Predicting Student Success from Socioeconomic Factors

COMP 4442 FINAL PROJECT

LUKE SONNANBURG

Research Question

Question

Can we apply factor analysis to data on student habits and backgrounds to predict student test scores?

Motivation

- Identify healthy/productive behaviors to promote
- Predict which students might need extra support
- Determine external factors that may be remedied with investment

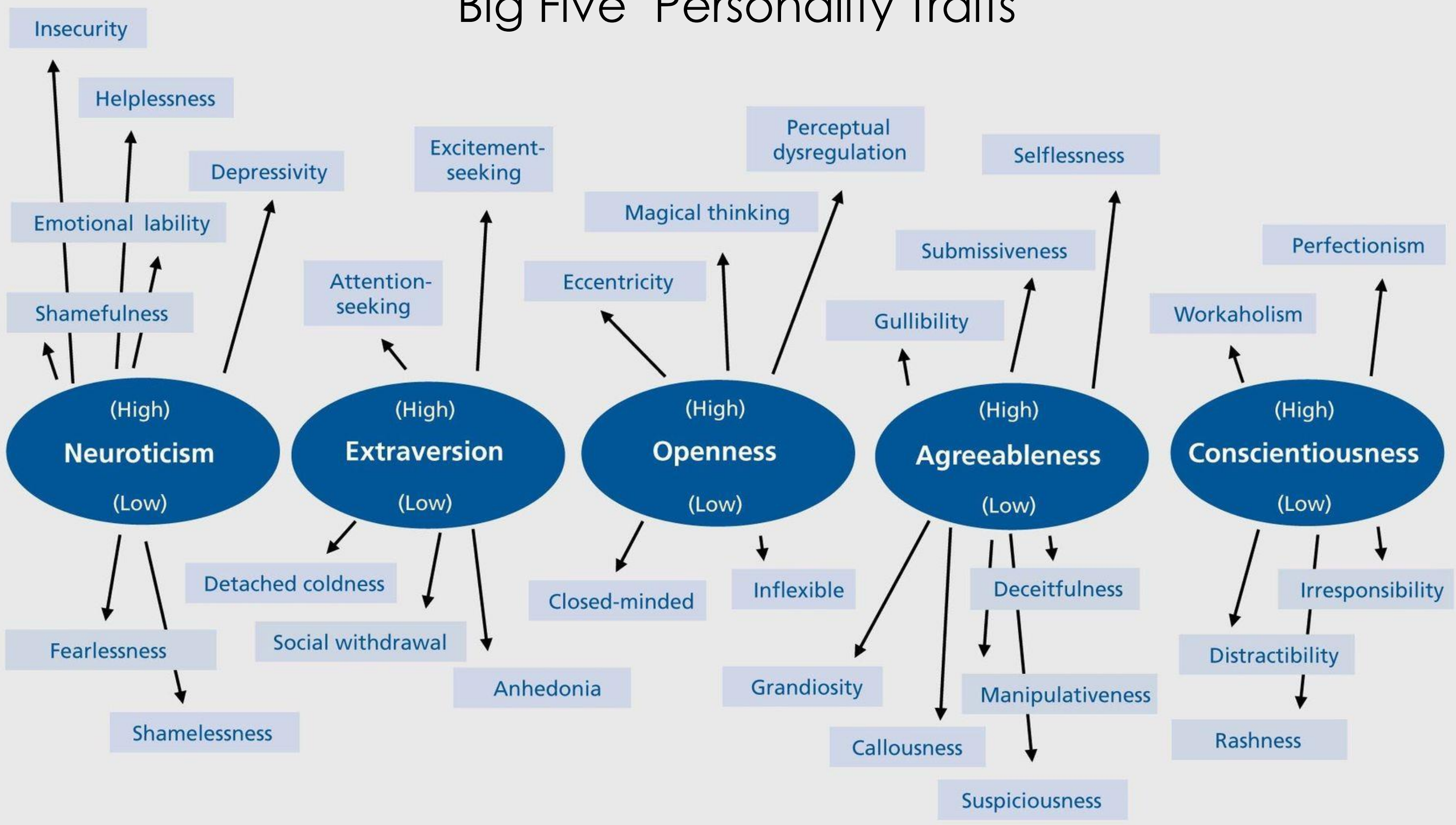
Student Performance Data Set

- Provided by University of Minho, Portugal, via the UCI Machine Learning Repository
- Captures data on two secondary education schools (grades 10-12)
- Contains data on student trimester test scores in math and Portuguese classes
    - For this analysis, the predicted variable is the average of all three grades
- Contains 30 other variables on student background and behavior

Factor Analysis: Overview

- Find variables with strong covariance
- Determine good number of "factors"
  - Factors = unobserved variables reflected by groups of observed variables moving together
- Make that many factors by combining weights of covariant observed variables
- Select a "rotation" that controls how variables are loaded into those factors

"Big Five" Personality Traits



Typical result of applying factor analysis to personality surveys

Sales Example: Correlation

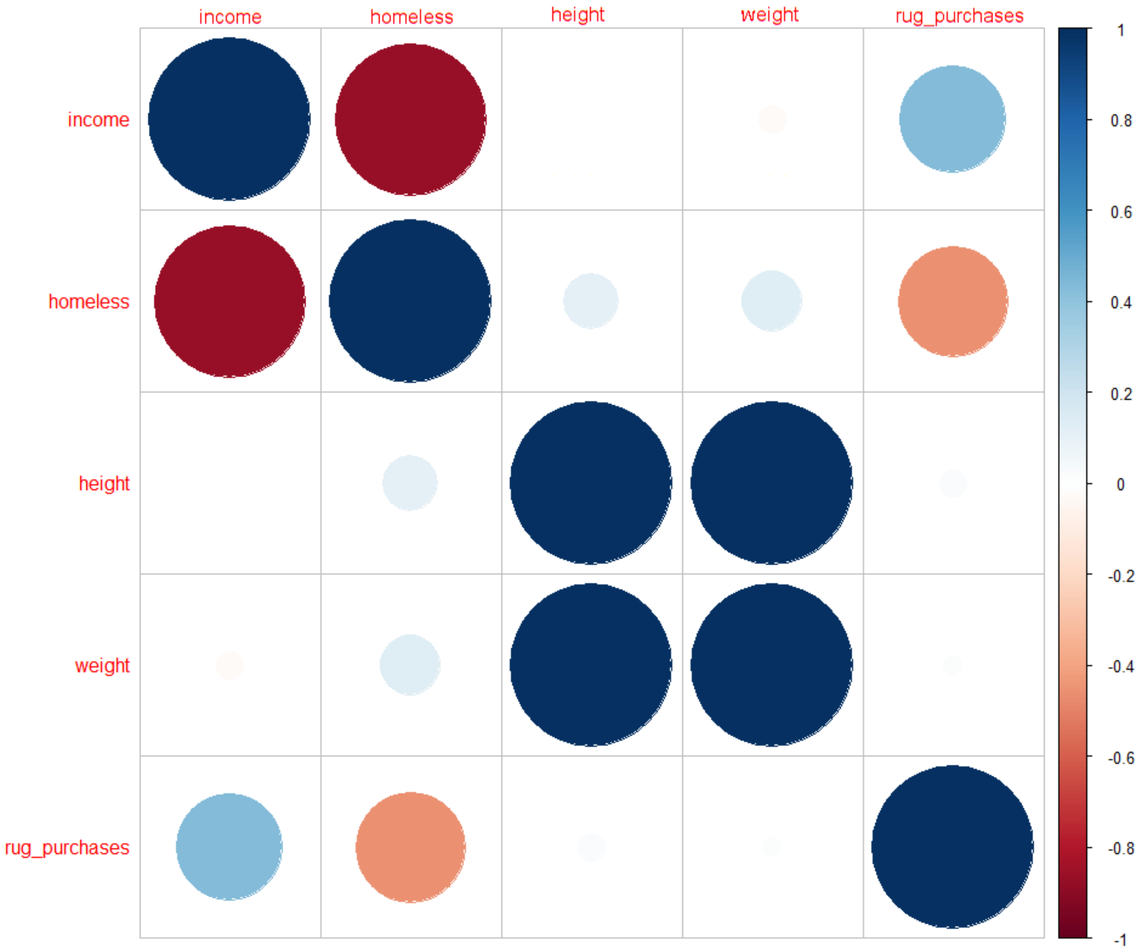
Suppose we're in the business of selling rugs.

We have four variables describing potential customers:

- Their income
- Whether they're homeless
- Height
- Weight

Income and homelessness have strong negative correlation

Height and weight have strong positive correlation



Sales Example: Scree Plot

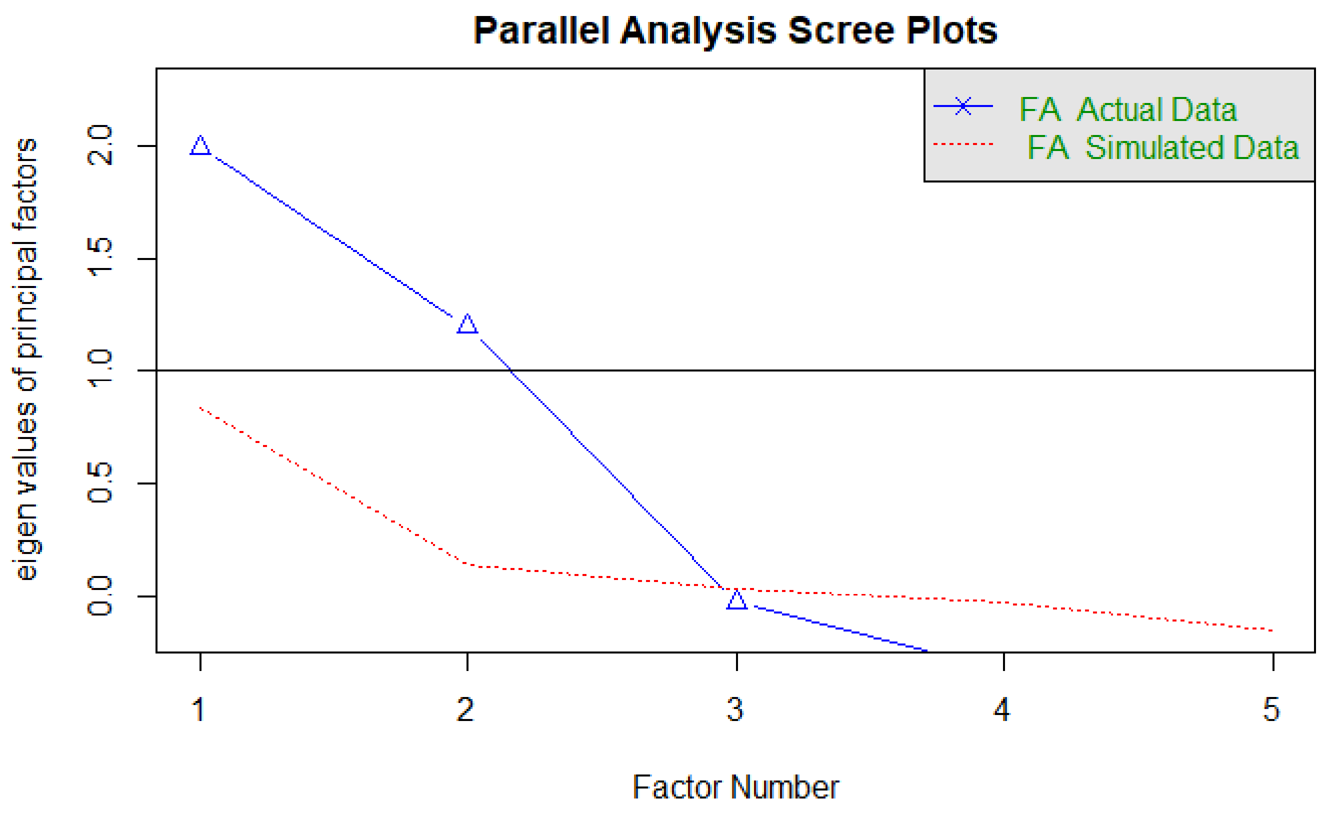
Scree plot:

Charts eigenvalues of the data's correlation matrix.

Eigenvalues >=1 imply greater predictive power than an individual variable.

Here two eigenvectors are >=1;

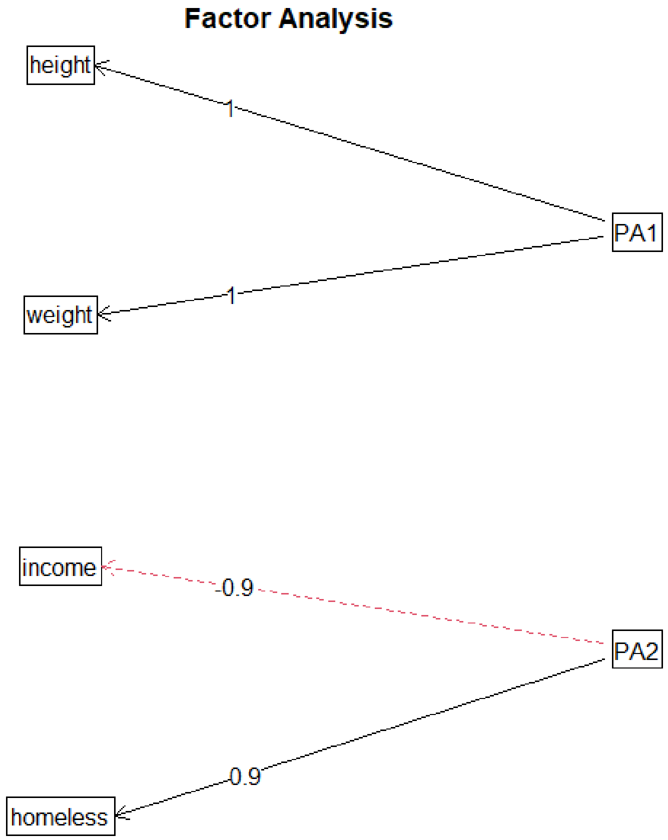
We'll use two factors.



Sales Example: Results

We've decided on two factors and earlier found two sets of covariant variables.

From here we can name our new factors.



Factor	Variables
Customer Size	Height, Weight
Too poor to buy rugs	-Income, Homelessness

```
Call:
lm(formula = purchased ~ ., data = demo.dat)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.4956 -0.3253 -0.1313  0.6283  1.5044 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)  0.66667    0.29043   2.295   0.0405 *
Size         0.05505    0.30077   0.183   0.8578  
TooPoor     -0.56137    0.31146  -1.802   0.0966 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.125 on 12 degrees of freedom
Multiple R-squared:  0.2147,	Adjusted R-squared:  0.08377 
F-statistic:  1.64 on 2 and 12 DF,  p-value: 0.2346
```

Factor "Rotation"

Rotating factors means reshuffling data to fit assumptions made about factors

Two commonly used types:

Varimax rotation

- Assumes factors are orthogonal; do not correlate
- Finds factors that explain maximum variance
- Tends to find simpler/easily understood factor loadings

Oblimin rotation

- Assumes factors can correlate
- Includes orthogonal solutions but rarely selects them
- Generally more realistic for complex systems/psychometry

Why Use Factor Analysis?

Our student dataset has values for...

- School attended
- Age and sex
- Family size
- Parents' marital status
- Parents' education
- Parents' jobs
- Family support (emotional, financial)
- Reason for attending this school
- Interest in higher education
- Extra tutoring
- Travel time to school
- Study habits
- Drinking habits
- Romantic involvement

...and more!

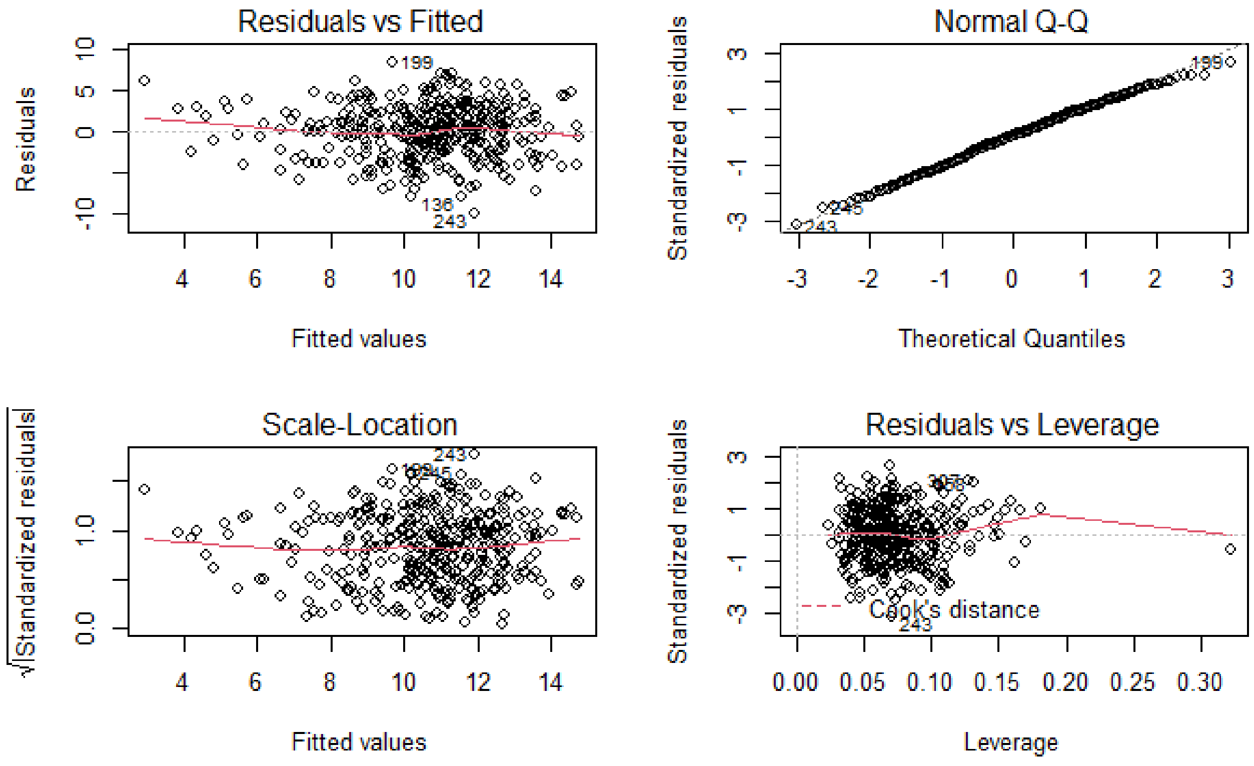
30 is too many variables to summarize simply and surely some group together.

Data Manipulation & Diagnostics

Typical factor analysis requires linear data, benefits from normally distributed data.

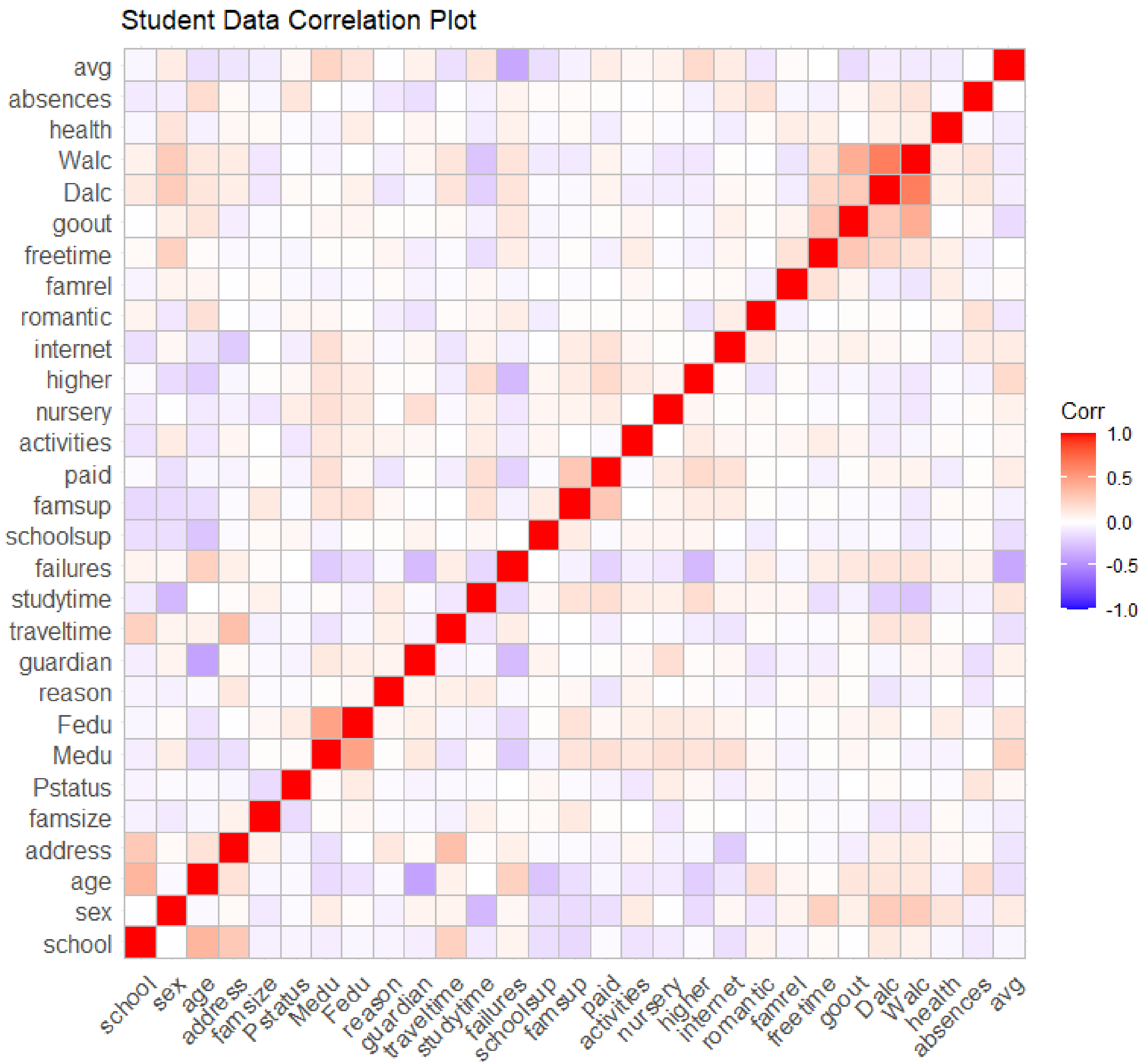
Categorical data was either dropped or coded into binaries where appropriate.

There are no significant problems in the diagnostic plots.



There aren't many obvious groupings of covariant variables.

Luckily, R takes care of making these connections for us.



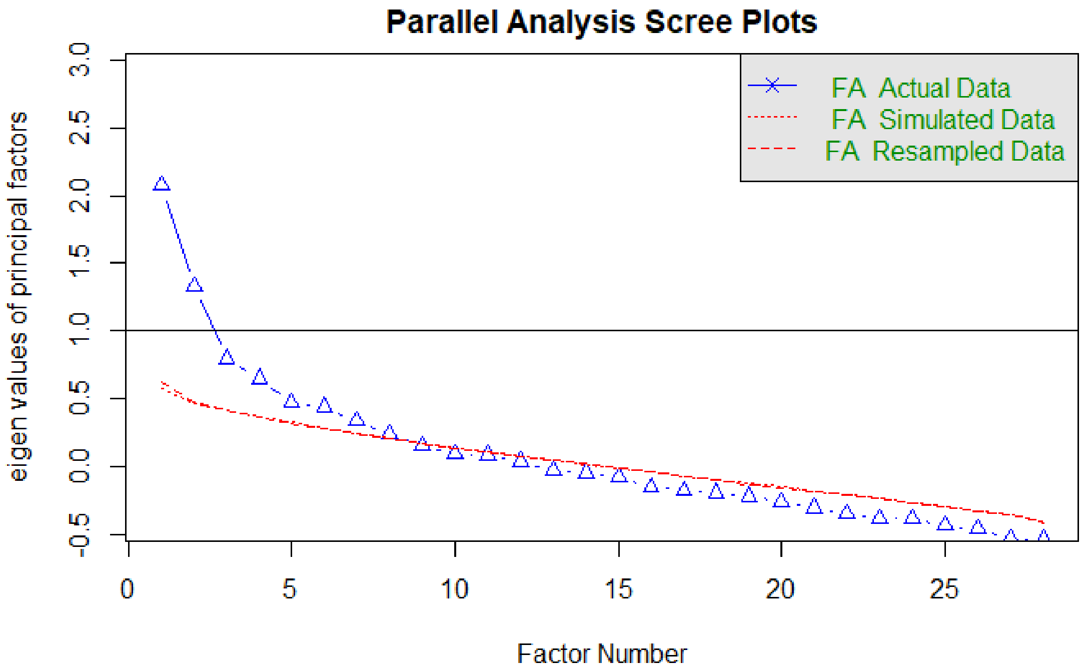
Student Data Scree Plot

Different approaches to Scree plots…

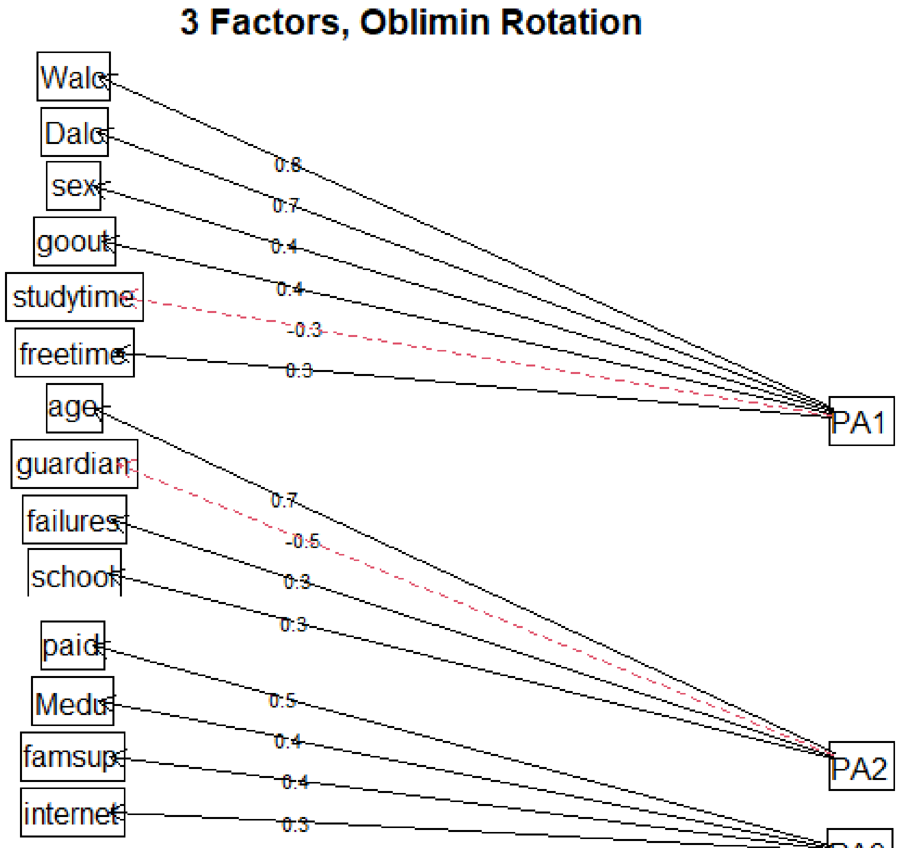
- Stop at $\lambda < 1$: gives us two factors
- Stop at first "leveling off": five factors

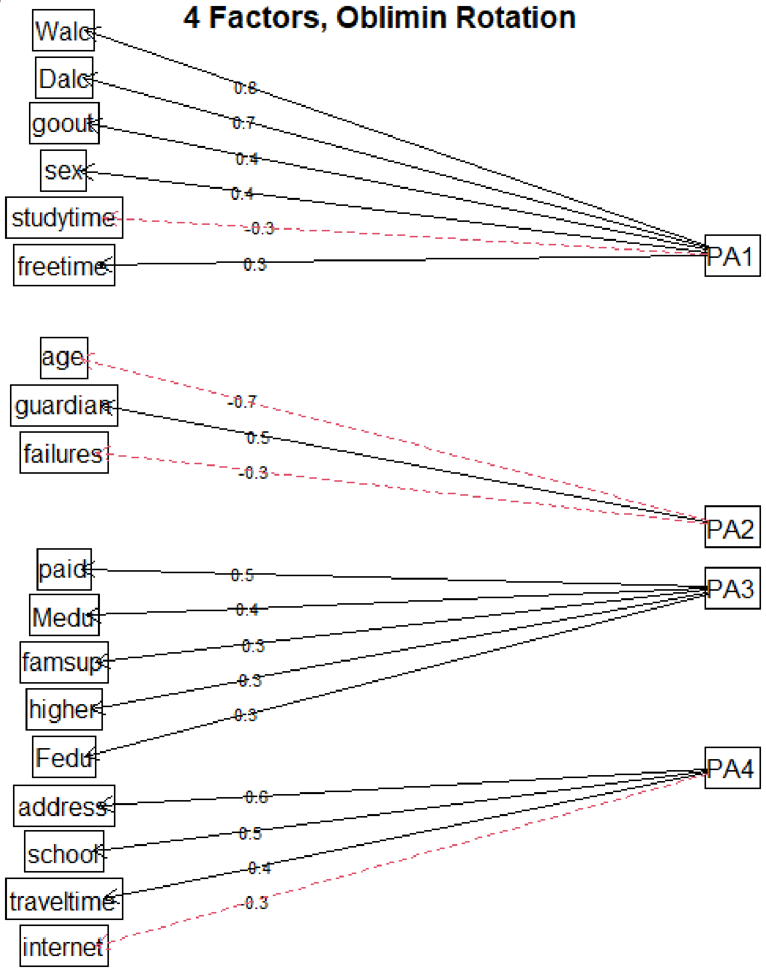
Best number of factors should lie in the interval [2,5]

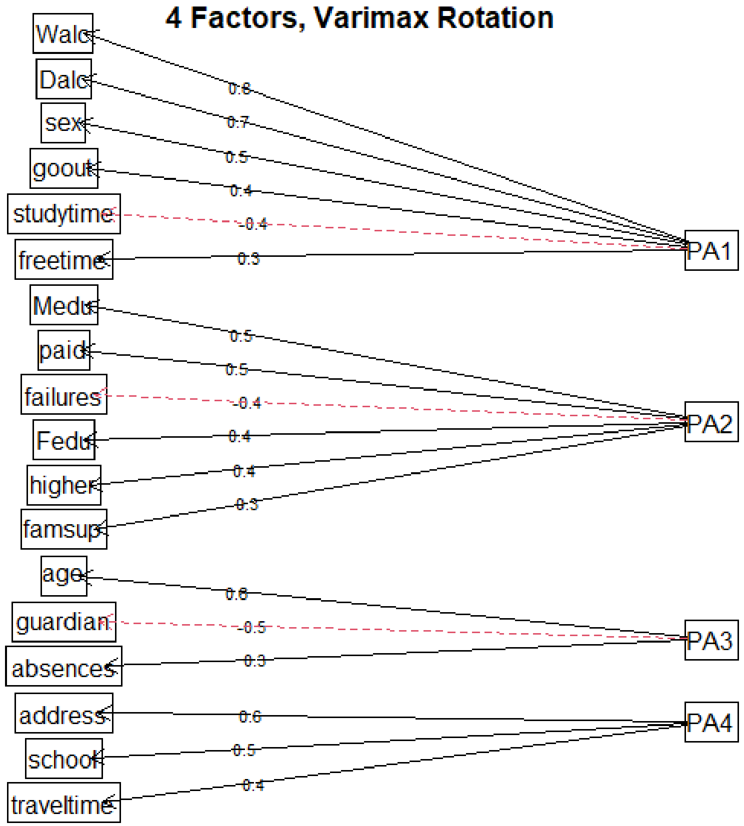
Tests were run on 2,3,4,5 factors using oblimin and varimax rotations.



Examples of Candidate Factorizations



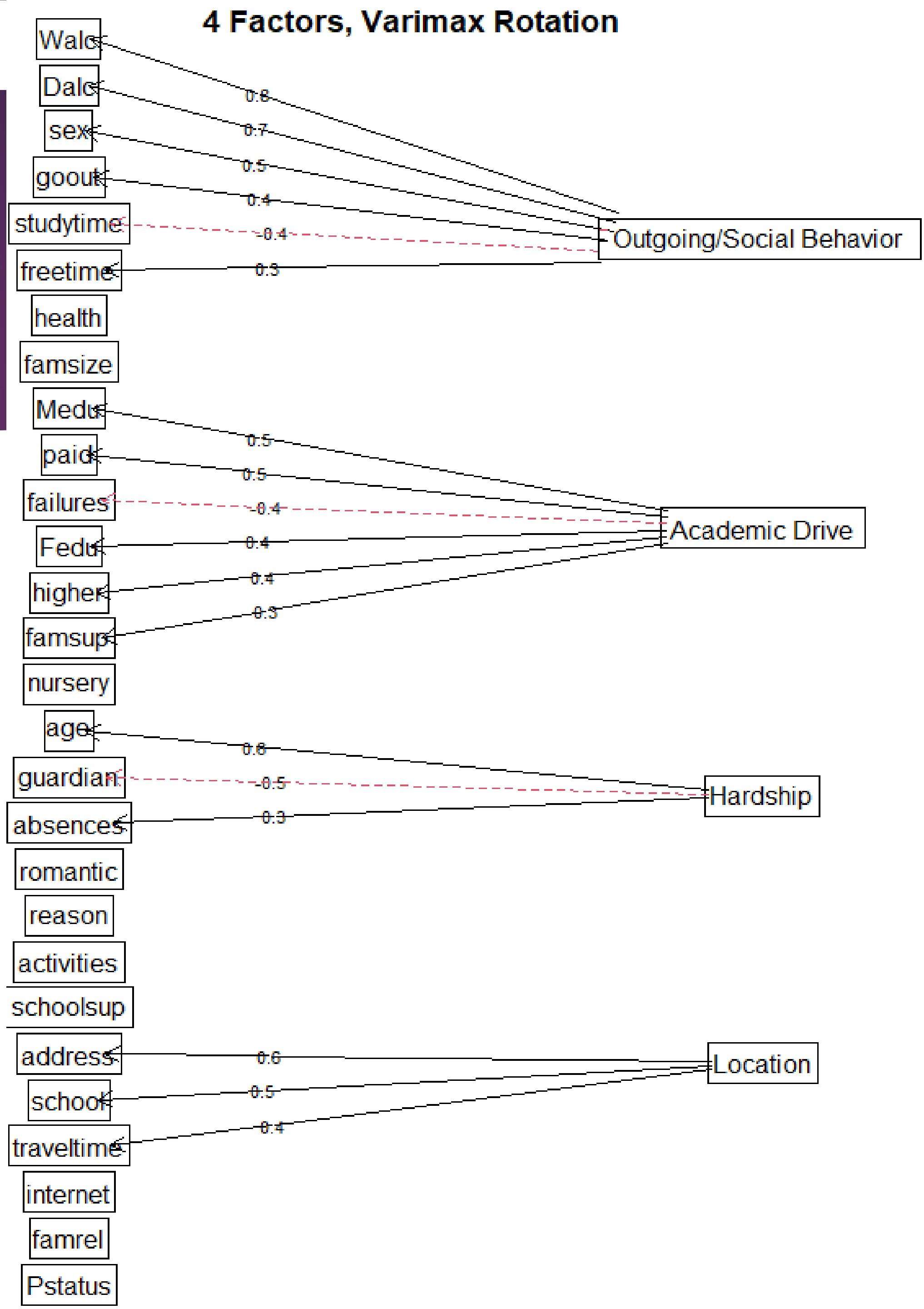




Selected Model

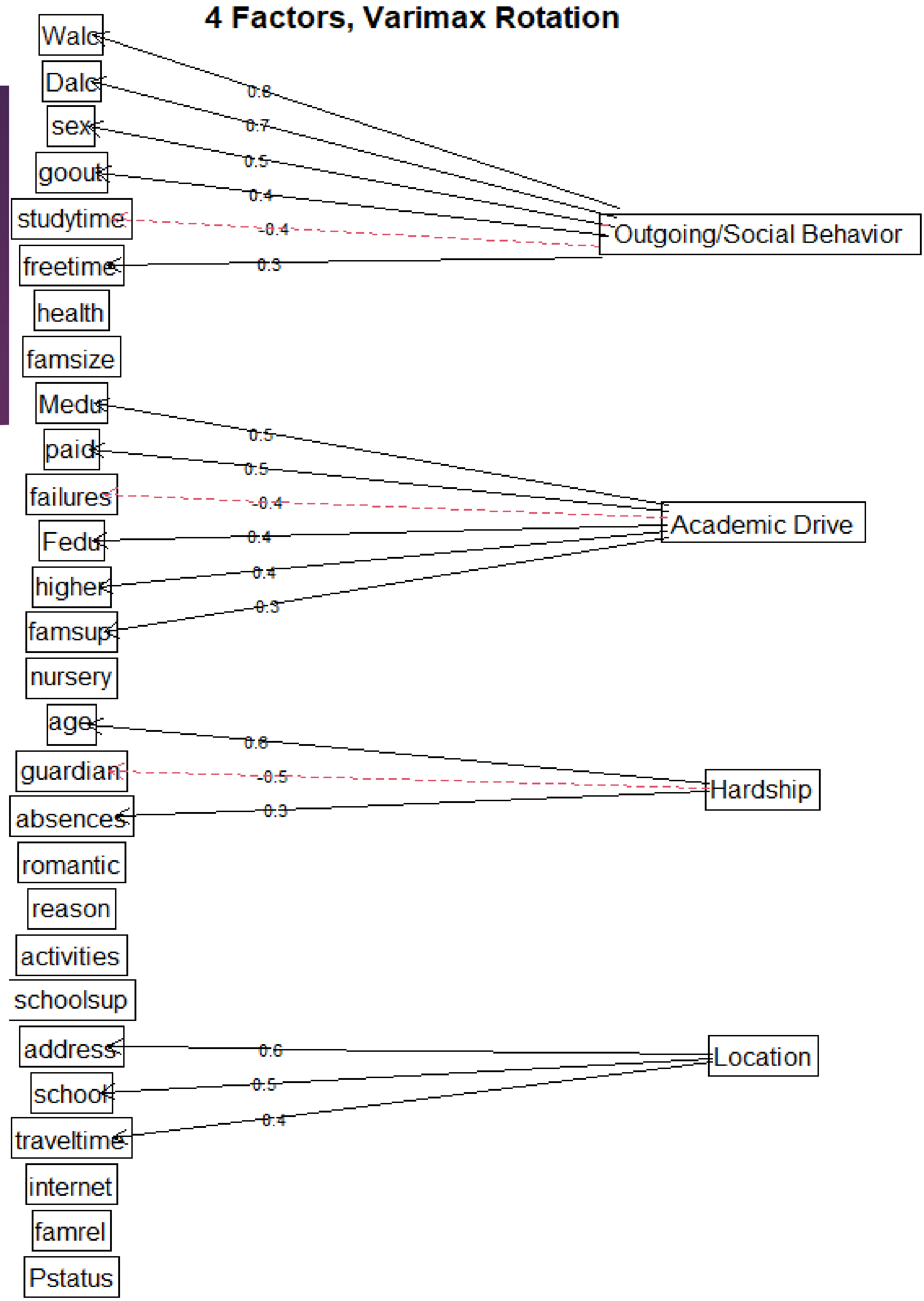
Models were compared and judged based on:

- % of academic score variance explained by factors
- How much the resulting factors "made sense"/could be easily understood
- Residual errors/R-squared values of regression performed on factors



Selected Model

Factor	Loaded variables
Outgoingness/Social Behavior	Weekend and daily alcohol consumption, how often student goes out, how little they study, how much free time they have, biological sex(?)
Academic Drive/Background	Parents' Education, paid tutoring, a lack of previous failures, aspirations to higher education, family's financial support
Hardship	Student's age, student being raised with non-biological parents, student number of absences
Rural Location	Student living in a rural location, going to the more isolated of the two schools observed, taking longer to travel to school



Selected Model... is not very good

.201% Cumulative

Variance Explained

```
Loadings:
           Outgoing/Social Behavior Academic Drive Hardship 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 Behavior Academic Drive Hardship Location
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
```

R-squared 0.103

Adjusted R-squared 0.0938

```
Call:
lm(formula = Scores ~ ., data = math.dat)

Residuals:
     Min       1Q   Median       3Q      Max
-10.0129  -2.1870   0.0083   2.4019   8.3688

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                10.67932    0.17707  60.312  < 2e-16 ***
`Outgoing/Social Behavior` -0.43872    0.26388  -1.663   0.0972 .
`Academic Drive`           -0.50418    0.20994  -2.402   0.0168 *
Hardship                    1.14243    0.22491   5.079 5.88e-07 ***
Location                   -0.07864    0.29822  -0.264   0.7922
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.519 on 390 degrees of freedom
Multiple R-squared:  0.103,     Adjusted R-squared:  0.0938
F-statistic:  11.2 on 4 and 390 DF,  p-value: 1.313e-08
```

What Went Wrong?

- Needing to coerce the data into a linear form to perform factor analysis probably hurt the analysis more than expected
- Most variables had weak correlation to begin with; only 2 factors have predictive power
- Kaiser-Meyer-Olkin factor of sample adequacy finds 0.64 MSA; larger sample might have helped

Citations

- P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7

- *Big five personality traits diagram*. (n.d.). Buffer.com. https://buffer.com/resources/how-the-big-five-personality-traits-can-help-you-build-a-more-effective-team/.