Comp 4442 Final Project

Luke Sonnanburg

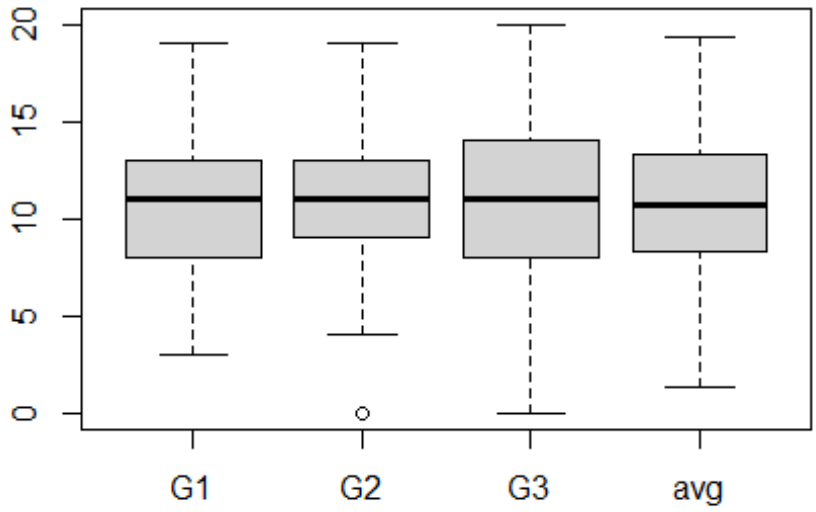
This notebook was used for data exploration, analysis, and generating visualizations, but was not originally intended to be shared. See accompanying PDF for full discussion of results.

Loading in data

```
# Importing Data
math<-read.csv("~/GitHub/Predicting-Student-Success/student-mat.csv", sep = ';') # Student data and scores in math
#port<-read.csv("~/student-por.csv", sep = ';') # Student data and scors in portuguese
```

Observing distribution of scores

```
math$avg <- (math$G1 + math$G2 + math$G3)/3
math.grades <- math[,c(31:34)]
boxplot(math.grades)
```



Getting data into the correct form

```
math$school <- as.numeric(ifelse(math$school =="GP", 0, 1)) # Gabriel Pereira = 0, Mousinho da Silveira = 1
```

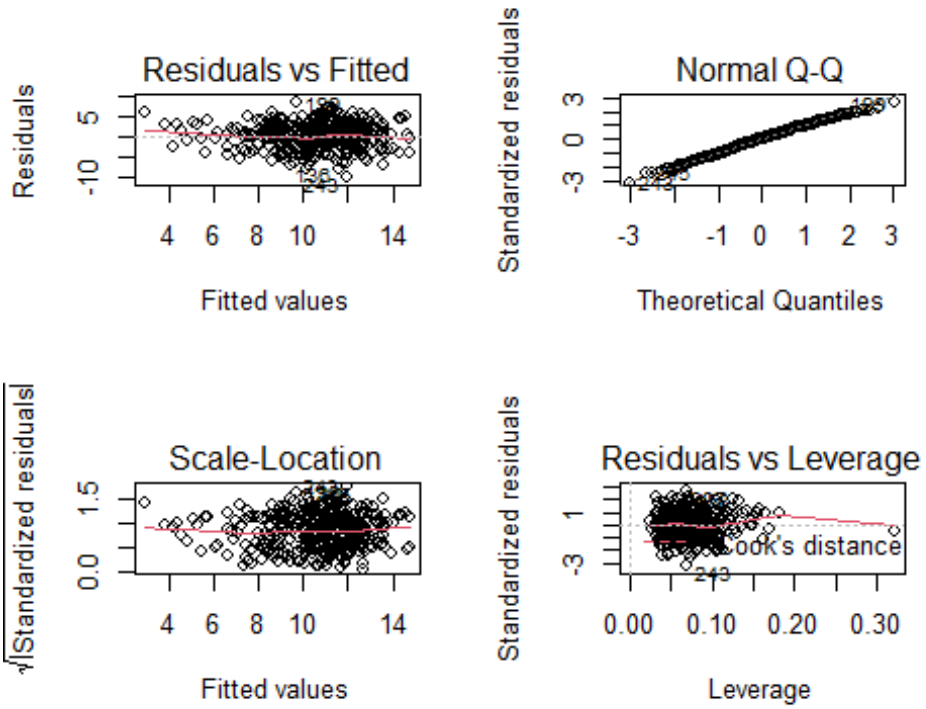
```
math$sex <- as.numeric(ifelse(math$sex == "F", 0, 1)) # Female = 0, Male = 1
math$address <- as.numeric(ifelse(math$address == "U", 0, 1)) # Urban = 0,
Rural = 1
math$famsize <- as.numeric(ifelse(math$famsize == "LE3", 0, 1)) # Less than 3
kids = 0, more = 1
math$Pstatus <- as.numeric(ifelse(math$Pstatus == "T", 0, 1)) # Parents
together = 0, parents apart = 1
math$Medu <- as.numeric(ifelse(math$Medu == 4, 1,0))
math$Fedu <- as.numeric(ifelse(math$Fedu == 4, 1,0))
math$guardian <- as.numeric(ifelse(math$guardian == "other", 0, 1)) #
recoding legal guardian as factors
math$reason <- ifelse(math$reason %in% c('reputation','course'), 1,0 ) # 1 ->
chose school for academic reasons
math$schoolsup <- as.numeric(ifelse(math$schoolsup =="no", 0, 1))
math$famsup <- as.numeric(ifelse(math$famsup =="no", 0, 1))
math$paid <- as.numeric(ifelse(math$paid =="no", 0, 1))
math$activities <- as.numeric(ifelse(math$activities =="no", 0, 1))
math$nursery <- as.numeric(ifelse(math$nursery =="no", 0, 1))
math$higher <- as.numeric(ifelse(math$higher =="no", 0, 1))
math$internet <- as.numeric(ifelse(math$internet =="no", 0, 1))
math$romantic <- as.numeric(ifelse(math$romantic =="no", 0, 1))
math <- math[-c(9:10,31:33)]
mathX <- math[-c(29)]
mathY <- math$avg
```

The variable inflation factor is a measure of excessive multicollinearity that may be a cause for concern. The values here are small enough not to worry about.

Checking assumptions for linear regression:

```
naive.model <- lm(avg~., data = math)
vif(naive.model)

##     school        sex        age    address    famsize    Pstatus
Medu
##   1.445697   1.409600   1.757240   1.340793   1.096765   1.103001
1.496942
##       Fedu     reason   guardian traveltime  studytime   failures
schoolsup
##   1.405227   1.104899   1.418038   1.228788   1.299861   1.357658
1.139820
##     famsup       paid activities    nursery     higher   internet
romantic
##   1.233314   1.289238   1.116316   1.120957   1.242830   1.166947
1.116847
##     famrel   freetime      goout       Dalc       Walc     health
absences
##   1.105096   1.258055   1.410180   1.909314   2.234650   1.101272
1.193737
```

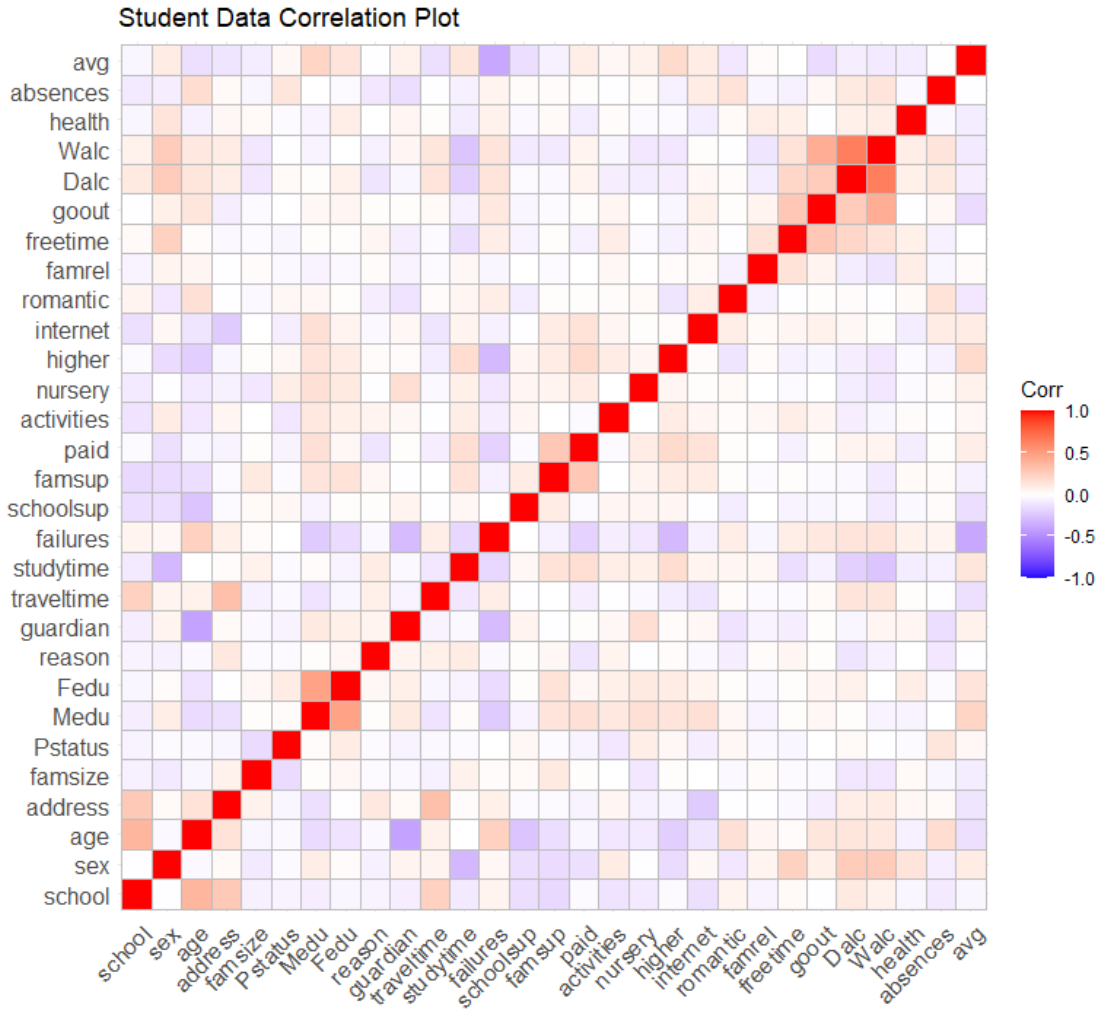
```
par(mfrow = c(2, 2))
plot(naive.model)
```



Assumptions for linear regression are met within reasonable parameters, there doesn't seem to be any need for removing outliers or transforming data.

Correlation between

```
math.cor <- cor(math)

ggcorrplot(math.cor, title = 'Student Data Correlation Plot')
```



KMO factor adequacy is an indication of how suitable data is for factor analysis.

```
KMO(math.cor)

## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = math.cor)
## Overall MSA =  0.62
## MSA for each item =
##     school        sex        age    address    famsize    Pstatus
Medu
##       0.58       0.65       0.61       0.59       0.57       0.51
0.66
##       Fedu     reason   guardian traveltime  studytime   failures
schoolsup
##       0.57       0.50       0.54       0.65       0.67       0.75
0.58
##     famsup       paid activities    nursery     higher   internet
```

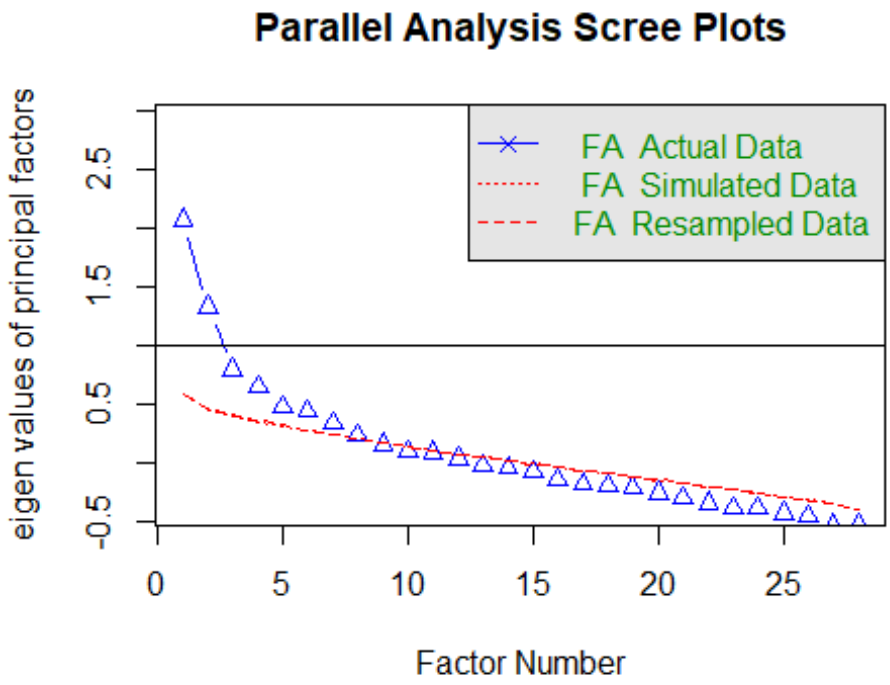
```
romantic
##       0.63       0.61       0.55       0.63       0.69       0.64
0.60
##     famrel   freetime      goout       Dalc       Walc     health
absences
##       0.47       0.59       0.58       0.67       0.63       0.51
0.50
##        avg
##       0.66
```

Opinions are divided on what a good minimum is for considering the use of factor analysis. By all accounts, greater than 0.60 indicates factor analysis should yield usable results.

The skree plot shows eigenvalues of principal factors. Greater eigenvalues indicate that a factor explains more of the output variable's variance. Generally eigenvalues >= 1 imply the factor explains more variance than a single variable would. In this case, the first two factors have eigenvalues >=1 and should obviously be included. A more subjective means of determining how many factors is where the scree plot levels off. In this case, the first place where the eigenvalues flatten (if only for a moment) is at the fourth factor. Since adding factors until the next "leveling off" would result in 8 factors where more are uninfluential than influential, I'll operate on the assumption of four factors being reasonable.

(https://www.theanalysisfactor.com/factor-analysis-1-introduction/)

```
parallel <- fa.parallel(mathX, fa='fa')
```



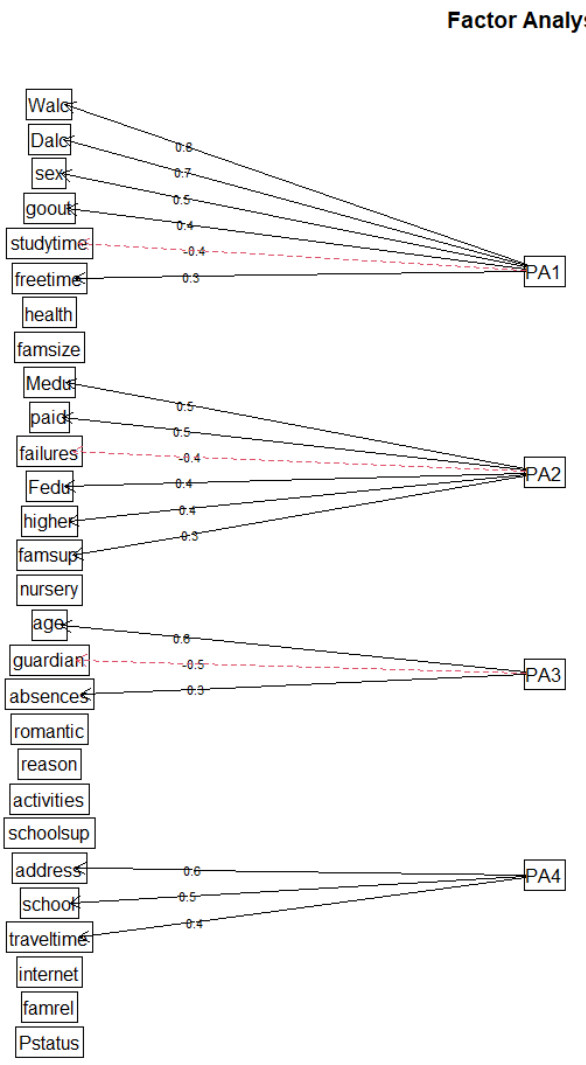
```
## Parallel analysis suggests that the number of factors =  7  and the number
of components =  NA

parallel

## Call: fa.parallel(x = mathX, fa = "fa")
## Parallel analysis suggests that the number of factors =  7  and the number
of components =  NA
##
##  Eigen Values of
##
##  eigen values of factors
##  [1]  2.07  1.33  0.79  0.65  0.47  0.44  0.33  0.23  0.15  0.09  0.08
0.03
## [13] -0.03 -0.05 -0.08 -0.15 -0.18 -0.20 -0.22 -0.26 -0.31 -0.35 -0.39 -
0.39
## [25] -0.44 -0.46 -0.54 -0.54
##
##  eigen values of simulated factors
##  [1]  0.59  0.45  0.40  0.35  0.31  0.27  0.24  0.21  0.17  0.14  0.11
0.07
## [13]  0.04  0.02 -0.01 -0.05 -0.07 -0.10 -0.12 -0.15 -0.18 -0.21 -0.23 -
0.26
## [25] -0.29 -0.32 -0.36 -0.40
##
##  eigen values of components
##  [1] 2.92 2.20 1.73 1.57 1.43 1.35 1.28 1.20 1.12 1.05 1.04 0.97 0.93 0.91
0.87
## [16] 0.78 0.75 0.73 0.72 0.67 0.63 0.58 0.54 0.52 0.49 0.40 0.35 0.28
##
##  eigen values of simulated components
## [1] NA
```

While the scree plot for portuguese test data suggests 5 factors may be more appropriate, the ideal scenario would be if both data sets produced similar factor loadings to paint a picture of what a "successful" student looks like. For portuguese I'll proceed with 4 to match math, but also experiment with 5.

```
math.varimaxfit.4 <- fa(r=mathX, nfactors = 4, rotate="varimax",fm="pa")

names(math.varimaxfit.4$loadings) <- c('a','b','c','d')
math.varimaxfit.4

## Factor Analysis using method =  pa
## Call: fa(r = mathX, nfactors = 4, rotate = "varimax", fm = "pa")
## Standardized loadings (pattern matrix) based upon correlation matrix
##              PA1   PA2   PA3   PA4     h2   u2 com
## school      0.03 -0.09  0.18  0.51 0.3038 0.70 1.3
## sex         0.49 -0.14 -0.20 -0.02 0.3025 0.70 1.5
## age         0.04 -0.24  0.63  0.29 0.5396 0.46 1.8
## address     0.00 -0.05 -0.06  0.61 0.3759 0.62 1.0
```

```
## famsize    -0.13  0.01 -0.02 -0.02 0.0169 0.98 1.1
## Pstatus     0.01  0.03  0.05 -0.06 0.0067 0.99 2.7
## Medu        0.07  0.49 -0.09 -0.15 0.2791 0.72 1.3
## Fedu        0.09  0.39 -0.12 -0.04 0.1756 0.82 1.3
## reason     -0.08 -0.01 -0.16  0.09 0.0393 0.96 2.2
## guardian    0.07  0.20 -0.48  0.00 0.2751 0.72 1.4
## traveltime  0.10 -0.08 -0.02  0.44 0.2124 0.79 1.2
## studytime  -0.40  0.22  0.08 -0.02 0.2196 0.78 1.7
## failures    0.18 -0.42  0.28  0.03 0.2826 0.72 2.2
## schoolsup  -0.09  0.04 -0.14 -0.10 0.0396 0.96 2.7
## famsup     -0.12  0.34  0.04 -0.09 0.1374 0.86 1.4
## paid       -0.07  0.47  0.17 -0.02 0.2576 0.74 1.3
## activities  0.02  0.09 -0.14 -0.05 0.0322 0.97 2.0
## nursery    -0.06  0.20 -0.08 -0.08 0.0588 0.94 1.9
## higher     -0.17  0.36 -0.11 -0.01 0.1703 0.83 1.7
## internet    0.07  0.21  0.09 -0.30 0.1472 0.85 2.2
## romantic   -0.03  0.02  0.29 -0.01 0.0846 0.92 1.0
## famrel     -0.02 -0.07 -0.03 -0.07 0.0127 0.99 2.5
## freetime    0.34 -0.07 -0.01 -0.08 0.1259 0.87 1.2
## goout       0.41  0.07  0.15 -0.06 0.2017 0.80 1.4
## Dalc        0.68  0.12  0.19  0.18 0.5477 0.45 1.4
## Walc        0.75  0.08  0.16  0.18 0.6313 0.37 1.2
## health      0.15 -0.07 -0.12  0.01 0.0417 0.96 2.4
## absences    0.08  0.03  0.30 -0.07 0.1037 0.90 1.3
##
##                        PA1  PA2  PA3  PA4
## SS loadings           1.90 1.36 1.21 1.15
## Proportion Var        0.07 0.05 0.04 0.04
## Cumulative Var        0.07 0.12 0.16 0.20
## Proportion Explained  0.34 0.24 0.22 0.20
## Cumulative Proportion 0.34 0.58 0.80 1.00
##
## Mean item complexity =  1.7
## Test of the hypothesis that 4 factors are sufficient.
##
## The degrees of freedom for the null model are  378  and the objective
function was  3.99 with Chi Square of  1532.92
## The degrees of freedom for the model are 272  and the objective function
was  1.49
##
## The root mean square of the residuals (RMSR) is  0.05
## The df corrected root mean square of the residuals is  0.06
##
## The harmonic number of observations is  395 with the empirical chi square
725.9  with prob <  5.3e-43
## The total number of observations was  395  with Likelihood Chi Square =
568.48  with prob <  4.5e-23
##
## Tucker Lewis Index of factoring reliability =  0.64
## RMSEA index =  0.052  and the 90 % confidence intervals are  0.047 0.059
```

```
## BIC =  -1057.78
## Fit based upon off diagonal values = 0.8
## Measures of factor score adequacy
##                                                  PA1  PA2  PA3  PA4
## Correlation of (regression) scores with factors   0.88 0.79 0.80 0.78
## Multiple R square of scores with factors          0.78 0.63 0.64 0.60
## Minimum correlation of possible factor scores     0.55 0.25 0.27 0.21

fa.diagram(math.varimaxfit.4)
```




```
math.varimaxfit.4$loadings

##
## Loadings:
##             PA1    PA2    PA3    PA4
## school                      0.176  0.514
## sex          0.491 -0.138 -0.205
```

```
## age               -0.239  0.628  0.295
## address                           0.608
## famsize    -0.127
## Pstatus
## Medu               0.494        -0.151
## Fedu               0.389 -0.119
## reason                   -0.157
## guardian           0.200 -0.481
## traveltime  0.104                 0.441
## studytime  -0.402  0.225
## failures    0.179 -0.415  0.279
## schoolsup                -0.143
## famsup     -0.117  0.337
## paid               0.473  0.170
## activities               -0.144
## nursery            0.203
## higher     -0.175  0.356 -0.113
## internet           0.215        -0.297
## romantic                  0.288
## famrel
## freetime    0.337
## goout       0.414         0.153
## Dalc        0.681  0.123  0.194  0.176
## Walc        0.755         0.158  0.175
## health      0.148        -0.124
## absences                  0.303
##
##                  PA1   PA2   PA3   PA4
## SS loadings    1.901 1.357 1.212 1.152
## Proportion Var 0.068 0.048 0.043 0.041
## Cumulative Var 0.068 0.116 0.160 0.201
```

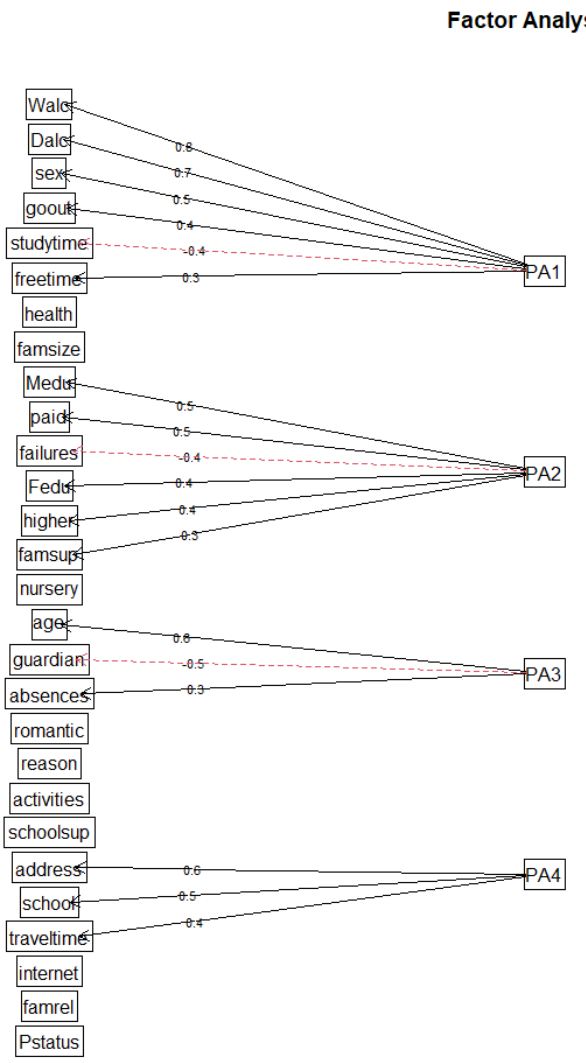
```
math.varimaxfit.5 <- fa(r=mathX, nfactors = 5, rotate="varimax",fm="pa")
math.varimaxfit.5
```

```
## Factor Analysis using method =  pa
## Call: fa(r = mathX, nfactors = 5, rotate = "varimax", fm = "pa")
## Standardized loadings (pattern matrix) based upon correlation matrix
##              PA1   PA2   PA3   PA4   PA5     h2   u2 com
## school      0.04  0.00  0.27  0.52 -0.04 0.3436 0.66 1.6
## sex         0.32  0.14 -0.12  0.02 -0.48 0.3720 0.63 2.1
## age         0.07 -0.16  0.77  0.25 -0.09 0.6974 0.30 1.4
## address     0.04 -0.05 -0.01  0.60  0.03 0.3684 0.63 1.0
## famsize    -0.11 -0.01 -0.02 -0.01  0.07 0.0173 0.98 1.9
## Pstatus     0.01  0.03  0.04 -0.06  0.01 0.0066 0.99 2.6
## Medu        0.05  0.79  0.06 -0.15  0.08 0.6590 0.34 1.1
## Fedu        0.07  0.55 -0.02 -0.02  0.04 0.3063 0.69 1.0
## reason     -0.10  0.04 -0.11  0.11 -0.02 0.0384 0.96 3.4
## guardian    0.03  0.19 -0.48  0.04  0.03 0.2648 0.74 1.3
## traveltime  0.12 -0.08  0.00  0.43 -0.03 0.2099 0.79 1.2
```

```
## studytime  -0.27  0.01  0.02 -0.05  0.40 0.2354 0.76 1.8
## failures    0.13 -0.32  0.27  0.02 -0.26 0.2562 0.74 3.3
## schoolsup  -0.06 -0.08 -0.22 -0.09  0.12 0.0842 0.92 2.4
## famsup     -0.01  0.13 -0.05 -0.13  0.34 0.1498 0.85 1.6
## paid        0.12  0.14  0.04 -0.10  0.49 0.2875 0.71 1.4
## activities -0.02  0.16 -0.10 -0.04 -0.04 0.0384 0.96 2.1
## nursery    -0.05  0.19 -0.07 -0.08  0.10 0.0589 0.94 2.4
## higher     -0.09  0.19 -0.16 -0.03  0.32 0.1681 0.83 2.4
## internet    0.10  0.13  0.03 -0.32  0.10 0.1413 0.86 1.8
## romantic    0.01  0.01  0.29 -0.04  0.05 0.0888 0.91 1.1
## famrel     -0.07  0.00  0.00 -0.06 -0.10 0.0188 0.98 2.6
## freetime    0.25  0.05  0.02 -0.08 -0.26 0.1366 0.86 2.3
## goout       0.42  0.01  0.09 -0.10 -0.06 0.1968 0.80 1.3
## Dalc        0.73 -0.02  0.07  0.10 -0.03 0.5434 0.46 1.1
## Walc        0.89 -0.11 -0.01  0.11 -0.02 0.8115 0.19 1.1
## health      0.09  0.02 -0.10  0.03 -0.16 0.0428 0.96 2.4
## absences    0.14 -0.07  0.23 -0.12  0.09 0.1017 0.90 2.8
##
##                        PA1  PA2  PA3  PA4  PA5
## SS loadings           1.87 1.29 1.26 1.13 1.10
## Proportion Var        0.07 0.05 0.04 0.04 0.04
## Cumulative Var        0.07 0.11 0.16 0.20 0.24
## Proportion Explained  0.28 0.19 0.19 0.17 0.16
## Cumulative Proportion 0.28 0.48 0.66 0.84 1.00
##
## Mean item complexity =  1.9
## Test of the hypothesis that 5 factors are sufficient.
##
## The degrees of freedom for the null model are  378  and the objective
function was  3.99 with Chi Square of  1532.92
## The degrees of freedom for the model are 248  and the objective function
was  1.12
##
## The root mean square of the residuals (RMSR) is  0.04
## The df corrected root mean square of the residuals is  0.05
##
## The harmonic number of observations is  395 with the empirical chi square
544.28  with prob <  2.9e-24
## The total number of observations was  395  with Likelihood Chi Square =
426.48  with prob <  1.3e-11
##
## Tucker Lewis Index of factoring reliability =  0.762
## RMSEA index =  0.043  and the 90 % confidence intervals are  0.036 0.05
## BIC =  -1056.28
## Fit based upon off diagonal values = 0.85
## Measures of factor score adequacy
##                                                  PA1  PA2  PA3  PA4  PA5
## Correlation of (regression) scores with factors   0.93 0.85 0.85 0.77 0.76
## Multiple R square of scores with factors          0.86 0.72 0.73 0.60 0.58
## Minimum correlation of possible factor scores     0.72 0.43 0.45 0.20 0.16
```

```
math.varimaxfit.4 <- fa(r=mathX, nfactors = 4, rotate="varimax",fm="pa")
math.varimaxfit.4

## Factor Analysis using method =  pa
## Call: fa(r = mathX, nfactors = 4, rotate = "varimax", fm = "pa")
## Standardized loadings (pattern matrix) based upon correlation matrix
##              PA1   PA2   PA3   PA4     h2   u2 com
## school      0.03 -0.09  0.18  0.51 0.3038 0.70 1.3
## sex         0.49 -0.14 -0.20 -0.02 0.3025 0.70 1.5
## age         0.04 -0.24  0.63  0.29 0.5396 0.46 1.8
## address     0.00 -0.05 -0.06  0.61 0.3759 0.62 1.0
## famsize    -0.13  0.01 -0.02 -0.02 0.0169 0.98 1.1
## Pstatus     0.01  0.03  0.05 -0.06 0.0067 0.99 2.7
## Medu        0.07  0.49 -0.09 -0.15 0.2791 0.72 1.3
## Fedu        0.09  0.39 -0.12 -0.04 0.1756 0.82 1.3
## reason     -0.08 -0.01 -0.16  0.09 0.0393 0.96 2.2
## guardian    0.07  0.20 -0.48  0.00 0.2751 0.72 1.4
## traveltime  0.10 -0.08 -0.02  0.44 0.2124 0.79 1.2
## studytime  -0.40  0.22  0.08 -0.02 0.2196 0.78 1.7
## failures    0.18 -0.42  0.28  0.03 0.2826 0.72 2.2
## schoolsup  -0.09  0.04 -0.14 -0.10 0.0396 0.96 2.7
## famsup     -0.12  0.34  0.04 -0.09 0.1374 0.86 1.4
## paid       -0.07  0.47  0.17 -0.02 0.2576 0.74 1.3
## activities  0.02  0.09 -0.14 -0.05 0.0322 0.97 2.0
## nursery    -0.06  0.20 -0.08 -0.08 0.0588 0.94 1.9
## higher     -0.17  0.36 -0.11 -0.01 0.1703 0.83 1.7
## internet    0.07  0.21  0.09 -0.30 0.1472 0.85 2.2
## romantic   -0.03  0.02  0.29 -0.01 0.0846 0.92 1.0
## famrel     -0.02 -0.07 -0.03 -0.07 0.0127 0.99 2.5
## freetime    0.34 -0.07 -0.01 -0.08 0.1259 0.87 1.2
## goout       0.41  0.07  0.15 -0.06 0.2017 0.80 1.4
## Dalc        0.68  0.12  0.19  0.18 0.5477 0.45 1.4
## Walc        0.75  0.08  0.16  0.18 0.6313 0.37 1.2
## health      0.15 -0.07 -0.12  0.01 0.0417 0.96 2.4
## absences    0.08  0.03  0.30 -0.07 0.1037 0.90 1.3
##
##                        PA1  PA2  PA3  PA4
## SS loadings           1.90 1.36 1.21 1.15
## Proportion Var        0.07 0.05 0.04 0.04
## Cumulative Var        0.07 0.12 0.16 0.20
## Proportion Explained  0.34 0.24 0.22 0.20
## Cumulative Proportion 0.34 0.58 0.80 1.00
##
## Mean item complexity =  1.7
## Test of the hypothesis that 4 factors are sufficient.
##
## The degrees of freedom for the null model are  378  and the objective
function was  3.99 with Chi Square of  1532.92
## The degrees of freedom for the model are 272  and the objective function
was  1.49
```

```
##
## The root mean square of the residuals (RMSR) is  0.05
## The df corrected root mean square of the residuals is  0.06
##
## The harmonic number of observations is  395 with the empirical chi square
725.9  with prob <  5.3e-43
## The total number of observations was  395  with Likelihood Chi Square =
568.48  with prob <  4.5e-23
##
## Tucker Lewis Index of factoring reliability =  0.64
## RMSEA index =  0.052  and the 90 % confidence intervals are  0.047 0.059
## BIC =  -1057.78
## Fit based upon off diagonal values = 0.8
## Measures of factor score adequacy
##                                                  PA1  PA2  PA3  PA4
## Correlation of (regression) scores with factors   0.88 0.79 0.80 0.78
## Multiple R square of scores with factors          0.78 0.63 0.64 0.60
## Minimum correlation of possible factor scores     0.55 0.25 0.27 0.21

fa.diagram(math.varimaxfit.4)
```




```
colnames(math.varimaxfit.4$loadings) <- c('Outgoing/Social', 'Academic
Drive/Background', 'Hardship', 'Rural Location')
math.varimaxfit.4$loadings
```

```
##
## Loadings:
##            Outgoing/Social Academic Drive/Background Hardship Rural
Location
## school                                                 0.176    0.514
## sex         0.491          -0.138                     -0.205
## age                        -0.239                      0.628    0.295
## address                                                        0.608
## famsize    -0.127
## Pstatus
## Medu                       0.494                              -0.151
## Fedu                       0.389                      -0.119
## reason                                                -0.157
```

```
## guardian                    0.200                    -0.481
## traveltime  0.104                                             0.441
## studytime  -0.402           0.225
## failures    0.179          -0.415                     0.279
## schoolsup                                             -0.143
## famsup     -0.117           0.337
## paid                        0.473                     0.170
## activities                                            -0.144
## nursery                     0.203
## higher     -0.175           0.356                    -0.113
## internet                    0.215                             -0.297
## romantic                                               0.288
## famrel
## freetime    0.337
## goout       0.414                                      0.153
## Dalc        0.681           0.123                      0.194    0.176
## Walc        0.755                                      0.158    0.175
## health      0.148                                     -0.124
## absences                                               0.303
##
##                Outgoing/Social Academic Drive/Background Hardship
## SS loadings              1.901                    1.357    1.212
## Proportion Var           0.068                    0.048    0.043
## Cumulative Var           0.068                    0.116    0.160
##                Rural Location
## SS loadings             1.152
## Proportion Var          0.041
## Cumulative Var          0.201

fa.diagram(math.varimaxfit.4)
```

Factor Analysis