综合技术服务业     1    0.0126 0.390 -118
## - 邮电通信业         1    0.0128 0.390 -118
## - 计算机应用服务业   1    0.0129 0.390 -118
## <none>                           0.377 -117
##
## Step:  AIC=-118.5
## 总效率值 ~ 加工制造业 + 邮电通信业 + 计算机应用服务业 +
##     科学研究业 + 综合技术服务业
##
##                     Df Sum of Sq   RSS  AIC
## - 科学研究业         1    0.0159 0.404 -119
## - 加工制造业         1    0.0159 0.404 -119
## - 综合技术服务业     1    0.0160 0.404 -119
## - 计算机应用服务业   1    0.0162 0.404 -119
## - 邮电通信业         1    0.0163 0.404 -119
## <none>                           0.388 -118
##
## Step:  AIC=-119.2
## 总效率值 ~ 加工制造业 + 邮电通信业 + 计算机应用服务业 +
##     综合技术服务业
##
##                     Df Sum of Sq   RSS  AIC
## <none>                           0.404 -119
## - 综合技术服务业     1    0.0450 0.449 -118
## - 邮电通信业         1    0.0974 0.501 -115
## - 计算机应用服务业   1    0.0999 0.504 -115
## - 加工制造业         1    0.1521 0.556 -112
```

进一步检测backward stepwise regression的结果，得出Adjusted R-squared=0.481 ，且 综合技术服务业 不为显着变量，因此backward stepwise regression还是存有缺陷，但Adjusted R-squared比forward stepwise regression高。

```
##
## Call:
## lm(formula = 总效率值 ~ 加工制造业 + 邮电通信业 +
##     计算机应用服务业 + 综合技术服务业, data = Q1L)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.17715 -0.08547 -0.00708  0.06170  0.26329
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.290      0.116    2.51   0.0191 *
## 加工制造业           0.446      0.145    3.07   0.0051 **
## 邮电通信业           5.181      2.110    2.46   0.0214 *
## 计算机应用服务业     4.849      1.950    2.49   0.0199 *
## 综合技术服务业       2.470      1.480    1.67   0.1075
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.127 on 25 degrees of freedom
## Multiple R-squared:  0.552,  Adjusted R-squared:  0.481
## F-statistic: 7.71 on 4 and 25 DF,  p-value: 0.000344
```

Both

Both stepwise regression則是雙向去挑選，此處分別從Null Regression 和 Full Regression當起點，得出不同結論： - From Null Regression

总效率值 科学研究业 + 加工制造业 + 综合技术服务业

```
## Start:  AIC=-103.1
## 总效率值 ~ 1
##
##                     Df Sum of Sq   RSS  AIC
## + 邮电通信业         1    0.2739 0.628 -112
## + 计算机应用服务业   1    0.2435 0.658 -111
## + 企业               1    0.1298 0.772 -106
## + 科学研究业         1    0.1216 0.780 -106
## + 人均GDP            1    0.1165 0.785 -105
## + 科研机构           1    0.0931 0.808 -104
## + 加工制造业         1    0.0632 0.838 -103
## <none>                            0.902 -103
## + 高等院校           1    0.0580 0.844 -103
## + 应用研究           1    0.0470 0.855 -103
## + 实验发展           1    0.0461 0.855 -103
## + 市场指数           1    0.0153 0.886 -102
## + 基础研究           1    0.0037 0.898 -101
## + 综合技术服务业     1    0.0001 0.901 -101
##
## Step:  AIC=-112
## 总效率值 ~ 邮电通信业
##
##                     Df Sum of Sq   RSS  AIC
```

```
## + 科学研究业         1    0.1180 0.510 -116
## + 企业               1    0.1153 0.512 -116
## + 科研机构           1    0.1068 0.521 -116
## + 加工制造业         1    0.0944 0.533 -115
## + 计算机应用服务业   1    0.0566 0.571 -113
## <none>                           0.628 -112
## + 人均GDP            1    0.0391 0.589 -112
## + 综合技术服务业     1    0.0102 0.617 -110
## + 高等院校           1    0.0100 0.618 -110
## + 应用研究           1    0.0066 0.621 -110
## + 市场指数           1    0.0059 0.622 -110
## + 实验发展           1    0.0044 0.623 -110
## + 基础研究           1    0.0000 0.628 -110
## - 邮电通信业         1    0.2739 0.902 -103
##
## Step:  AIC=-116.3
## 总效率值 ~ 邮电通信业 + 科学研究业
##
##                     Df Sum of Sq   RSS  AIC
## + 加工制造业         1    0.0882 0.421 -120
## + 计算机应用服务业   1    0.0745 0.435 -119
## + 人均GDP            1    0.0380 0.472 -117
## <none>                           0.510 -116
## + 综合技术服务业     1    0.0217 0.488 -116
## + 科研机构           1    0.0117 0.498 -115
## + 基础研究           1    0.0044 0.505 -114
## + 实验发展           1    0.0037 0.506 -114
## + 高等院校           1    0.0031 0.506 -114
## + 市场指数           1    0.0005 0.509 -114
## + 应用研究           1    0.0002 0.509 -114
## + 企业               1    0.0001 0.509 -114
## - 科学研究业         1    0.1180 0.628 -112
## - 邮电通信业         1    0.2703 0.780 -106
##
## Step:  AIC=-120
## 总效率值 ~ 邮电通信业 + 科学研究业 + 加工制造业
##
##                     Df Sum of Sq   RSS  AIC
## - 邮电通信业         1    0.0252 0.447 -120
## <none>                           0.421 -120
## + 实验发展           1    0.0205 0.401 -120
## + 计算机应用服务业   1    0.0175 0.404 -119
## + 综合技术服务业     1    0.0173 0.404 -119
## + 基础研究           1    0.0131 0.408 -119
## + 人均GDP            1    0.0112 0.410 -119
## + 市场指数           1    0.0044 0.417 -118
## + 应用研究           1    0.0042 0.417 -118
## + 科研机构           1    0.0017 0.420 -118
## + 高等院校           1    0.0009 0.420 -118
## + 企业               1    0.0003 0.421 -118
## - 加工制造业         1    0.0882 0.510 -116
## - 科学研究业         1    0.1118 0.533 -115
##
## Step:  AIC=-120.2
```

```
## 总效率值 ~ 科学研究业 + 加工制造业
##
##                      Df Sum of Sq   RSS  AIC
## + 综合技术服务业     1     0.042 0.404 -121
## <none>                           0.447 -120
## + 邮电通信业         1     0.025 0.421 -120
## + 计算机应用服务业   1     0.023 0.424 -120
## + 实验发展           1     0.019 0.428 -120
## + 基础研究           1     0.014 0.432 -119
## + 人均GDP            1     0.011 0.436 -119
## + 市场指数           1     0.009 0.438 -119
## + 科研机构           1     0.004 0.443 -118
## + 高等院校           1     0.003 0.444 -118
## + 应用研究           1     0.003 0.444 -118
## + 企业               1     0.001 0.445 -118
## - 加工制造业         1     0.333 0.780 -106
## - 科学研究业         1     0.392 0.838 -103
##
## Step:  AIC=-121.2
## 总效率值 ~ 科学研究业 + 加工制造业 + 综合技术服务业
##
##                      Df Sum of Sq   RSS  AIC
## <none>                           0.404 -121
## + 实验发展           1     0.024 0.380 -121
## - 综合技术服务业     1     0.042 0.447 -120
## + 科研机构           1     0.013 0.392 -120
## + 应用研究           1     0.011 0.393 -120
## + 基础研究           1     0.008 0.397 -120
## + 高等院校           1     0.007 0.398 -120
## + 人均GDP            1     0.003 0.402 -119
## + 企业               1     0.002 0.402 -119
## + 邮电通信业         1     0.000 0.404 -119
## + 计算机应用服务业   1     0.000 0.404 -119
## + 市场指数           1     0.000 0.404 -119
## - 加工制造业         1     0.373 0.777 -104
## - 科学研究业         1     0.431 0.835 -101
```

进一步检测结果，得出Adjusted R-squared=0.5 ，且 综合技术服务业 不为显着变量

```
##
## Call:
## lm(formula = 总效率值 ~ 科学研究业 + 加工制造业 +
##     综合技术服务业, data = Q1L)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.17716 -0.08703 -0.00757  0.06177  0.25968
##
## Coefficients:
##                Estimate Std. Error t value  Pr(>|t|)
## (Intercept)       5.290      0.905    5.85 0.0000036 ***
## 科学研究业       -5.001      0.950   -5.26 0.0000168 ***
## 加工制造业       -4.554      0.930   -4.90 0.0000441 ***
## 综合技术服务业   -2.533      1.535   -1.65       0.11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.125 on 26 degrees of freedom
## Multiple R-squared:  0.552,  Adjusted R-squared:   0.5
## F-statistic: 10.7 on 3 and 26 DF,  p-value: 0.0000946
```

- From Full Regression

$$总效率值 = 加工制造业 + 邮电通信业 + 计算机应用服务业 + 综合技术服务业$$

```
## Start:  AIC=-107.3
## 总效率值 ~ 基础研究 + 应用研究 + 实验发展 + 科研机构 +
##     高等院校 + 企业 + 加工制造业 + 邮电通信业 +
##     计算机应用服务业 + 科学研究业 + 综合技术服务业 +
##     市场指数 + 人均GDP
##
##                     Df Sum of Sq   RSS  AIC
## - 基础研究           1   0.00002 0.330 -109
## - 实验发展           1   0.00002 0.330 -109
## - 应用研究           1   0.00004 0.330 -109
## - 企业               1   0.00830 0.338 -109
## - 高等院校           1   0.00833 0.338 -109
## - 科研机构           1   0.00840 0.338 -109
## - 科学研究业         1   0.01004 0.340 -108
## - 综合技术服务业     1   0.01018 0.340 -108
## - 加工制造业         1   0.01018 0.340 -108
## - 邮电通信业         1   0.01036 0.340 -108
## - 计算机应用服务业   1   0.01046 0.340 -108
## - 市场指数           1   0.01309 0.343 -108
## - 人均GDP            1   0.01722 0.347 -108
## <none>                           0.330 -107
##
## Step:  AIC=-109.3
## 总效率值 ~ 应用研究 + 实验发展 + 科研机构 + 高等院校 +
##     企业 + 加工制造业 + 邮电通信业 + 计算机应用服务业 +
##     科学研究业 + 综合技术服务业 + 市场指数 +
##     人均GDP
```

```
##
##                      Df Sum of Sq   RSS  AIC
## - 实验发展            1   0.00007 0.330 -111
## - 企业                1   0.01004 0.340 -110
## - 高等院校            1   0.01007 0.340 -110
## - 科研机构            1   0.01015 0.340 -110
## - 科学研究业          1   0.01061 0.340 -110
## - 综合技术服务业      1   0.01076 0.340 -110
## - 加工制造业          1   0.01077 0.340 -110
## - 邮电通信业          1   0.01096 0.341 -110
## - 计算机应用服务业    1   0.01106 0.341 -110
## - 应用研究            1   0.01158 0.341 -110
## - 市场指数            1   0.01320 0.343 -110
## - 人均GDP             1   0.01721 0.347 -110
## <none>                            0.330 -109
## + 基础研究            1   0.00002 0.330 -107
##
## Step:  AIC=-111.3
## 总效率值 ~ 应用研究 + 科研机构 + 高等院校 + 企业 +
##     加工制造业 + 邮电通信业 + 计算机应用服务业 +
##     科学研究业 + 综合技术服务业 + 市场指数 +
##     人均GDP
##
##                      Df Sum of Sq   RSS  AIC
## - 企业                1   0.00999 0.340 -112
## - 高等院校            1   0.01003 0.340 -112
## - 科研机构            1   0.01011 0.340 -112
## - 科学研究业          1   0.01092 0.341 -112
## - 综合技术服务业      1   0.01108 0.341 -112
## - 加工制造业          1   0.01110 0.341 -112
## - 邮电通信业          1   0.01129 0.341 -112
## - 计算机应用服务业    1   0.01141 0.341 -112
## - 市场指数            1   0.01328 0.343 -112
## - 人均GDP             1   0.01840 0.348 -112
## - 应用研究            1   0.01955 0.349 -112
## <none>                            0.330 -111
## + 实验发展            1   0.00007 0.330 -109
## + 基础研究            1   0.00007 0.330 -109
##
## Step:  AIC=-112.4
## 总效率值 ~ 应用研究 + 科研机构 + 高等院校 + 加工制造业 +
##     邮电通信业 + 计算机应用服务业 + 科学研究业 +
##     综合技术服务业 + 市场指数 + 人均GDP
##
##                      Df Sum of Sq   RSS  AIC
## - 高等院校            1   0.01181 0.351 -113
## - 应用研究            1   0.01629 0.356 -113
## - 市场指数            1   0.01750 0.357 -113
## - 科学研究业          1   0.01783 0.357 -113
## - 综合技术服务业      1   0.01807 0.358 -113
## - 加工制造业          1   0.01808 0.358 -113
## - 邮电通信业          1   0.01831 0.358 -113
## - 计算机应用服务业    1   0.01843 0.358 -113
## <none>                            0.340 -112
```

```
## - 科研机构           1   0.02428 0.364 -112
## - 人均GDP            1   0.03029 0.370 -112
## + 企业               1   0.00999 0.330 -111
## + 基础研究           1   0.00003 0.340 -110
## + 实验发展           1   0.00002 0.340 -110
##
## Step:  AIC=-113.4
## 总效率值 ~ 应用研究 + 科研机构 + 加工制造业 +
##     邮电通信业 + 计算机应用服务业 + 科学研究业 +
##     综合技术服务业 + 市场指数 + 人均GDP
##
##                     Df Sum of Sq   RSS  AIC
## - 应用研究           1   0.00919 0.361 -115
## - 市场指数           1   0.00969 0.361 -115
## - 科研机构           1   0.01328 0.365 -114
## - 科学研究业         1   0.01377 0.365 -114
## - 加工制造业         1   0.01390 0.365 -114
## - 综合技术服务业     1   0.01393 0.365 -114
## - 邮电通信业         1   0.01412 0.366 -114
## - 计算机应用服务业   1   0.01421 0.366 -114
## - 人均GDP            1   0.02188 0.373 -114
## <none>                           0.351 -113
## + 高等院校           1   0.01181 0.340 -112
## + 企业               1   0.01177 0.340 -112
## + 实验发展           1   0.00056 0.351 -112
## + 基础研究           1   0.00055 0.351 -112
##
## Step:  AIC=-114.6
## 总效率值 ~ 科研机构 + 加工制造业 + 邮电通信业 +
##     计算机应用服务业 + 科学研究业 + 综合技术服务业 +
##     市场指数 + 人均GDP
##
##                     Df Sum of Sq   RSS  AIC
## - 市场指数           1   0.00561 0.366 -116
## - 科学研究业         1   0.01616 0.377 -115
## - 加工制造业         1   0.01631 0.377 -115
## - 综合技术服务业     1   0.01635 0.377 -115
## - 人均GDP            1   0.01651 0.377 -115
## - 邮电通信业         1   0.01653 0.377 -115
## - 计算机应用服务业   1   0.01662 0.377 -115
## - 科研机构           1   0.01868 0.379 -115
## <none>                           0.361 -115
## + 应用研究           1   0.00919 0.351 -113
## + 实验发展           1   0.00652 0.354 -113
## + 高等院校           1   0.00470 0.356 -113
## + 企业               1   0.00468 0.356 -113
## + 基础研究           1   0.00005 0.361 -113
##
## Step:  AIC=-116.2
## 总效率值 ~ 科研机构 + 加工制造业 + 邮电通信业 +
##     计算机应用服务业 + 科学研究业 + 综合技术服务业 +
##     人均GDP
##
##                     Df Sum of Sq   RSS  AIC
```

```
## - 人均GDP            1   0.01090 0.377 -117
## - 科学研究业         1   0.01482 0.381 -117
## - 加工制造业         1   0.01496 0.381 -117
## - 综合技术服务业     1   0.01503 0.381 -117
## - 邮电通信业         1   0.01520 0.381 -117
## - 计算机应用服务业   1   0.01525 0.381 -117
## - 科研机构           1   0.01584 0.382 -117
## <none>                           0.366 -116
## + 市场指数           1   0.00561 0.361 -115
## + 应用研究           1   0.00511 0.361 -115
## + 实验发展           1   0.00505 0.361 -115
## + 高等院校           1   0.00195 0.364 -114
## + 企业               1   0.00194 0.364 -114
## + 基础研究           1   0.00003 0.366 -114
##
## Step:  AIC=-117.3
## 总效率值 ~ 科研机构 + 加工制造业 + 邮电通信业 +
##     计算机应用服务业 + 科学研究业 + 综合技术服务业
##
##                      Df Sum of Sq   RSS  AIC
## - 科研机构           1   0.01073 0.388 -118
## - 科学研究业         1   0.01244 0.390 -118
## - 加工制造业         1   0.01255 0.390 -118
## - 综合技术服务业     1   0.01264 0.390 -118
## - 邮电通信业         1   0.01279 0.390 -118
## - 计算机应用服务业   1   0.01287 0.390 -118
## <none>                           0.377 -117
## + 人均GDP            1   0.01090 0.366 -116
## + 实验发展           1   0.00741 0.370 -116
## + 应用研究           1   0.00369 0.373 -116
## + 高等院校           1   0.00145 0.376 -115
## + 企业               1   0.00143 0.376 -115
## + 基础研究           1   0.00128 0.376 -115
## + 市场指数           1   0.00000 0.377 -115
##
## Step:  AIC=-118.5
## 总效率值 ~ 加工制造业 + 邮电通信业 + 计算机应用服务业 +
##     科学研究业 + 综合技术服务业
##
##                      Df Sum of Sq   RSS  AIC
## - 科学研究业         1   0.01586 0.404 -119
## - 加工制造业         1   0.01589 0.404 -119
## - 综合技术服务业     1   0.01604 0.404 -119
## - 计算机应用服务业   1   0.01622 0.404 -119
## - 邮电通信业         1   0.01625 0.404 -119
## <none>                           0.388 -118
## + 实验发展           1   0.01576 0.372 -118
## + 科研机构           1   0.01073 0.377 -117
## + 应用研究           1   0.00738 0.380 -117
## + 人均GDP            1   0.00579 0.382 -117
## + 基础研究           1   0.00477 0.383 -117
## + 高等院校           1   0.00349 0.384 -117
## + 企业               1   0.00049 0.387 -116
## + 市场指数           1   0.00000 0.388 -116
```

```
##
## Step:  AIC=-119.2
## 总效率值 ~ 加工制造业 + 邮电通信业 + 计算机应用服务业 +
##     综合技术服务业
##
##                    Df Sum of Sq   RSS  AIC
## <none>                           0.404 -119
## + 实验发展          1    0.0237 0.380 -119
## + 科学研究业        1    0.0159 0.388 -118
## + 科研机构          1    0.0141 0.390 -118
## - 综合技术服务业    1    0.0450 0.449 -118
## + 应用研究          1    0.0109 0.393 -118
## + 基础研究          1    0.0081 0.396 -118
## + 高等院校          1    0.0068 0.397 -118
## + 人均GDP           1    0.0035 0.400 -118
## + 企业              1    0.0022 0.402 -117
## + 市场指数          1    0.0000 0.404 -117
## - 邮电通信业        1    0.0974 0.501 -115
## - 计算机应用服务业  1    0.0999 0.504 -115
## - 加工制造业        1    0.1521 0.556 -112
```

进一步检测结果，得出Adjusted R-squared=0.481 ，且 综合技术服务业 不为显着变量

```
##
## Call:
## lm(formula = 总效率值 ~ 加工制造业 + 邮电通信业 +
##     计算机应用服务业 + 综合技术服务业, data = Q1L)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.17715 -0.08547 -0.00708  0.06170  0.26329
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.290      0.116    2.51   0.0191 *
## 加工制造业          0.446      0.145    3.07   0.0051 **
## 邮电通信业          5.181      2.110    2.46   0.0214 *
## 计算机应用服务业    4.849      1.950    2.49   0.0199 *
## 综合技术服务业      2.470      1.480    1.67   0.1075
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.127 on 25 degrees of freedom
## Multiple R-squared:  0.552,  Adjusted R-squared:  0.481
## F-statistic: 7.71 on 4 and 25 DF,  p-value: 0.000344
```

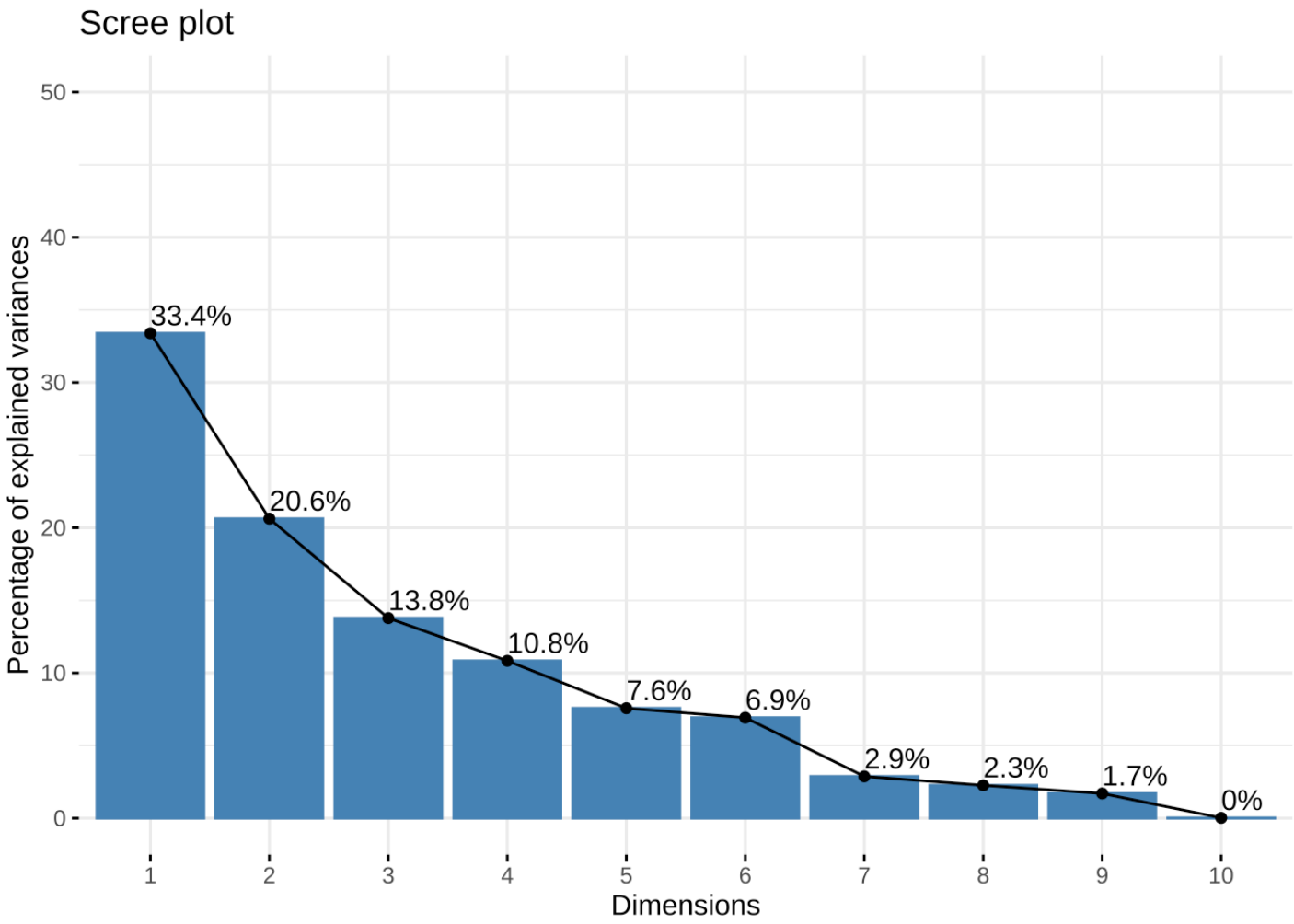
综上所述，both stepwise regression从只有截距项的模型出发所获得的模型，其解释程度最高。

1.2 对X1到X13变量做主成分分析

下表可以看出13个变量降为5至6个维度（5至6个主成分）后才可解释90%的变异。

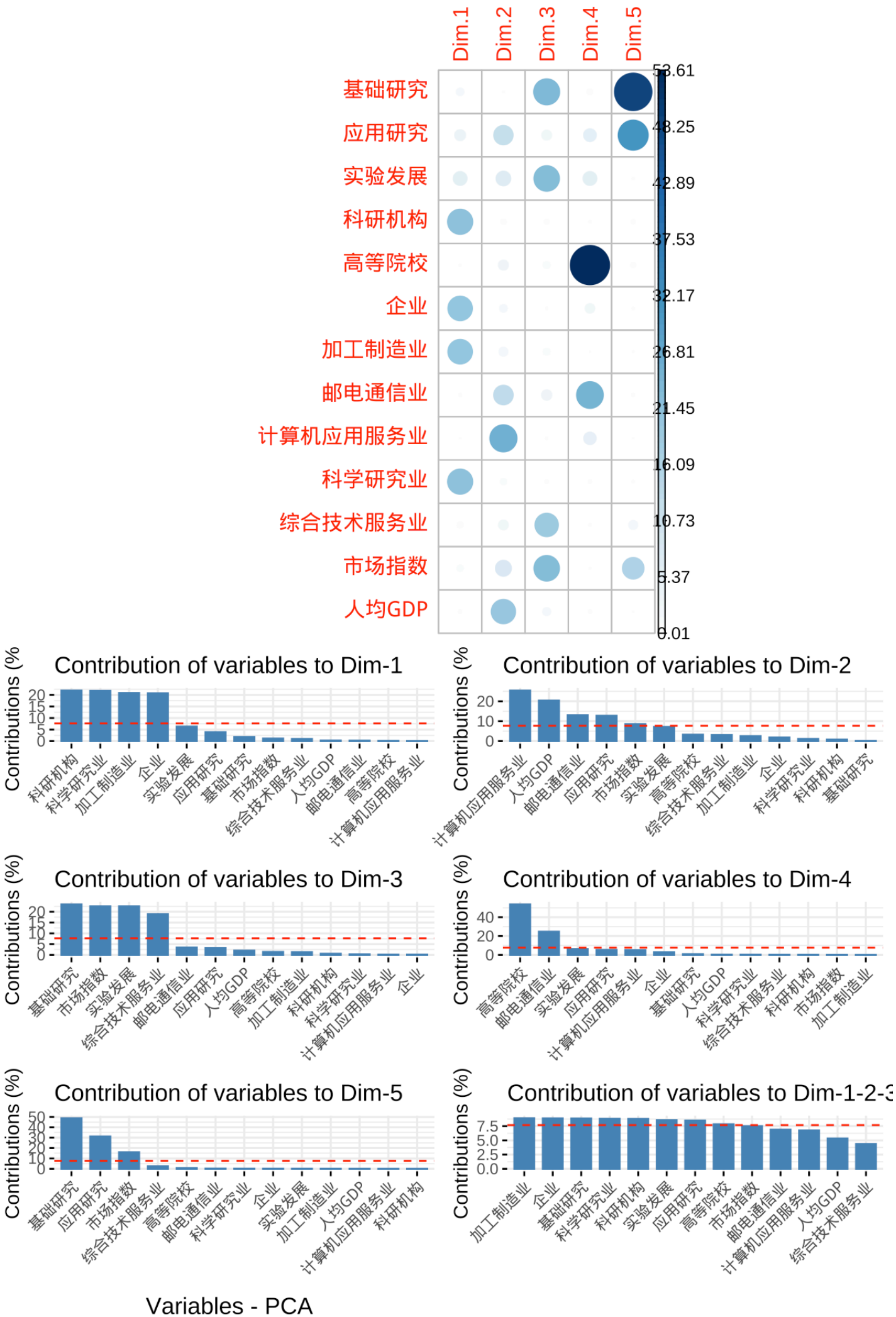
```
##           eigenvalue variance.percent cumulative.variance.percent
## Dim.1  4.34072249041    33.3901730031                       33.39
## Dim.2  2.68124785690    20.6249835146                       54.02
## Dim.3  1.79103353279    13.7771810215                       67.79
## Dim.4  1.40873091874    10.8363916826                       78.63
## Dim.5  0.98456199343     7.5735537956                       86.20
## Dim.6  0.90021309292     6.9247160994                       93.13
## Dim.7  0.37431479006     2.8793445389                       96.01
## Dim.8  0.29457323877     2.2659479906                       98.27
## Dim.9  0.22175673477     1.7058210367                       99.98
## Dim.10 0.00282410731     0.0217239024                      100.00
## Dim.11 0.00002118547     0.0001629652                      100.00
## Dim.12 0.00000004009     0.0000003084                      100.00
## Dim.13 0.00000001833     0.0000001410                      100.00
```

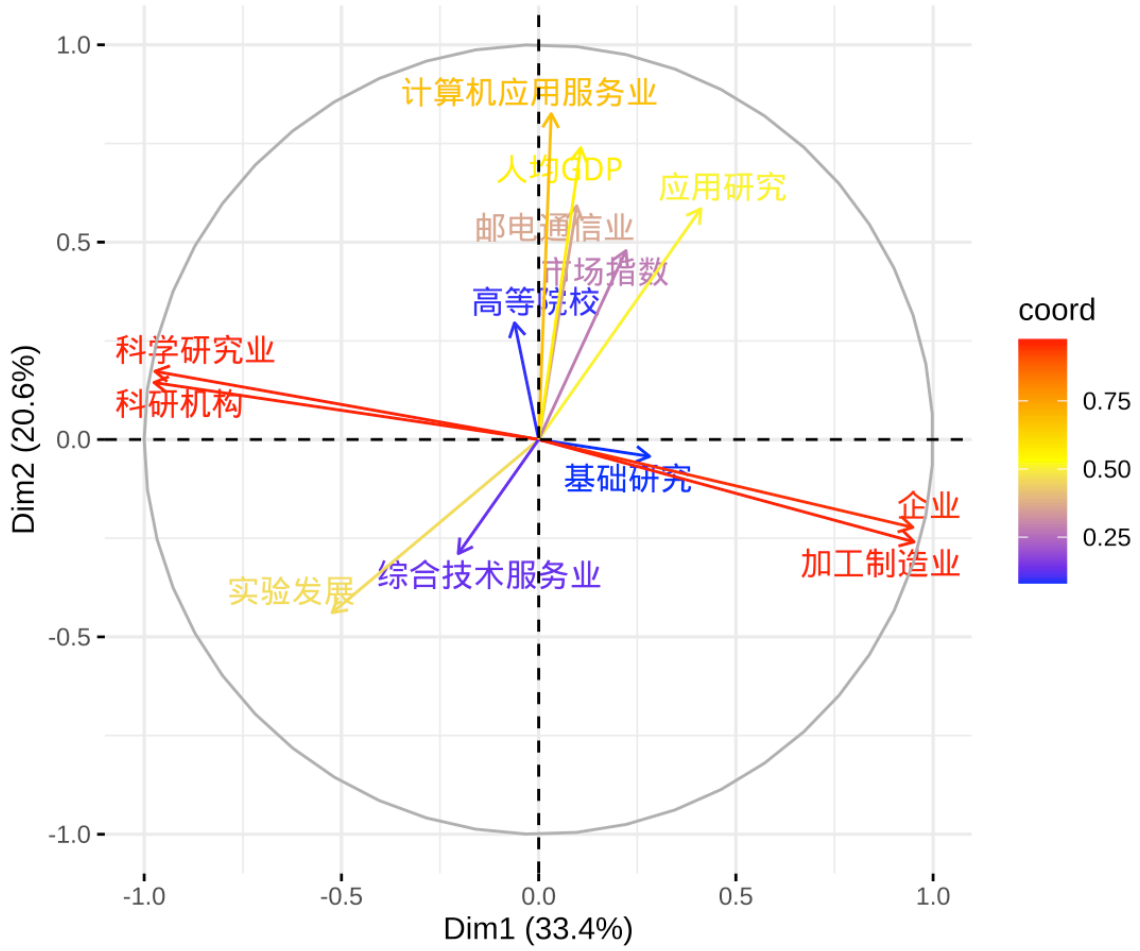
下图为 PCA 的陡坡图， 横轴为维度(主成分)，纵轴为解释变异百分比。



下为不同主成分中，原始变量的贡献。

```
##                       Dim.1    Dim.2    Dim.3    Dim.4    Dim.5
## 基础研究           0.28122 -0.04273  0.64771 -0.11182  0.69316
## 应用研究           0.41113  0.58448  0.23823 -0.27527 -0.55456
## 实验发展          -0.52342 -0.43862 -0.63536  0.29926 -0.03314
## 科研机构          -0.97512  0.14437  0.10235  0.04722  0.01180
## 高等院校          -0.06119  0.29550 -0.15967 -0.86904  0.08120
## 企业               0.94867 -0.22256 -0.05196  0.20387 -0.03460
## 加工制造业         0.95150 -0.25937 -0.15423 -0.01574 -0.02663
## 邮电通信业         0.09656  0.59196  0.24897  0.58998  0.04806
## 计算机应用服务业   0.03169  0.82547  0.05281  0.26855  0.01659
## 科学研究业        -0.97286  0.17278  0.07477 -0.06202  0.03802
## 综合技术服务业    -0.20366 -0.28888  0.58132  0.04902 -0.15792
## 市场指数           0.22150  0.47842 -0.63541 -0.03171  0.39679
## 人均GDP            0.10672  0.73922 -0.19125  0.06460  0.01996
```


Dim.1
Dim.2
Dim.3
Dim.4
Dim.5
基础研究
应用研究
实验发展
科研机构
高等院校
企业
加工制造业
邮电通信业
计算机应用服务业
科学研究业
综合技术服务业
市场指数
人均GDP
53.61
48.25
42.89
37.53
32.17
26.81
21.45
16.09
10.73
5.37
0.01
Contribution of variables to Dim-1
Contributions (%
科研机构
科学研究业
加工制造业
企业
实验发展
应用研究
基础研究
市场指数
综合技术服务业
人均GDP
邮电通信业
高等院校
计算机应用服务业
Contribution of variables to Dim-2
计算机应用服务业
人均GDP
邮电通信业
应用研究
市场指数
实验发展
高等院校
综合技术服务业
加工制造业
企业
科学研究业
科研机构
基础研究
Contribution of variables to Dim-3
基础研究
市场指数
实验发展
综合技术服务业
邮电通信业
应用研究
人均GDP
高等院校
加工制造业
科研机构
科学研究业
计算机应用服务业
企业
Contribution of variables to Dim-4
高等院校
邮电通信业
实验发展
应用研究
计算机应用服务业
企业
基础研究
人均GDP
科学研究业
综合技术服务业
科研机构
市场指数
加工制造业
Contribution of variables to Dim-5
基础研究
应用研究
市场指数
综合技术服务业
高等院校
邮电通信业
科学研究业
企业
实验发展
加工制造业
人均GDP
计算机应用服务业
科研机构
Contribution of variables to Dim-1-2-3
加工制造业
企业
基础研究
科学研究业
科研机构
实验发展
应用研究
高等院校
市场指数
邮电通信业
计算机应用服务业
人均GDP
综合技术服务业
Variables - PCA



2. 根据附件中55个国家和地区男子和女子一些跑步项目2005年的纪录数据。

2.1 分别用样本协方差矩阵和样本相关矩阵对数据进行主成分分析，给出主成分得分并解释，并对比男子和女子的分析结果

样本协方差矩阵

以下为男子的分析结果，一个主成分就能解释变异程度达98.24%。

PC	Var.
1	0.9824
2	0.0134
3	0.0027
4	0.0013
5	0.0001
6	0.0001
7	0.0000

8	0.0000

以下为女子的分析结果，一个主成分就能解释变异程度达98.41%。

PC	Var.
1	0.9841
2	0.0144
3	0.0010
4	0.0004
5	0.0001
6	0.0000
7	0.0000

样本相关矩阵

以下为男子的分析结果，需要两个主成分解释变异程度才能达至91.39%，与样本协方差矩阵分析中一个主成分便能解释变异至98.24%相差很大，原因在于样本相关矩阵会消除变量测量单位不同的问题。

PC	Var.
1	0.8288
2	0.0851
3	0.0289
4	0.0278
5	0.0119
6	0.0091
7	0.0071
8	0.0013

以下为女子的分析结果，需要两个主成分解释变异程度才能达至91.37%，与样本协方差矩阵分析中一个主成分便能解释变异至98.41%相差很大，原因在于样本相关矩阵会消除变量测量单位不同的问题。

PC	Var.
1	0.8233
2	0.0904

3	0.0410
4	0.0177
5	0.0130
6	0.0089
7	0.0057

综上所述，若要解释变异程度大于90%，且主成分数量固定，以此比较男子和女子的分析结果。样本协方差矩阵中，女子的解释变异程度略微高于男子；样本相关矩阵中，则是男子的解释变异程度略微高于女子。

2.2 若统一用米/秒单位表示各项纪录的平均速度，用速度协方差矩阵做主成分分析，与前面结果进行对比，你认为哪一个分析得更好， 为什么？

以下使用速度协方差矩阵 下为男子的分析结果，一个主成分解释变异程度达94.67%。

PC	Var.
1	0.9467
2	0.0463
3	0.0044
4	0.0025
5	0.0000
6	0.0000
7	0.0000
8	0.0000

下为女子的分析结果，一个主成分解释变异程度达97.34%。

PC	Var.
1	0.9734
2	0.0198
3	0.0049
4	0.0019
5	0.0000

6	0.0000
7	0.0000

与上题结果相比，男子和女子的一个主成分解释变异程度均下降，然而一样都是女子的一个主成分之解释变异程度高于男主，在此题中，男女之间的解释变异程度差异增大。我认为此题的分析结果更好，因为此题有考虑到去统一不同变量间的单位问题，因此较为合理。

2.3 分别用时间和速度速度以及样本协方差矩阵和样本相关矩阵，对男、女径赛纪录做因子分析，计算因子得分，并解释和比较所得结果

首先对男子竞赛的样本协方差矩阵因子分析，并设定以Varimax去旋转矩阵，且以PCA方法计算因子分析（此处就不附加MLE方法的因子分析结果了）。 下面为Loading Matrix，可以以此建构 $X = AF + \epsilon$ ，且能看出累积贡献比率（可解释的变异程度）。 由Loading Matrix可以看出因子1为中长米因子、因子2为短米因子、因子3为中米因子。

```
##
## Loadings:
##              RC1   RC2   RC3
## X100米.秒    0.356 0.860 0.248
## X200米.秒    0.304 0.845 0.380
## X400米.秒    0.538 0.769
## X800米.分    0.604 0.435 0.615
## X1500米.分   0.679 0.453 0.517
## X5000米.分   0.848 0.420 0.285
## X10000米.分 0.876 0.388 0.251
## 马拉松.分    0.890 0.332 0.228
##
##                  RC1   RC2   RC3
## SS loadings    3.612 2.877 1.054
## Proportion Var 0.451 0.360 0.132
## Cumulative Var 0.451 0.811 0.943
```

下为计算出的因子得分，利于后续计算。

```
##                   RC1     RC2     RC3
## X100米.秒    -0.24158  0.5961 -0.1744
## X200米.秒    -0.43855  0.4909  0.3936
## X400米.秒     0.18395  0.5697 -0.9641
## X800米.分    -0.22177 -0.2731  1.2865
## X1500米.分   -0.03896 -0.2007  0.8148
## X5000米.分    0.41432 -0.1114 -0.2243
## X10000米.分   0.49433 -0.1298 -0.3558
## 马拉松.分     0.55815 -0.1779 -0.4135
```

下为将因子得分带入资料中所得出的因子矩阵，可用来做进一步的分析。

```
##              F1       F2        F3           F.rank
##  [1,] -0.49180  0.14244  0.272719  0.142658     26
```

```
##  [2,] -0.37860 -1.18256  0.228557  0.600529     51
##  [3,] -0.34039 -0.07590  0.007625  0.190703     29
##  [4,] -0.85969 -0.18120 -0.131532  0.498746     46
##  [5,]  1.90901 -1.04877 -0.360564 -0.462154      8
##  [6,]  0.20556 -1.11376 -0.875980  0.449497     42
##  [7,] -0.01688 -1.35038  0.168633  0.499988     47
##  [8,]  0.15823 -0.50067 -0.232798  0.148048     27
##  [9,] -0.38779 -0.30119  0.494342  0.231247     36
## [10,] -0.84439  0.29370  1.551654  0.074514     22
## [11,]  3.65031  2.44144  0.125685 -2.695438      1
## [12,] -0.66763  0.02773  3.185637 -0.137205     16
## [13,] -0.40840  0.37706 -0.459597  0.115710     24
## [14,] -0.21258  0.90677 -2.348147  0.084189     23
## [15,]  1.40691 -1.27469  0.977641 -0.323096     12
## [16,] -0.46149  0.13294 -0.309290  0.213253     32
## [17,] -0.51073 -0.55750 -0.629256  0.545174     50
## [18,] -0.61363 -0.65981  0.026737  0.541622     49
## [19,] -0.09952 -1.17866 -0.935294  0.628481     52
## [20,]  0.42926 -0.77452 -0.688196  0.186717     28
## [21,] -0.03355  1.26494  0.025954 -0.470489      7
## [22,]  0.17185 -0.70780 -0.228549  0.220013     33
## [23,] -0.32113  0.39140 -0.007264  0.005181     20
## [24,]  0.53557 -0.29340  1.359382 -0.334420     11
## [25,] -0.84273  0.66568 -0.350889  0.198028     30
## [26,] -0.14565  0.25724  0.674295 -0.122930     18
## [27,] -0.01046 -0.62248 -1.819306  0.497306     45
## [28,] -0.44223 -1.15186  1.055317  0.503516     48
## [29,] -1.50532  0.48091 -0.682454  0.631872     53
## [30,] -0.84383  0.95679  0.143598  0.018204     21
## [31,] -1.68737  2.25320  1.897857 -0.318838     13
## [32,]  0.05216  1.48975 -1.583599 -0.372002      9
## [33,]  1.14253  0.02976 -0.191899 -0.530929      6
## [34,]  1.36838 -1.86491  1.371893 -0.134527     17
## [35,] -1.16923 -0.49850  1.824234  0.494148     44
## [36,]  0.02376  2.39945 -0.688486 -0.831003      5
## [37,] -0.37545  0.11716 -1.004272  0.275415     37
## [38,] -0.36019  0.19450 -0.932268  0.228508     34
## [39,] -0.03240 -0.03991 -1.420642  0.229593     35
## [40,]  1.58994  0.09109  0.636036 -0.884214      4
## [41,] -0.58934  0.99733  1.700273 -0.336885     10
## [42,]  0.11503 -1.05393 -0.482061  0.414813     40
## [43,] -0.26173 -0.61910 -0.659994  0.453908     43
## [44,] -0.75680  0.56412  0.031052  0.142243     25
## [45,] -0.19167 -0.42482 -0.863954  0.374780     39
## [46,]  3.14964  1.46364  0.342800 -2.113094      2
## [47,]  1.49921  0.32202  0.448309 -0.902700      3
## [48,] -0.95397  0.26851 -0.454199  0.417318     41
## [49,] -0.39332 -0.08523  0.087242  0.208435     31
## [50,] -0.34262  0.03117 -1.092874  0.304943     38
## [51,]  0.06760  0.55814  0.343861 -0.293540     15
## [52,]  0.76432 -0.44793  0.874378 -0.316933     14
## [53,] -0.81469  1.26920 -0.127676 -0.077022     19
## [54,]  0.12795 -2.37862 -0.294670  0.888114     54
```

再来将男子的样本相关矩阵和速度矩阵做因子分析，分析结果均和样本协方差矩阵因子分析相同，我认为原因可能是因为因子分析的过程中，都是以相关矩阵求解出初始公因子，且F、ϵ都符合一些特定条件，才会导致这个结果。女子亦是。

以下是女子因子分析結果，一樣以Varimax去旋转矩阵，且以PCA方法计算因子分析（此处就不附加MLE方法的因子分析结果了）。 下面为Loading Matrix，可以以此建构 $X = AF + \epsilon$ ，且能看出累积贡献比率（可解释的变异程度）。 由Loading Matrix可以看出因子1为短米因子、因子2为長米因子、因子3为中米因子。

```
##
## Loadings:
##             RC2   RC1   RC3
## X100米.秒  0.847 0.412 0.228
## X200米.秒  0.859 0.406 0.247
## X400米.秒  0.861 0.252 0.362
## X800米.分  0.536 0.582 0.548
## X1500米.分 0.415 0.836 0.319
## X3000米.分 0.358 0.828 0.387
## 马拉松.分  0.335 0.435 0.824
##
##                  RC2   RC1   RC3
## SS loadings    2.897 2.310 1.475
## Proportion Var 0.414 0.330 0.211
## Cumulative Var 0.414 0.744 0.955
```

下为计算出的因子得分，利于后续计算。

```
##                 RC2      RC1     RC3
## X100米.秒    0.4901 -0.01874 -0.3435
## X200米.秒    0.4962 -0.06147 -0.2918
## X400米.秒    0.5286 -0.48158  0.1928
## X800米.分   -0.0441  0.03298  0.3836
## X1500米.分 -0.1919  0.84073 -0.4641
## X3000米.分 -0.2731  0.77033 -0.2588
## 马拉松.分   -0.2848 -0.47427  1.3577
```

下为将因子得分带入资料中所得出的因子矩阵，可用来做进一步的分析。

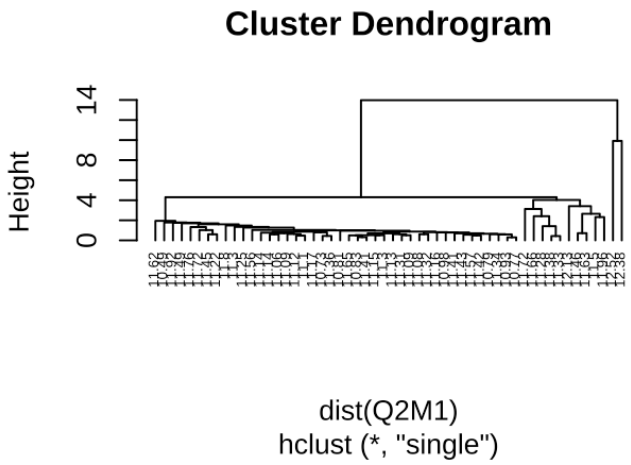
```
##              F1       F2       F3       F.rank
##  [1,]  0.23268  0.31955 -0.38879 -0.12539     20
##  [2,] -0.99933  0.07411 -0.38347  0.49233     44
##  [3,] -0.54371 -0.42590  0.21304  0.33580     38
##  [4,] -0.35759 -0.10193 -0.50937  0.30278     37
##  [5,] -0.34038  0.19115  1.52927 -0.25638     17
##  [6,] -0.63437  0.38771 -0.48979  0.24924     36
##  [7,] -0.74008 -0.46957  0.12827  0.45475     42
##  [8,]  1.04186 -0.05821 -0.69048 -0.27898     14
##  [9,] -0.89273 -1.04137 -0.04153  0.75603     51
## [10,] -0.88304  1.15234 -0.21251  0.03157     23
## [11,]  2.65175  0.15014  3.14007 -1.89521      2
## [12,]  0.29100  1.25262 -0.08634 -0.53992      7
## [13,] -1.37623  0.38477 -0.67412  0.61259     49
## [14,]  0.63562 -0.59555 -0.21009 -0.02333     22
## [15,]  0.20341  1.21457  0.11225 -0.53268      8
```

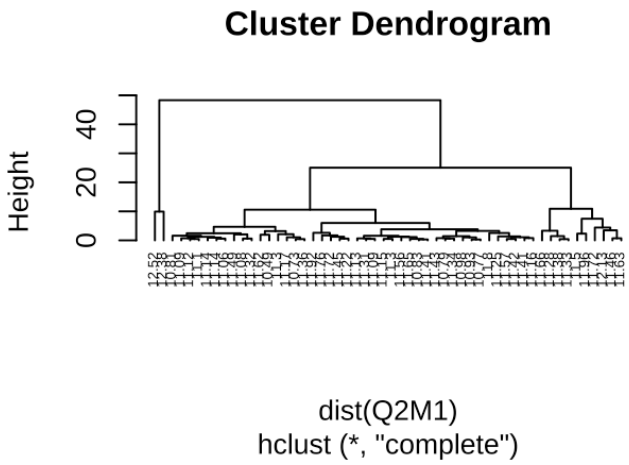
```
## [16,] -0.76057 -0.04459  0.02479  0.33965     39
## [17,] -1.76203  0.06517  0.20078  0.69697     50
## [18,] -1.72585  0.06217 -0.32656  0.79884     53
## [19,] -0.64698 -0.23335 -0.96136  0.57351     47
## [20,] -1.07086 -0.06042  0.56168  0.36100     41
## [21,]  1.39790  0.21280  0.68970 -0.83192      5
## [22,]  0.32400 -0.74250 -0.20225  0.16080     29
## [23,]  0.96536  0.29959 -0.20481 -0.47676     10
## [24,] -0.34538 -0.15199  0.39598  0.11475     27
## [25,]  0.42878 -0.76902 -0.54367  0.19998     30
## [26,] -0.04896  0.32935  0.17596 -0.13146     19
## [27,]  0.15300 -2.07056  0.47405  0.54442     46
## [28,]  0.48386  0.02077 -1.16012  0.03939     24
## [29,]  1.04045 -0.82889 -1.24721  0.11094     26
## [30,]  0.95005 -0.01344 -0.59932 -0.27479     15
## [31,]  2.62923 -0.61019 -1.69114 -0.55530      6
## [32,]  1.67567 -0.06876 -0.77917 -0.53050      9
## [33,] -0.11006  0.28542  1.15249 -0.30555     13
## [34,]  0.96034 -0.33853  0.64242 -0.44128     11
## [35,] -0.73334  0.70957 -0.75548  0.23964     33
## [36,]  0.77881 -0.31097  0.02754 -0.23625     18
## [37,] -0.16552 -0.42318 -0.60174  0.35093     40
## [38,]  0.19162 -0.27792 -0.59088  0.14352     28
## [39,]  0.77283 -0.75349 -0.62559  0.06356     25
## [40,]  0.28073 -0.23693  4.69039 -1.07614      3
## [41,] -0.06357  0.53165  1.00824 -0.37891     12
## [42,] -1.09960 -0.24364 -0.13520  0.59075     48
## [43,]  0.32865 -0.87932 -0.38041  0.24542     35
## [44,] -0.23856 -0.77361 -0.55427  0.49320     45
## [45,] -1.27174 -0.61211 -0.09912  0.78472     52
## [46,]  0.42480  5.44000 -0.26775 -2.00479      1
## [47,]  1.77107  1.11495 -1.21477 -0.88465      4
## [48,] -0.76707 -0.39089 -0.07790  0.48481     43
## [49,] -0.24291 -0.35425  0.06573  0.21319     31
## [50,]  0.17766 -0.71030 -0.32435  0.24009     34
## [51,] -0.76553  0.85818  0.59705 -0.09659     21
## [52,] -0.44878  1.15559  0.26654 -0.26366     16
## [53,]  0.28641 -1.40575  0.66755  0.21410     32
## [54,] -2.04273 -0.21506  0.26573  0.90114     54
```

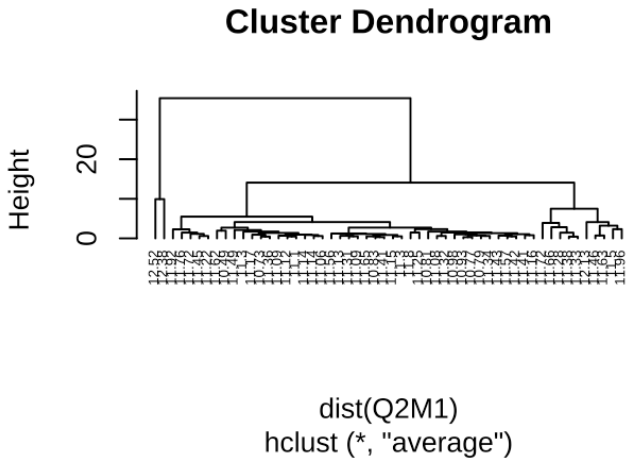
2.4 用聚类分析的方法研究这组数据

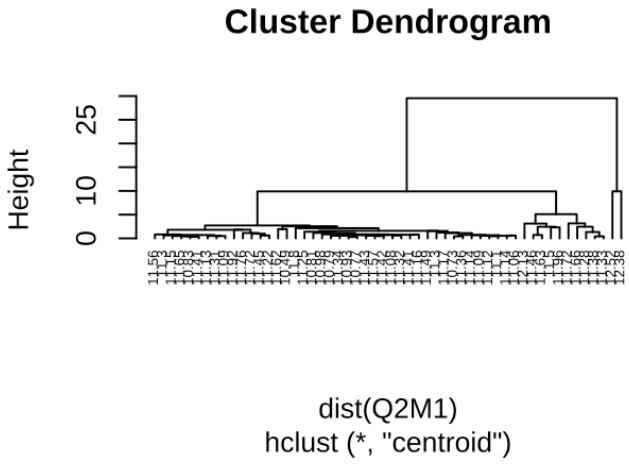
Hierarchical clustering

下为男子的阶层式分群，可见不同方法的分群结果不同，还是得依照资料结构而决定分群方法。

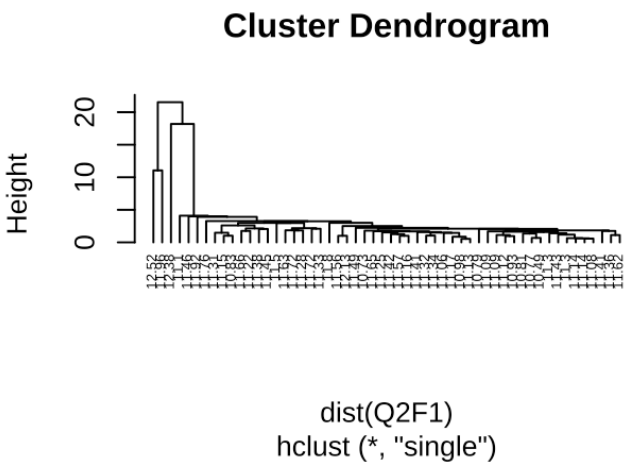


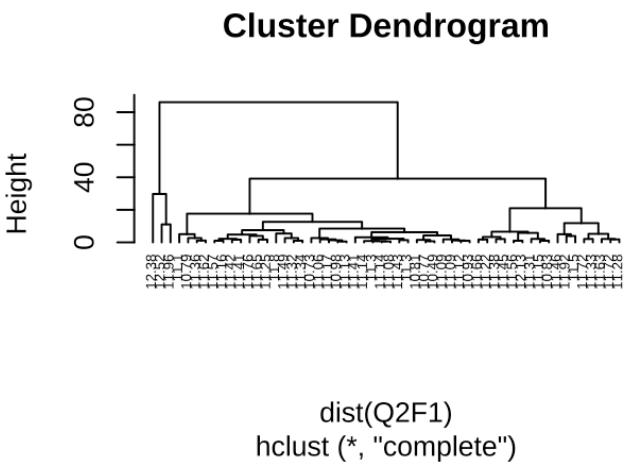


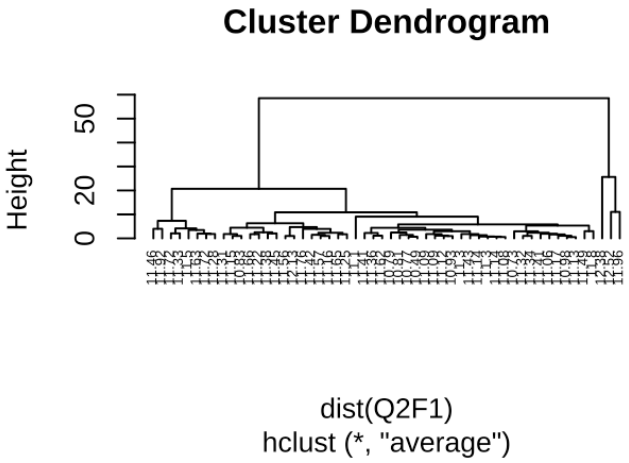


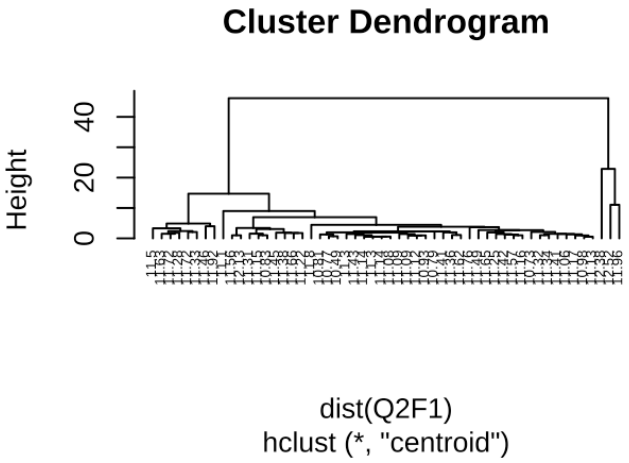


下为女子的阶层式分群



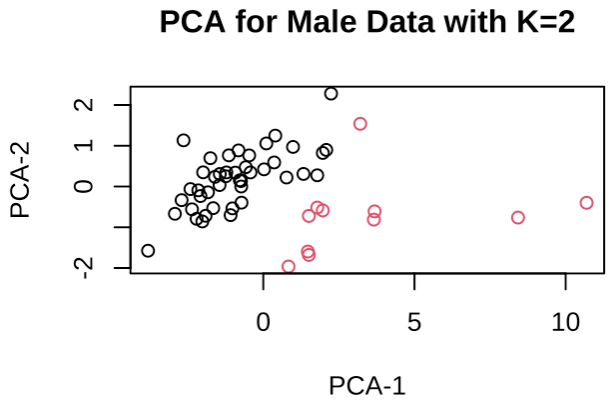


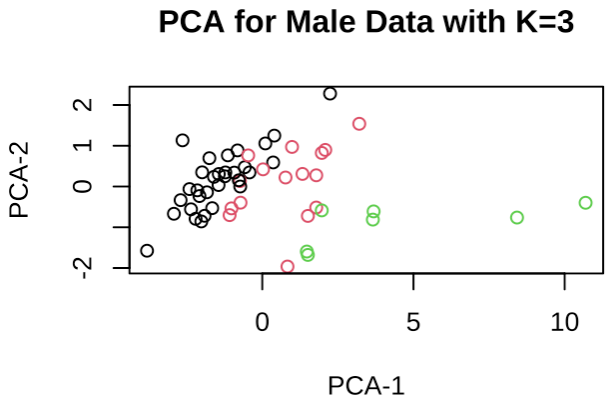


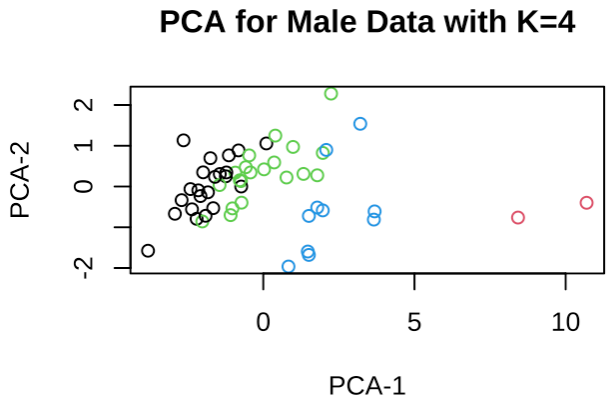


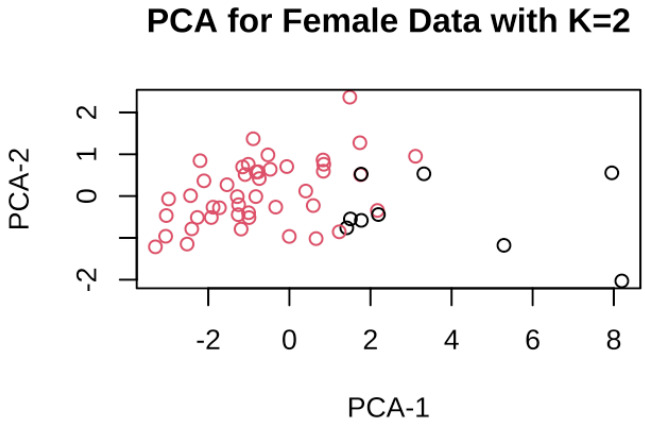
K-means

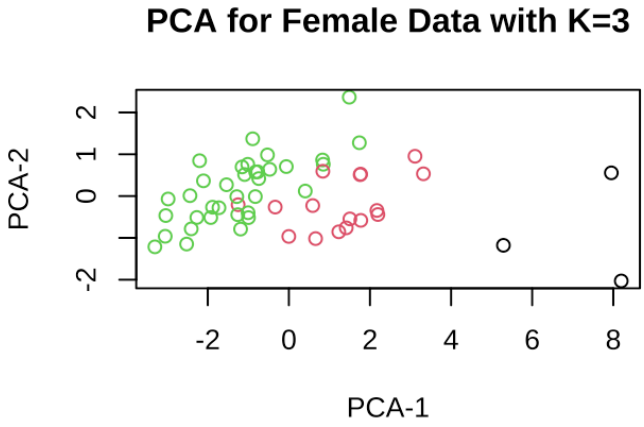
1. 使用PCA降维后再做K-means 以下分别为男子跟女子的分群结果，均在k=2时，分群界线最明显，较无交叠。

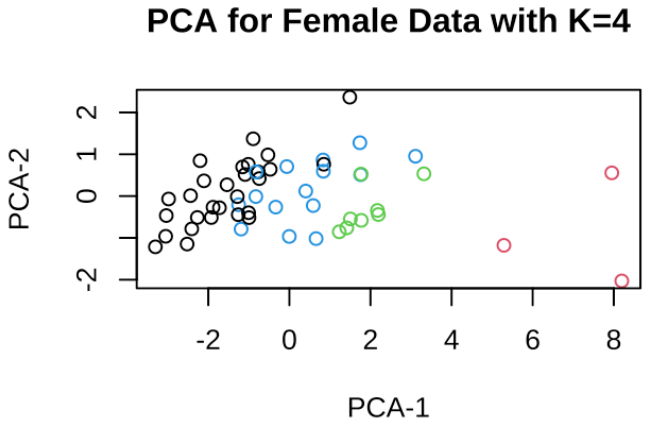












2. 使用MDS降维后再做K-means 以下分别为男子跟女子的分群结果，在k>2的分群效果都比PCA来得

好，尤其是女子，直到k=4都分群得很清楚。