(Bonus) Homework #5: Factor analysis

* **Due** Dec 16 by 1pm

* **Points** 100

This assignment is optional.  If completed, it is worth up to +10% addition to your course grade.

In the first part of this homework assignment, you’ll perform PCA on a US crime data set and explore the relationships between the variables in the data. In the second part, you’ll use EFA to analyze the responses from a survey on managerial skills.

See the attached zip file for all required files, including the R markdown template with complete instructions and a sample output file.

US Crime Data Analysis (35 points)

First, load the data provided in the uscrime.txt file into an R data frame and display it as below (5 points). The dataset contains the following variables (extracted from http://lib.stat.cmu.edu/DASL/Datafiles/USCrime.html). :

R: Crime rate: # of offenses reported to police per million population

Age: The number of males of age 14-24 per 1000 population

Ed: Mean # of years of schooling x 10 for persons of age 25 or older

Ex0: 1960 per capita expenditure on police by state and local government

Ex1: 1959 per capita expenditure on police by state and local government

LF: Labor force participation rate per 1000 civilian urban males age 14-24

M: The number of males per 1000 females

N: State population size in hundred thousands

U1: Unemployment rate of urban males per 1000 of age 14-24

U2: Unemployment rate of urban males per 1000 of age 35-39

W: Median value of transferable goods and assets or family income in tens of $

X: The number of families per 1000 earning below 1/2 the median income

Once loaded, the data should look like:

##        R Age  Ed Ex0 Ex1  LF    M   N  U1 U2   W   X

## 1   79.1 151  91  58  56 510  950  33 108 41 394 261

## 2  163.5 143 113 103  95 583 1012  13  96 36 557 194

## 3   57.8 142  89  45  44 533  969  18  94 33 318 250

## 4  196.9 136 121 149 141 577  994 157 102 39 673 167

## 5  123.4 141 121 109 101 591  985  18  91 20 578 174

## 6   68.2 121 110 118 115 547  964  25  84 29 689 126

## 7   96.3 127 111  82  79 519  982   4  97 38 620 168

## 8  155.5 131 109 115 109 542  969  50  79 35 472 206

## 9   85.6 157  90  65  62 553  955  39  81 28 421 239

## 10  70.5 140 118  71  68 632 1029   7 100 24 526 174

## 11 167.4 124 105 121 116 580  966 101  77 35 657 170

## 12  84.9 134 108  75  71 595  972  47  83 31 580 172

## 13  51.1 128 113  67  60 624  972  28  77 25 507 206

## 14  66.4 135 117  62  61 595  986  22  77 27 529 190

## 15  79.8 152  87  57  53 530  986  30  92 43 405 264

## 16  94.6 142  88  81  77 497  956  33 116 47 427 247

## 17  53.9 143 110  66  63 537  977  10 114 35 487 166

## 18  92.9 135 104 123 115 537  978  31  89 34 631 165

## 19  75.0 130 116 128 128 536  934  51  78 34 627 135

## 20 122.5 125 108 113 105 567  985  78 130 58 626 166

## 21  74.2 126 108  74  67 602  984  34 102 33 557 195

## 22  43.9 157  89  47  44 512  962  22  97 34 288 276

## 23 121.6 132  96  87  83 564  953  43  83 32 513 227

## 24  96.8 131 116  78  73 574 1038   7 142 42 540 176

## 25  52.3 130 116  63  57 641  984  14  70 21 486 196

## 26 199.3 131 121 160 143 631 1071   3 102 41 674 152

## 27  34.2 135 109  69  71 540  965   6  80 22 564 139

## 28 121.6 152 112  82  76 571 1018  10 103 28 537 215

## 29 104.3 119 107 166 157 521  938 168  92 36 637 154

## 30  69.6 166  89  58  54 521  973  46  72 26 396 237

## 31  37.3 140  93  55  54 535 1045   6 135 40 453 200

## 32  75.4 125 109  90  81 586  964  97 105 43 617 163

## 33 107.2 147 104  63  64 560  972  23  76 24 462 233

## 34  92.3 126 118  97  97 542  990  18 102 35 589 166

## 35  65.3 123 102  97  87 526  948 113 124 50 572 158

## 36 127.2 150 100 109  98 531  964   9  87 38 559 153

## 37  83.1 177  87  58  56 638  974  24  76 28 382 254

## 38  56.6 133 104  51  47 599 1024   7  99 27 425 225

## 39  82.6 149  88  61  54 515  953  36  86 35 395 251

## 40 115.1 145 104  82  74 560  981  96  88 31 488 228

## 41  88.0 148 122  72  66 601  998   9  84 20 590 144

## 42  54.2 141 109  56  54 523  968   4 107 37 489 170

## 43  82.3 162  99  75  70 522  996  40  73 27 496 224

## 44 103.0 136 121  95  96 574 1012  29 111 37 622 162

## 45  45.5 139  88  46  41 480  968  19 135 53 457 249

## 46  50.8 126 104 106  97 599  989  40  78 25 593 171

## 47  84.9 130 121  90  91 623 1049   3 113 40 588 160

Now perform PCA with scaling, and plot the resulting Scree plot (10 points).

## Standard deviations:

##  [1] 2.2292902 1.5270237 1.3527914 1.1013981 0.7770701 0.5649796 0.4949622

##  [8] 0.4700725 0.3582280 0.2773785 0.2364407 0.0719401

##

## Rotation:

##             PC1         PC2          PC3         PC4          PC5

## R    0.26685963  0.08673898 -0.116821578  0.64689323 -0.088574854

## Age -0.31162886 -0.07141851 -0.160223186  0.43938861 -0.313768592

## Ed   0.34242058 -0.32422100  0.109941139 -0.10235683 -0.017166226

## Ex0  0.40393369  0.15691222 -0.152718666  0.15791580 -0.146995179

## Ex1  0.40497071  0.15166825 -0.156179671  0.12915958 -0.181184557

## LF   0.15298518 -0.48983303 -0.002055818  0.16214051  0.608179770

## M    0.09777702 -0.35558275  0.453041959  0.39008302  0.009873998

## N    0.18259610  0.39194925 -0.288349349  0.02838305  0.611000190

## U1   0.02570843  0.24164638  0.657892543  0.02500586  0.078895617

## U2   0.05753086  0.49216742  0.416886848  0.10981777  0.080150339

## W    0.42156549 -0.03062254  0.020902473 -0.14357819 -0.117548506

## X   -0.37318024  0.11433546 -0.073677465  0.35186709  0.263070432

##             PC6           PC7         PC8         PC9        PC10

## R   -0.25261335  0.4778158919  0.00833614 -0.19481698  0.27537438

## Age  0.65334808  0.0219074509 -0.33727878  0.11583879 -0.08904480

## Ed   0.17512474  0.5972371852  0.24987837  0.26950172 -0.43505518

## Ex0 -0.05751002 -0.3338959261  0.03570042  0.33192761 -0.02003496

## Ex1 -0.04551056 -0.3182562134  0.06867715  0.35903369 -0.06884604

## LF  -0.08233415 -0.1028063231 -0.50280458  0.24104658  0.10673679

## M    0.11599050 -0.4222481274  0.36470849 -0.34917502 -0.22289082

## N    0.49971878  0.0009399851  0.23802975 -0.18263490 -0.07451698

## U1   0.26132012  0.0955520294  0.09228313  0.34766157  0.53709218

## U2  -0.15793770  0.0559828359 -0.47154888 -0.07791235 -0.55286517

## W    0.11662316 -0.0024477084 -0.25434122 -0.53842433  0.24194757

## X   -0.31062910  0.0247489509  0.28230551  0.08181438 -0.05033295

##             PC11          PC12

## R   -0.272936635  2.878398e-02

## Age  0.134592638 -6.236659e-03

## Ed   0.202850003 -4.379262e-02

## Ex0  0.069423774 -7.147606e-01

## Ex1  0.110710462  6.965033e-01

## LF   0.001246253  2.871379e-02

## M   -0.100514013  5.014691e-03

## N   -0.101854676  1.008454e-02

## U1   0.076936911  9.124091e-03

## U2  -0.003785031  1.214733e-03

## W    0.599783075 -3.997247e-06

## X    0.680450937 -1.338735e-02

Answer the following questions: \*What does the scree plot tell you? (2 points)

\*What can you say about the data from the scree plot? (3 points)

Draw a biplot of the PCA results (5 points)

Answer questions: \* (2 points) From the biplot, what can you say about Ex0 (1960 per capita police expenditure) and Ex1 (1959 per capita police expenditure)?

(2 points) From the biplot, what can you say about R (crime rate) and Ex0 (1960 per capita police expenditure)?

(2 points) From the biplot, what can you say about W (median assets/income) and Age (proportion of males 14-24)?

(4 points) R and U1 point to almost exactly the same direction. Does this mean they are highly correlated? Why? Plot a scatterplot of these two variables and compute the Pearson’s correlation coeffecient to justify your answer.

Managerial Skills Survey Analysis (30 points)

The included CSV file “manage.csv” contains survey data of these 11 questions:

I show confidence in my staff

I let my staff know they are doing well

I give feedback to staff on how well they are working

I would personally compliment staff if they did outstanding work

I believe in setting goals and achieving them

I achieve the things I want to get done in a day

I never try to put off until tomorrow what I can finish today

I plan the use of my time well

I remain clear headed when too many demands are made upon me

I rarely overlook important factors when plans are made

I handle complex problems efficiently

We can view each question as an apparent skill. The question we seek to answer through EFA is: Are there underlying fundamental “latent” skills that “produce” or “explain” these 11 observed skills?

First, load the data into a R data frame. You can’t do EFA on this data frame directly, because the answers to these survey questions are not numerical. You’ll need to convert the data frame into a numeric data frame according to the following scale (6 points):

Survey AnswerCorresponding Numeric Value

VERY LITTLE EXTENT1

LITTLE EXTENT2

SOME EXTENT3

MODERATE EXTENT4

CONSIDERABLE EXTENT5

GREAT EXTENT6

TO A VERY GREAT EXTENT7

Now perform exploratory factor analysis with “promax” rotation. Experiment with a minimum of 1,2 and 3 factors. Show the outputs of the factor analysis (12 points).

Can you find a set of factors that have “simple” structure (e.g. according to Thurstone’s criteria covered in class) while fitting the data well? Include a justification for your choice (6 points) Give your interpretation of each factor (6 points)?

What to submit:

A ZIP (not .rar or other format) file named 'si618hw5\_***youruniquename***.zip' containing:

* The R Markdown file you wrote named 'si618hw5report\_***youruniquename***.Rmd'
* The HTML report generated by your R Markdown file.