Analysis on Dataset “test”

by Yimi Zhao

1 Introduction of Dataset

The “test” is a small size dataset which contains 20 observations across 8 variables. The data format is:

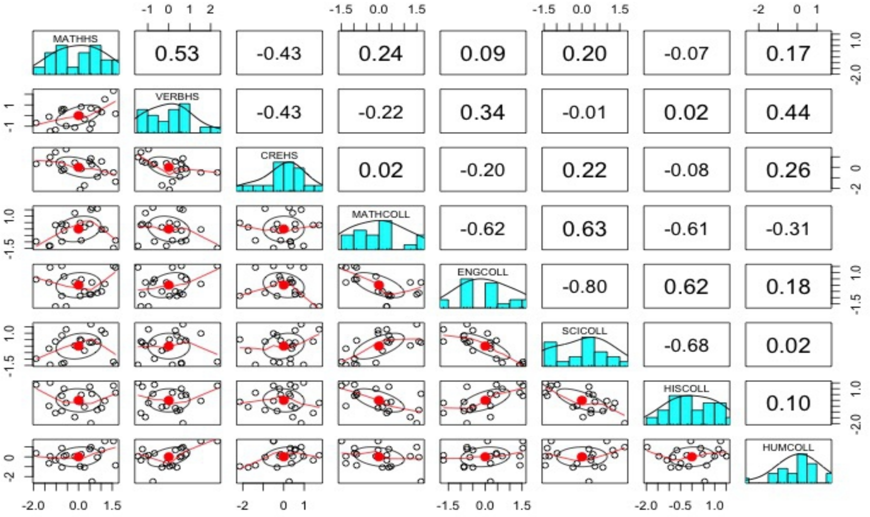
test mathhs, verbhs, crehs, mathcoll, engcoll, scicoll, hiscoll, humcoll

[“hs” = high school SAT %-tile, “coll” = college GPA, “math” = math, “verb” = verbal, “cre” = creative writing, “eng” = English, “sci” = science, “his” = history, “hum” = other humanities]

```
> head(test)
  MATHHS VERBHS CREHS MATHCOLL ENGCOLL SCICOLL HISCOLL HUMCOLL
1     27     62    46      2.5     2.0     2.9     1.2     1.4
2     21     38    76      2.4     1.3     2.3     1.0     2.1
3     27     65    57      0.9     4.0     0.1     2.7     3.3
4     45     68    49      1.0     3.1     2.4     2.8     2.7
5     15     55    58      0.5     2.9     1.6     1.7     1.8
6     54     66    66      3.8     2.1     4.0     0.0     3.4
```

2 Data Analysis

Our objective here is to describe the inter-relationships among the variables. From the output of the correlations among the variables, we see it shows some fairly high values which indicate a significant correlation of the variables. For example, the college GPA of English has a strong negative correlation(corr=-0.8) with the GPA of Science, which indicates the better the English GPA the poorer the Science GPA. Similarly, GPA of History has a fairly strong negative relation with the GPA of Science as well. On the contrary, GPA of Math has a fairly strong positive relation(corr=0.63) with the GPA of Science, which means the better the Math GPA the better the Science GPA.



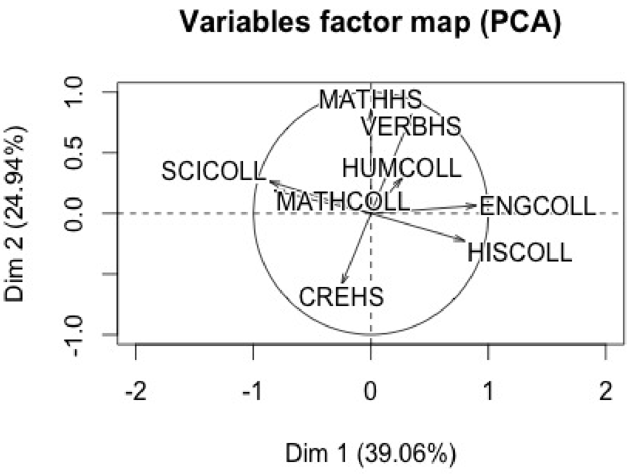
```
> round(cor(test),2)
          MATHHS  VERBHS  CREHS MATHCOLL ENGCOLL SCICOLL HISCOLL HUMCOLL
MATHHS     1.00    0.53   -0.43   0.24     0.09    0.20   -0.07    0.17
VERBHS     0.53    1.00   -0.43  -0.22     0.34   -0.01    0.02    0.44
CREHS     -0.43   -0.43    1.00   0.02    -0.20    0.22   -0.08    0.26
MATHCOLL   0.24   -0.22    0.02   1.00    -0.62    0.63   -0.61   -0.31
ENGCOLL    0.09    0.34   -0.20   0.62     1.00   -0.80    0.62    0.18
SCICOLL    0.20   -0.01    0.22   0.63    -0.80    1.00   -0.68    0.02
HISCOLL   -0.07    0.02   -0.08  -0.61     0.62   -0.68    1.00    0.10
HUMCOLL    0.17    0.44    0.26  -0.31     0.18    0.02    0.10    1.00
```

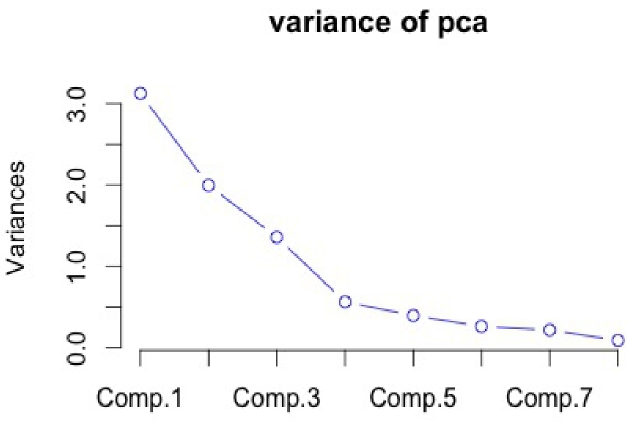
Because of the correlations of the variables, we consider to describe this dataset through a more parsimonious model. Next, we are going to try different ways to analyze the data.

2.1 PCA(Principal Component Analysis)

Since there are fairly strong correlations exist among the variable. Naturally, we would consider summay the dataset with less variables. We then apply Principal Component Analysis(PCA) to conduct the analysis. In order to decide how many principal components should be retained, we use "*screeplot()*" function to make a scree plot of variance of each component.

```
> test.pca<- princomp(test, cor=T)
> summary(test.pca)
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4     Comp.5     Comp.6     Comp.7     Comp.8
Standard deviation     1.7676276 1.4125470 1.1651744 0.74949161 0.62764516 0.51093058 0.46553846 0.29852994
Proportion of Variance 0.3905634 0.2494111 0.1697039 0.07021721 0.04924231 0.03263126 0.02709076 0.01114002
Cumulative Proportion  0.3905634 0.6399745 0.8096785 0.87989566 0.92913797 0.96176923 0.98885998 1.00000000
> screeplot(test.pca, type="lines")
> library(FactoMineR)
> test.pca2<-PCA(test)
> test.pca$loadings
Loadings:
          Comp.1  Comp.2  Comp.3  Comp.4  Comp.5  Comp.6  Comp.7  Comp.8
MATHHS             0.607           0.600          -0.166  -0.458   0.168
VERBHS     0.196   0.578  -0.185  -0.366          -0.336   0.492   0.319
CREHS     -0.141  -0.417  -0.565   0.333  -0.292  -0.452           0.288
MATHCOLL  -0.466   0.134   0.195   0.376  -0.343   0.363   0.578
ENGCOLL    0.508           0.124  -0.593  -0.151                  -0.585
SCICOLL   -0.489   0.187  -0.223           0.325  -0.365          -0.655
HISCOLL    0.452  -0.160           0.485   0.578           0.441
HUMCOLL    0.152   0.208  -0.740                   0.604          -0.120
```





Based on the outputs, we tend to **<u>retain the first four components which account for about 88% of the variance</u>**. There are missing values in the result of test.pca$loadings. However, we can see the first component is mainly related to college GPA while the second component is primarily about high school SAT. Therefore, we may call "Comp 1" as "college GPA" and "Comp2" as "HS SAT".

2.2 Factor Analysis

To further describe the inter-relationships among the variables, we try factor analysis on the dataset.

```
> library(psych)
> corMat<-cor(test)
> (test.fa<-fa(r=corMat, nfactors=3,rotate="varimax",fm="pa"))
Factor Analysis using method =  pa
Call: fa(r = corMat, nfactors = 3, rotate = "varimax", fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix
            PA1   PA2   PA3    h2   u2 com
MATHHS     0.15  0.68  0.11  0.50 0.498 1.1
VERBHS    -0.13  0.77  0.44  0.80 0.197 1.7
CREHS      0.17 -0.73  0.38  0.70 0.297 1.6
MATHCOLL   0.76  0.07 -0.28  0.66 0.345 1.3
ENGCOLL   -0.83  0.23  0.11  0.76 0.241 1.2
SCICOLL    0.94  0.03  0.17  0.92 0.077 1.1
HISCOLL   -0.74 -0.04  0.03  0.55 0.448 1.0
HUMCOLL   -0.11  0.09  0.84  0.72 0.280 1.1

                       PA1  PA2  PA3
SS loadings           2.79 1.66 1.18
Proportion Var        0.35 0.21 0.15
Cumulative Var        0.35 0.56 0.70
Proportion Explained  0.50 0.29 0.21
Cumulative Proportion 0.50 0.79 1.00

Mean item complexity =  1.3
Test of the hypothesis that 3 factors are sufficient.

The degrees of freedom for the null model are  28  and the objective function was  4.66
The degrees of freedom for the model are 7  and the objective function was  0.49

The root mean square of the residuals (RMSR) is  0.04
The df corrected root mean square of the residuals is  0.08

Fit based upon off diagonal values = 0.99
Measures of factor score adequacy
                                                PA1  PA2  PA3
Correlation of scores with factors             0.97 0.92 0.90
Multiple R square of scores with factors       0.95 0.85 0.82
Minimum correlation of possible factor scores  0.90 0.70 0.63
```

According to the output, the values that we consider large are in boldface, using about 0.5 as the cutoff. Based on this criterion the following statements may be made:

1) The first factor is mainly about college GPAs, and the second one is related to high school SAT score. Clearly, this result is similar with the conclusion we gained from previous principal component analysis.

2) Factor 1 shows a clear comparison between the two groups: Math&Science and English&History. Students with better Math&Science GPAs tend to be not so good at English&History, which is very true in reality.

3) Similarly, factor 2 indicates negative relations among SAT scores of Creative writing, Math and Verbal. Math and creative writing SAT scores are negatively related is plausible. However, the negative relation between creative writing and verbal does not seem quite reasonable.

4) Factor 3 is primarily a measure of other humanities. We also see GPA of Math is negatively related to Humanities, which indicates students who are good at humanity subjects probably are poor at Math.

The communality for a given variable can be interpreted as the proportion of variation in that variable explained by the three factors. Therefore, the results below suggest that the factor analysis does the best job of explaining variation in Science GPA(0.923), Verbal SAT(0.803), English GPA (0.759)and Humanities GPA(0.720). Moreover, this three-factor model explained roughly 70% of the overall variation of the dataset. And we could say this model works fairly well with the data.

```
#proportion of variation of each variable explained by the three factors
> round(test.fa$communality,3)
  MATHHS   VERBHS    CREHS   MATHCOLL  ENGCOLL  SCICOLL  HISCOLL  HUMCOLL
   0.502    0.803    0.703     0.655    0.759    0.923    0.552    0.720
# overall model fit(percentage of variation explained in our model)
>sum(test.fa$communality)/8 # 8 is the number of parameters in our model
[1] 0.7021619
```