Predicting Student Success from Socioeconomic Factors

Luke Sonnanburg

Abstract

This paper explores the use of factor analysis in predicting the academic success of Portuguese secondary school students studying mathematics. 395 students at two schools were surveyed for primarily self-reported information about their life and habits. This data was used to construct metavariable factors describing students. Students' social behavior, academic interest and support, past hardship, and rural locations found to be possibly influential on student grades. However, findings are inconclusive due to poor fit of model stemming from weakly correlated variables. Includes discussion of how to perform, and when not to perform, factor analysis.

Introduction

In evaluating education outcomes, academic grades are one of the simplest means of measuring and understanding student outcomes. While the matter of whether test and homework scores are an accurate reflection of student learning is beyond the scope of this paper, if we assume that they are then it would be useful to determine factors that predict good scores. On the macro level, investments can be made to bring school environments more in line with positive outcomes. For students, it can be used to identify individuals who may benefit from additional tutoring, guidance, or other outreach.

For this kind of application, it would be useful if the factors that predict student success were also easily understood. To accomplish that, in this report we attempt to apply factor analysis to data pertaining to 10-12 grade students from two Portuguese secondary schools. If successful, we will glean metavariables that capture broad descriptors of students and circumstances that correlate to academic success and failure and use it to prescribe strategies for education going forward.

Data Set

The data was originally collected for a similar study by researchers investigating causes of Portugal's high rates of student failure and dropout. The data contains three variables (G1, G2, G3) representing trimester-based mathematics grades and 30 other variables representing the same students' identity (sex, age), home life (family relationship, location), financial situation (familial or institutional support, paid tutoring), academic experience and aspirations, and social behavior (drinking, partying). Since these variables were only recorded once and are likely mostly stable, and this analysis is intended to consider practical, actionable variables, the grades of the first two trimesters are not considered separately, and all grades are combined into a variable representing average scores over the course of the year. The data was collected from 395 students at two different schools. Certain variables were changed from two-level factors to integers in order to be able to apply linear factor analysis. For full definitions of all 30 variables, see table 1 in the appendix.

Exploratory Analysis

A cursory survey of the data does not look good for students as is. Grades range from 0 to 20, with 10 considered passing. The average grade is 10.679. Only 58% of students pass at all. Just over 6% score averages above 16, which is considered "very good". For score distribution, see fig. 1 in the appendix. Most numeric data is normally distributed; an exception is Dalc, representing weekday alcohol consumption, which skews very low, and student health, which was reported to be very high. About a third of mothers have had a higher education, as opposed to roughly a quarter of fathers. There are other patterns that may be interesting, but for this paper's methodology of factor analysis, it is not critical to study these variables individually.

Factor Analysis Overview

For this study, we are employing factor analysis. Factor analysis is a tool used to reduce high-dimension data to a simpler form to interpret and predict data. It accomplishes this by finding explanatory variables with strong covariance and collecting them into groups intended to represent unobserved metavariables

that capture some unobserved trait in the subject. Upon grouping them, it hopefully makes some kind of intuitive sense that the variables would move together in this way, enough that it's possible to give the group a name that's descriptive and easy to understand.

To illustrate with a common example, consider personality tests or questionnaires:
A questionnaire might ask seemingly unrelated questions about pragmatism, independence, curiosity, and a preference for routine. The test's authors may have had no specific intention for those questions, but it's likely that high curiosity moves with low preference for routine. If the other two covary in similar patterns, they may all be grouped into one "factor", or unobserved metavariable. High independence and curiosity, combined with low pragmatism and preference for routine, might be called something like "Openness to new experiences". Conversely, you could associate them with the opposite polarities i.e., high preference for routine and low curiosity, and call it something like "rigidity".

The ideal result will be significantly fewer factors than there were starting variables, grouped in ways that can be easily named.

Methodology: Factor Analysis Application

Factor analysis has fewer requirements than many methods of statistical analysis. Non-numeric data has already been manipulated to be linear, satisfying its only strict demand. However, there is an issue. For reasons that will be discussed, factor analysis requires some variables be highly covariant to yield interesting results. In the correlation plot (figure 2) there seem to be few variables with a covariance with a magnitude of .5 or higher. Weekend and daily alcohol consumption are correlated, previous failures are negatively correlated with several variables, but few other strong relationships are easy to visually pick out.

There is a measure of how good a fit factor analysis is on a data set, the Kaiser-Meyer-Olkin factor adequacy test. This data scores at a measure of sample adequacy of .62, which is considered a "mediocre" fit, narrowly above the .50-.59 range, which is considered "miserable". This is a direct result of the correlation matrix for our data being at mostly similar levels of covariance, leading to it being difficult to pick out groups of variables that move together.
To conduct factor analysis, you need to first determine the number of factors that you could end up with. Methods for determining this number come from the data's scree plot (figure 3), which represents the eigenvalues of the data's correlation matrix. In general, vectors with greater magnitude values imply greater eigenvalues. For each eigenvalue equal to or greater than 1, there exists a factor that we can extract from our data that explains more of the output variable's variation than any one explanatory variable individually can. There are several approaches to interpreting a scree plot. Most conservatively, you can use only as many factors as there are eigenvalues greater than 1; in our case this would mean 2 factors. Another strategy is to cut off the number after an eigenvalue that starts a trend of data leveling off, as in the case of eigenvalue 5 being barely any greater than eigenvalue 6. The correct approach is somewhat subjective, but from these two schools of thought it seems likely that the "best" number of factors lies somewhere between 2 and 5.

The only decision left to make in standard factor analysis is how to rotate the data to fit into factors. Data rotations are methods of adjusting the coordinates of data in relation to each other so that they cluster differently. In factor analysis, this is generally done so that they cluster to maximize variance, separating variables into factors more strictly and increasing their predictive power of the output variable. There are many types of rotation, but they are generally related to one of two common rotations:

Varimax rotations are done assuming that, while individual *variables* may covary, separate *factors* do not covary at all. This forces the coordinates representing factor data to be orthogonal, absolutely maximizing the variance between factors. This tends to yield easily understood results, but may not be realistic representations of complicated subjects such as students, who have many traits that may covary.

Oblimin rotations still attempt to maximize variance, but does not enforce this orthogonality. Factors are still allowed to covary. Oblimin rotations are inclusive of varimax solutions, but usually won't conclude with them.

For this analysis, we exhaustively tried every model with 2, 3, 4, and 5 factors, using varimax and oblimin rotations. Candidate factor analysis were judged based on how well they fit the data they came from, and how viable it was to name the generated factors.

Results

The selected model (figure 4) utilizes four factors and varimax rotation.
Those factors are (reporting variables with >30% covariance with the factor they're loaded into):

- Outgoing/Social behavior: This includes alcohol consumption, amount of free time, lack of dedicated study time, frequency with which the student goes out with friends, and whether the student's sex is female. This has a negative correlation with good grades.
- Academic Drive/Background: Parents with higher education, parents who can support education in the home, taking extra paid lessons, a desire to pursue higher education, and a lack of previous academic failures. This has a strong positive correlation with good grades.
- Hardship: Students being older, having more absences, and being raised by someone other than their biological parents. This has a negative correlation with good grades.
- Rural location: Students having a rural address, going to the more rural school (Mousinho da Silveira), or taking longer to travel to school. This has a negative correlation with good grades.

The summary of a simple least-squares linear regression model relating student grades to these factors can be seen in figure 6. Statistically significant influences on grades where the student body mean grade of 10.6793, whether or not students have academic drive/an academic familial background, and whether students fall into the "hardship" category.

To measure how well the model fits the data it was constructed from, we can refer to the multiple R-squared of .09424 and adjusted R-squared of 0.08495.

Conclusion

Simply put, the resulting model is not satisfactory. An R-squared value of .09424 suggests that the students' grades are not well-explained by the selected factors. However, out of the candidate models, it is one of the best. Also, this model has one of the easier arrays of factors to make coherent sense of.

It is surprising to see that biological sex should be grouped into the same factor as social behavior that predicts grades, but upon reviewing the correlation graph that covariance is an apparent pattern in this data.

The second factor is likely the least surprising: students who want to succeed and have a history of performing well, who also have a family that can support them academically and financially, are going to have an edge over other students.

"Hardship" is a somewhat ambiguous name selected on the basis that being raised by non-family members, missing school, and being old for your class are all reasonable results of various unfortunate life obstacles.

Rural location makes sense as a factor, though how much of this factor's influence might come from the inconveniences of traveling significantly farther to school and other services, and how much might come simply from the more remote school having fewer resources, is unclear from this.

If I were to recommend action based on these results, they are simply:

- Encourage good, routine study habits
- Offer more paid lessons, with discounts or incentives for students facing known hardship
- Offer tools to help parents be involved in their child's education at home

But ultimately, it appears that factor analysis is simply not a good fit for this data due to both the muddled, almost-uniform covariance of variables, and the need to manipulate the data to be linear so that factor analysis may be performed. Other methodologies can likely find effective predictions with this same data. In order to find underlying factors that can easily summarize traits of effective students, it would be more useful to collect observations of a more standardized nature.

Appendix

Variable	Explanation
School*	Highschool student attends; 0 = Gabriel Pereira, 1 = Mousinho de Silveira
Sex*	Student's biological sex; 0 = male, 1 = female
Address*	Geographic descriptor of student's home; 0 = urban, 1 = rural
Famsize*	If student's immediately family has at least 3 children, 1. Otherwise, 0.
Pstatus*	Marital status of student's parents; 0 = together, 1 = separated
Medu, Fedu*	Mother/Father's education level; 1 = higher education, otherwise 0.
Guardian*	If student's legal guardian is a biological parent 1, otherwise 0
Traveltime	Time student takes to travel to school; integer measuring units of 15 minutes
Studytime	Integer representing student's self-reported time spent studying
Failures	Number of student's previous failed courses
Schoolsup*	If a student receives extra educational support from their school: 1, otherwise 0
Famsup*	If a student receives extra educational support from their family: 1, otherwise 0
Paid*	If a student received extra paid classes: 1, otherwise 0
Activities*	If a student participates in extra-curricular activities: 1, otherwise 0
Nursery*	If a student previously attended nursery school: 1, otherwise 0
Higher*	If a student intends to pursue higher education: 1, otherwise 0
Internet*	If a student has internet access at home: 1, otherwise 0
Romantic*	If a student is involved in a romantic relationship: 1, otherwise 0
Famrel	Student's self-reported quality of family relationship, scale of 1 – low to 5 – high
Freetime	Student's amount of free time after school, scale of 1 – low to 5 – high
goout	Frequency student goes out with friends, scale of 1 – low to 5 - high
Dalc	Workday alcohol consumption, scale of 1 – low to 5 - high
Walc	Weekend alcohol consumption, scale of 1 – low to 5 - high
Health	Student's current health status, scale of 1 – bad to 5 - good
Absences	Integer representing number of days of school student missed
Avg	Student's average math grades across an entire year, scores range 0-20 with 10 considered passing

*Table 1: Data Definitions. * = variables who were recoded from factors to integers.*

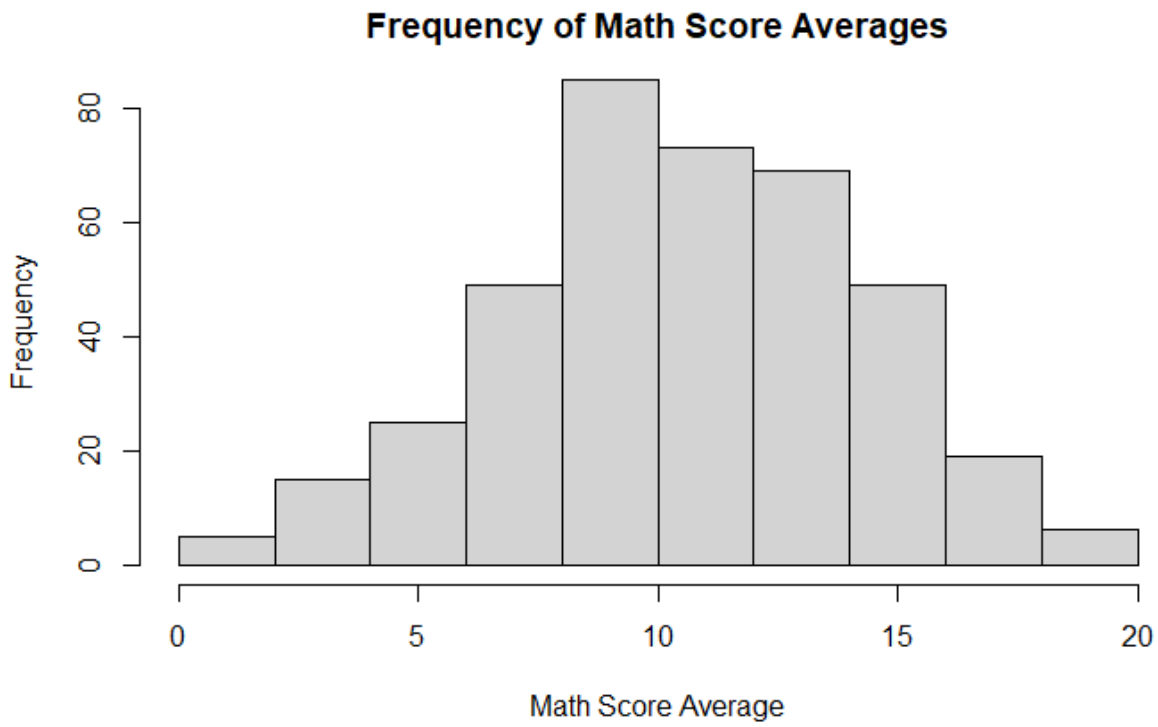

Figure 1: Average grade distribution

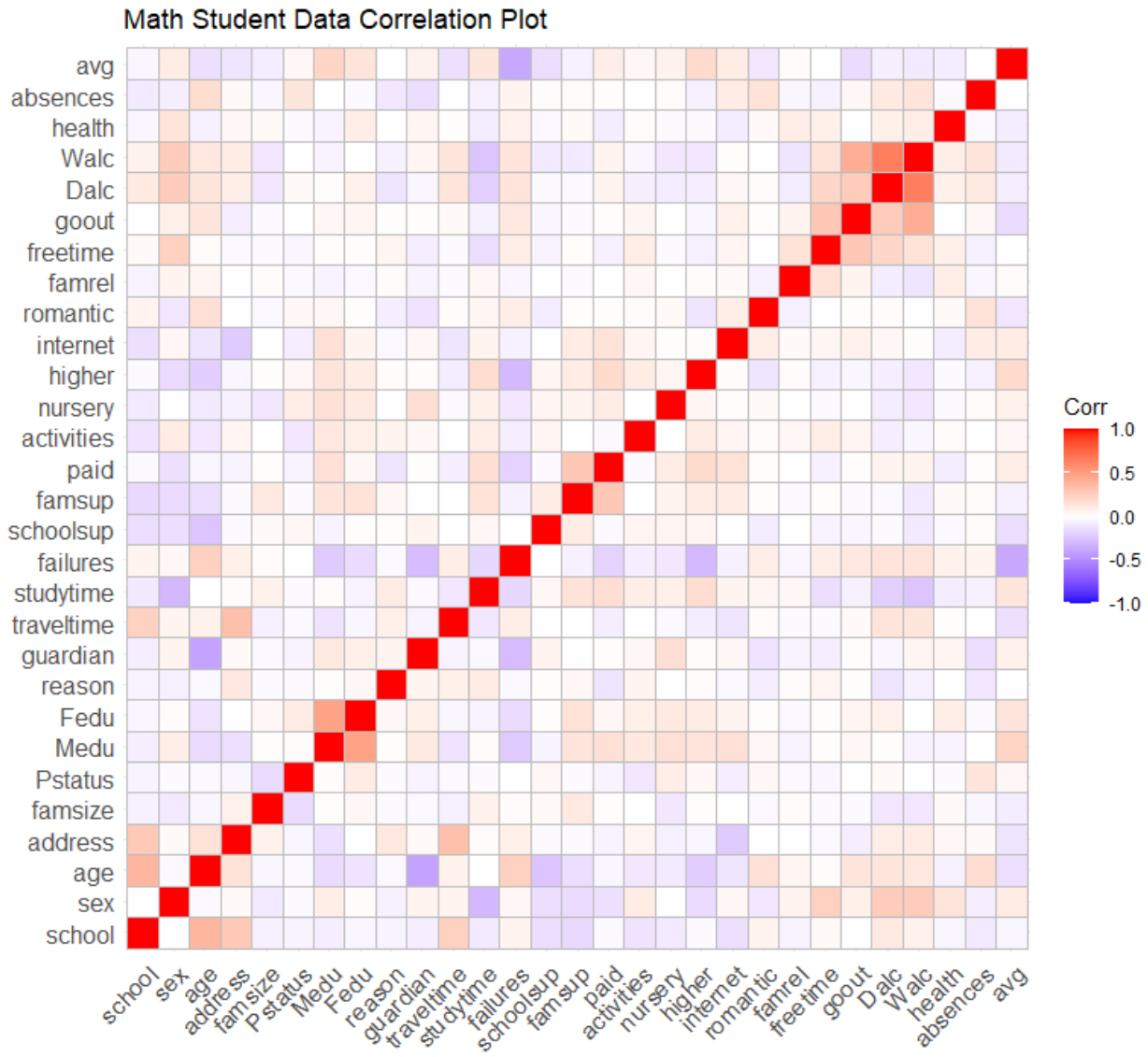

Figure 2: Student data correlation plot

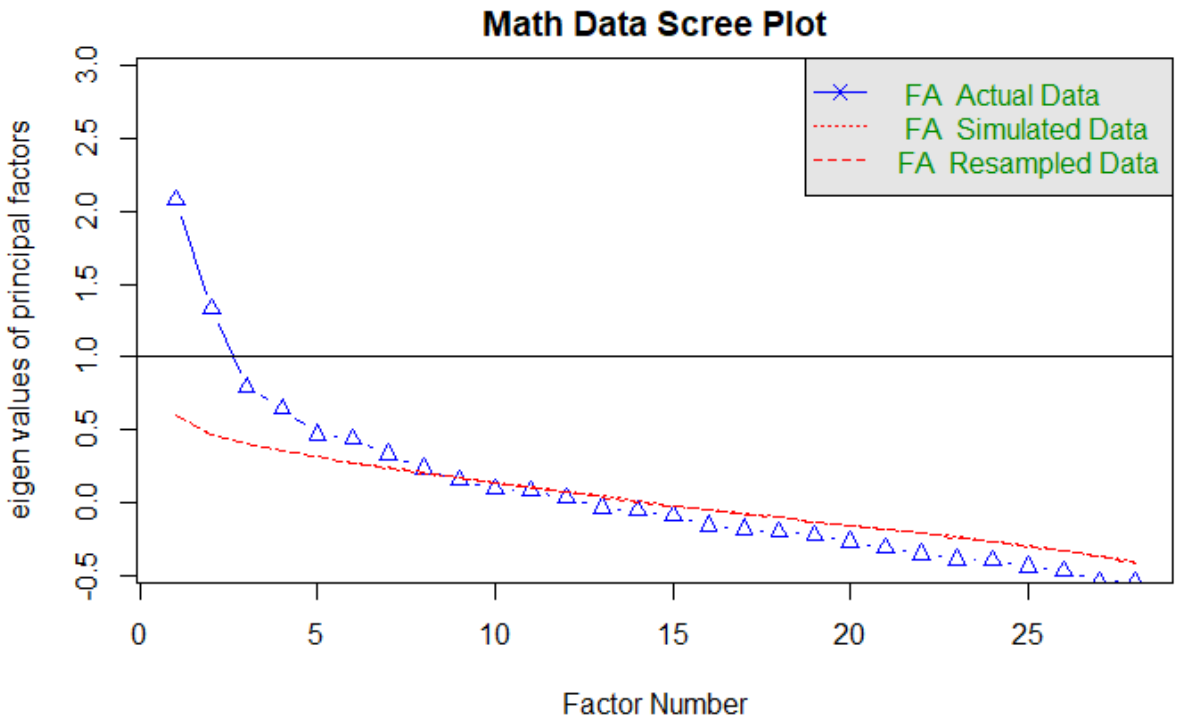

Figure 3: Scree plot of correlation eigenvalues

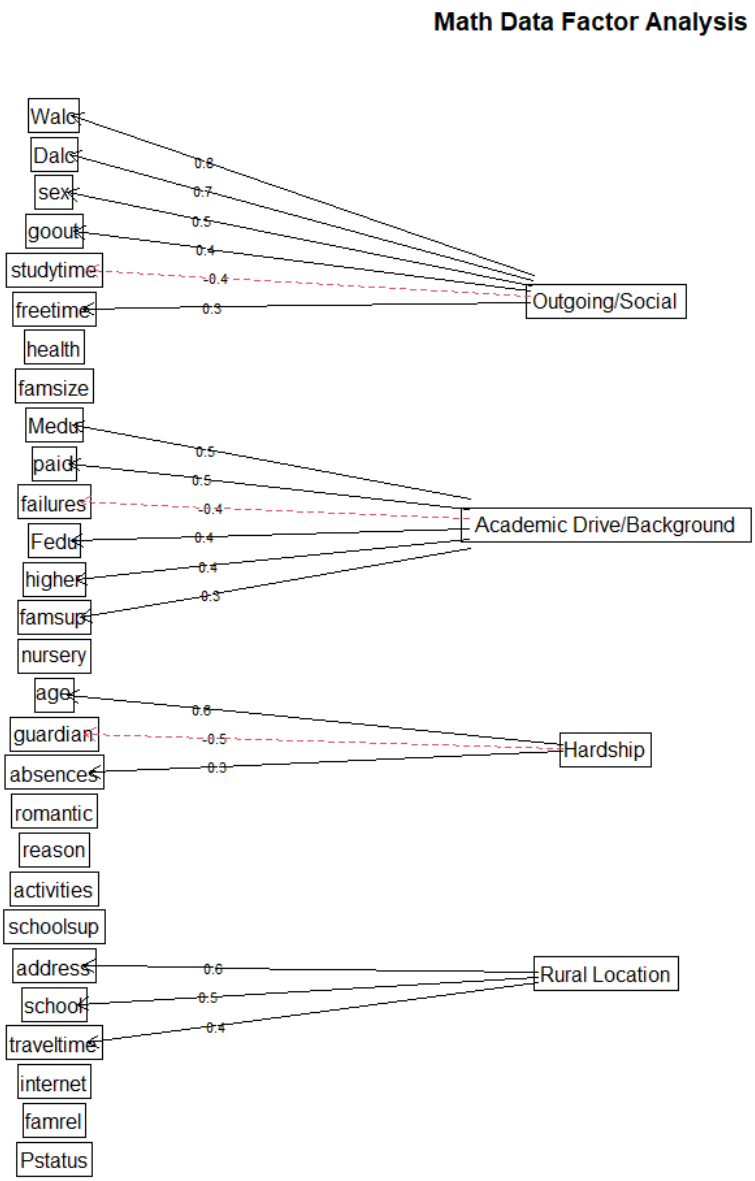

Figure 4: Resulting factors

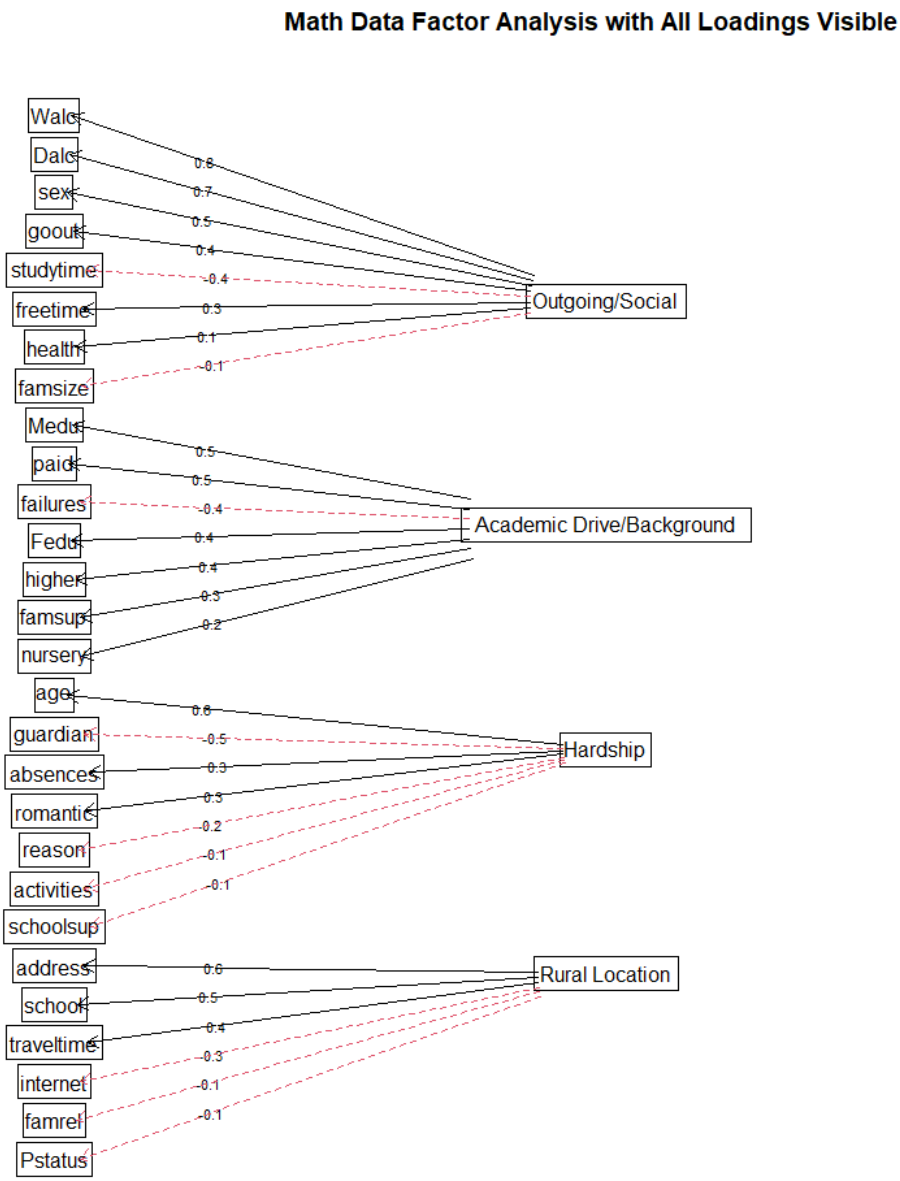

Figure 5: Resulting factors showing all loadings

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

Residuals:
    Min      1Q  Median      3Q     Max
-8.8989 -2.1749  0.0318  2.4684  9.2677

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                10.6793     0.1779  60.020  < 2e-16 ***
`Outgoing/Social`          -0.2604     0.2031  -1.282  0.20054
`Academic Drive/Background`  1.1223     0.2270   4.943 1.14e-06 ***
Hardship                   -0.5931     0.2252  -2.633  0.00879 **
`Rural Location`           -0.3343     0.2320  -1.441  0.15038
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.536 on 390 degrees of freedom
Multiple R-squared:  0.09424,   Adjusted R-squared:  0.08495
F-statistic: 10.14 on 4 and 390 DF,  p-value: 8.032e-08
```

Figure 6: Summary of the terrible resulting model

Works Cited

Cortez, Paulo & Silva, Alice. (2008). Using data mining to predict secondary school student performance. EUROSIS.