

R Notebook: Multiple Regression Model of Student Academic Achievement

Tristen Bristow

April 6, 2017

Abstract

The interest of this study is in developing a prediction model of student success based on measured factors of success in Mathematics and Portuguese. Multiple regression is applied to develop a regression classifier based on student-provided factors that relate to living conditions and education conditions. Though regression tree modeling would appear at first to be the correct approach, the number and high-cardinality nature of many of the variables in this data makes such an approach less feasible in practice.

```
set.seed(1)
library(car)
library(boot)
c <- read.table("student-por.csv", sep=";", header=TRUE)
c <- data.frame(c)
d <- read.table("student-mat.csv", sep=";", header=TRUE)
d <- data.frame(d)
e <- rbind(c,d)
```

Data Cleaning

Both Math and Portuguese sets are merged, alternate column titles are applied, and all student grades are averaged across three grade entries.

```
names(e) <- c("school", "sex", "age", "address", "family size", "parents
cohab.", "mom's education",
              "dad's education", "mom's job", "dad's job", "reason",
              "guardian", "travel", "study",
```

```

      "failures", "education support", "family
support", "paid", "activities", "nursery", "higher",
      "internet", "romantic", "family bond", "free
time", "social", "workday alch.", "weekend alch.", "health",
      "absences", "Grade 1", "Grade 2", "Grade 3")

Grade <- (e$`Grade 1` + e$`Grade 2` + e$`Grade 3`) / 3
e$`Grade 1` <- NULL
e$`Grade 2` <- NULL
e$`Grade 3` <- NULL
e = cbind(e, Grade)
attach(e)

```

Exploratory Investigation

From initial inspection it is clear education success is quantified by the Grade variable. Results of inspection indicate general normality of this output variable.

The class distributions of explanatory variables, 'dad's job' and 'mom's job', appear to show questionable value by inspection of the summary table. This is indicated by the limited difference between class-levels, except for the vaguely defined class, 'other', showing the survey question isn't well-defined or reliable an indicator.

An inspection of the VIF's (Variance Inflation Factors) of model parameters is performed to check for multicollinearity in the data-set.

```
summary(e)
```

##	school	sex	age	address	family size	parents	cohab.
##	GP:772	F:591	Min. :15.00	R:285	GT3:738	A:121	
##	MS:272	M:453	1st Qu.:16.00	U:759	LE3:306	T:923	
##			Median :17.00				
##			Mean :16.73				
##			3rd Qu.:18.00				
##			Max. :22.00				
##	mom's education		dad's education		mom's job		dad's job
##	Min. :0.000		Min. :0.000		at_home :194		at_home : 62
##	1st Qu.:2.000		1st Qu.:1.000		health : 82		health : 41
##	Median :3.000		Median :2.000		other :399		other :584
##	Mean :2.603		Mean :2.388		services:239		services:292
##	3rd Qu.:4.000		3rd Qu.:3.000		teacher :130		teacher : 65
##	Max. :4.000		Max. :4.000				

```

##          reason      guardian      travel      study
## course      :430    father:243    Min.      :1.000    Min.      :1.00
## home        :258    mother:728    1st Qu.:1.000    1st Qu.:1.00
## other       :108    other : 73    Median :1.000    Median :2.00
## reputation:248                Mean   :1.523    Mean    :1.97
##                3rd Qu.:2.000    3rd Qu.:2.00
##                Max.    :4.000    Max.    :4.00
##      failures      education support family support paid
activities
## Min.      :0.0000    no :925                no :404                no :824    no :528
## 1st Qu.:0.0000    yes:119                yes:640                yes:220    yes:516
## Median :0.0000
## Mean     :0.2644
## 3rd Qu.:0.0000
## Max.     :3.0000

## nursery  higher  internet  romantic  family bond  free time
## no :209    no : 89    no :217    no :673    Min.      :1.000    Min.      :
1.000
## yes:835    yes:955    yes:827    yes:371    1st Qu.:4.000    1st
Qu.:3.000
##                Median :4.000    Median :
3.000
##                Mean   :3.936    Mean    :
3.201
##                3rd Qu.:5.000    3rd
Qu.:4.000
##                Max.    :5.000    Max.    :
5.000

##      social      workday alch.  weekend alch.      health
## Min.      :1.000    Min.      :1.000    Min.      :1.000    Min.      :1.000
## 1st Qu.:2.000    1st Qu.:1.000    1st Qu.:1.000    1st Qu.:3.000
## Median :3.000    Median :1.000    Median :2.000    Median :4.000
## Mean     :3.156    Mean     :1.494    Mean     :2.284    Mean     :3.543
## 3rd Qu.:4.000    3rd Qu.:2.000    3rd Qu.:3.000    3rd Qu.:5.000
## Max.     :5.000    Max.     :5.000    Max.     :5.000    Max.     :5.000
##      absences      Grade
## Min.      : 0.000    Min.      : 1.333
## 1st Qu.: 0.000    1st Qu.: 9.333
## Median : 2.000    Median :11.333
## Mean     : 4.435    Mean     :11.267
## 3rd Qu.: 6.000    3rd Qu.:13.333
## Max.     :75.000    Max.     :19.333

```

```
hist(Grade)
```

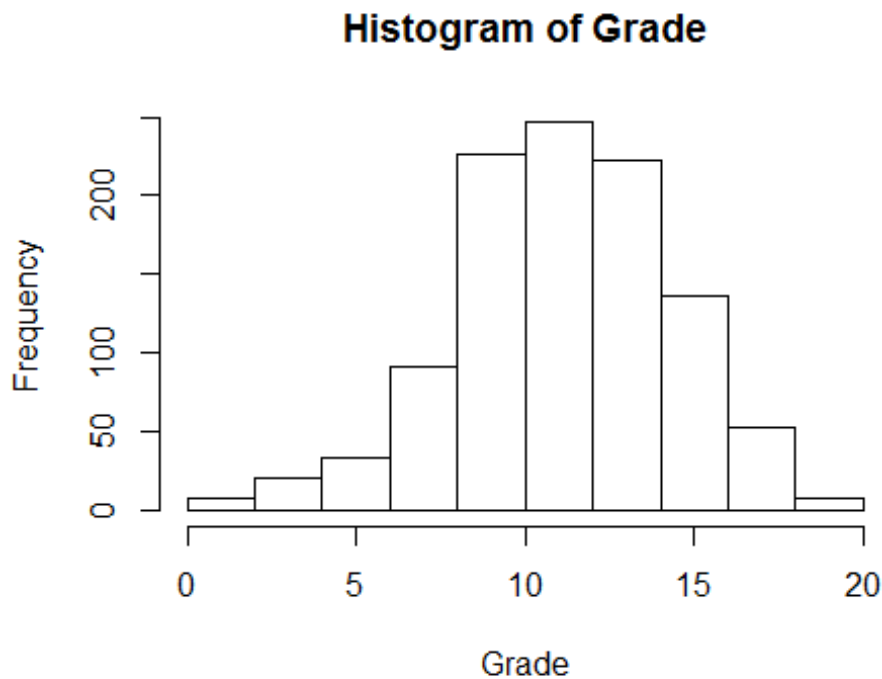


Fig 1: Student-grade frequency distribution.

```
str(e)
```

```
## 'data.frame':    1044 obs. of  31 variables:
## $ school          : Factor w/ 2 levels "GP","MS": 1 1 1 1 1 1 1 1
1 1 ...
## $ sex             : Factor w/ 2 levels "F","M": 1 1 1 1 1 2 2 1 2
2 ...
## $ age            : int  18 17 15 15 16 16 16 17 15 15 ...
## $ address        : Factor w/ 2 levels "R","U": 2 2 2 2 2 2 2 2 2
2 ...
## $ family size     : Factor w/ 2 levels "GT3","LE3": 1 1 2 1 1 2 2
1 2 1 ...
## $ parents cohab.  : Factor w/ 2 levels "A","T": 1 2 2 2 2 2 2 1 1
2 ...
## $ mom's education : int  4 1 1 4 3 4 2 4 3 3 ...
## $ dad's education : int  4 1 1 2 3 3 2 4 2 4 ...
## $ mom's job       : Factor w/ 5 levels "at_home","health",...: 1 1
1 2 3 4 3 3 4 3 ...
## $ dad's job       : Factor w/ 5 levels "at_home","health",...: 5 3
3 4 3 3 3 5 3 3 ...
## $ reason          : Factor w/ 4 levels "course","home",...: 1 1 3 2
```

```

2 4 2 2 2 2 ...
## $ guardian      : Factor w/ 3 levels "father","mother",...: 2 1 2
2 1 2 2 2 2 2 ...
## $ travel        : int  2 1 1 1 1 1 1 2 1 1 ...
## $ study         : int  2 2 2 3 2 2 2 2 2 2 ...
## $ failures      : int  0 0 0 0 0 0 0 0 0 0 ...
## $ education support: Factor w/ 2 levels "no","yes": 2 1 2 1 1 1 1 2
1 1 ...
## $ family support  : Factor w/ 2 levels "no","yes": 1 2 1 2 2 2 1 2
2 2 ...
## $ paid           : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1
1 1 ...
## $ activities     : Factor w/ 2 levels "no","yes": 1 1 1 2 1 2 1 1
1 2 ...
## $ nursery        : Factor w/ 2 levels "no","yes": 2 1 2 2 2 2 2 2
2 2 ...
## $ higher         : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2
2 2 ...
## $ internet       : Factor w/ 2 levels "no","yes": 1 2 2 2 1 2 2 1
2 2 ...
## $ romantic       : Factor w/ 2 levels "no","yes": 1 1 1 2 1 1 1 1
1 1 ...
## $ family bond    : int  4 5 4 3 4 5 4 4 4 5 ...
## $ free time      : int  3 3 3 2 3 4 4 1 2 5 ...
## $ social         : int  4 3 2 2 2 2 4 4 2 1 ...
## $ workday alch.  : int  1 1 2 1 1 1 1 1 1 1 ...
## $ weekend alch.  : int  1 1 3 1 2 2 1 1 1 1 ...
## $ health         : int  3 3 3 5 5 5 3 1 1 5 ...
## $ absences       : int  4 2 6 0 0 6 0 2 0 0 ...
## $ Grade          : num  7.33 10.33 12.33 14 12.33 ...

```

`names(e)`

```

## [1] "school"      "sex"          "age"
## [4] "address"     "family size"  "parents cohab."
## [7] "mom's education" "dad's education" "mom's job"
## [10] "dad's job"   "reason"       "guardian"
## [13] "travel"      "study"        "failures"
## [16] "education support" "family support" "paid"
## [19] "activities"  "nursery"      "higher"
## [22] "internet"    "romantic"     "family bond"
## [25] "free time"   "social"       "workday alch."
## [28] "weekend alch." "health"       "absences"
## [31] "Grade"

```

```

x1 <- e[c(31, 1 : 10)]
pairs(x1)

```

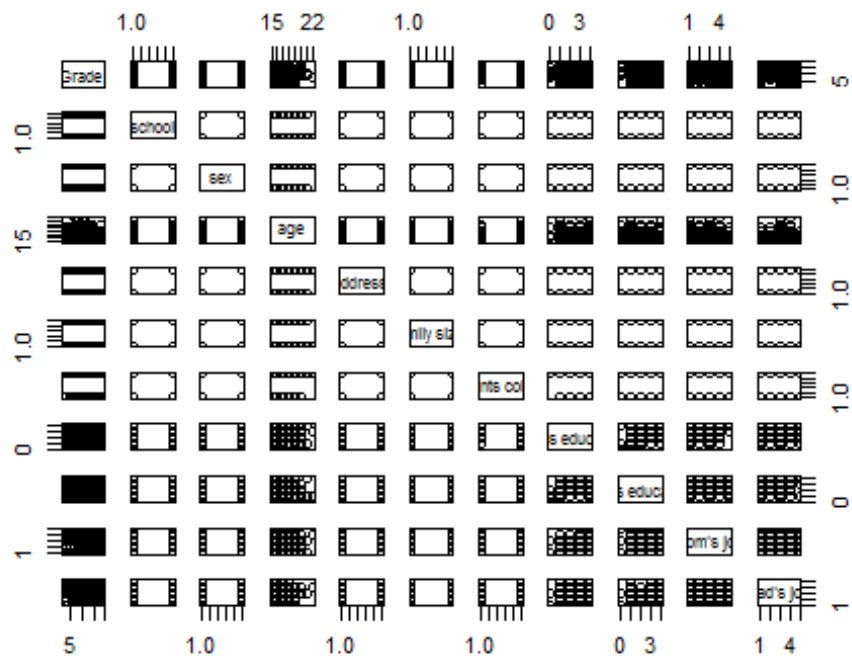


Table 1 : Scatter plot, Grade verses independent predictors 1 through 10.

```
x2 <- e[c(31, 10 : 20)]
pairs(x2)
```

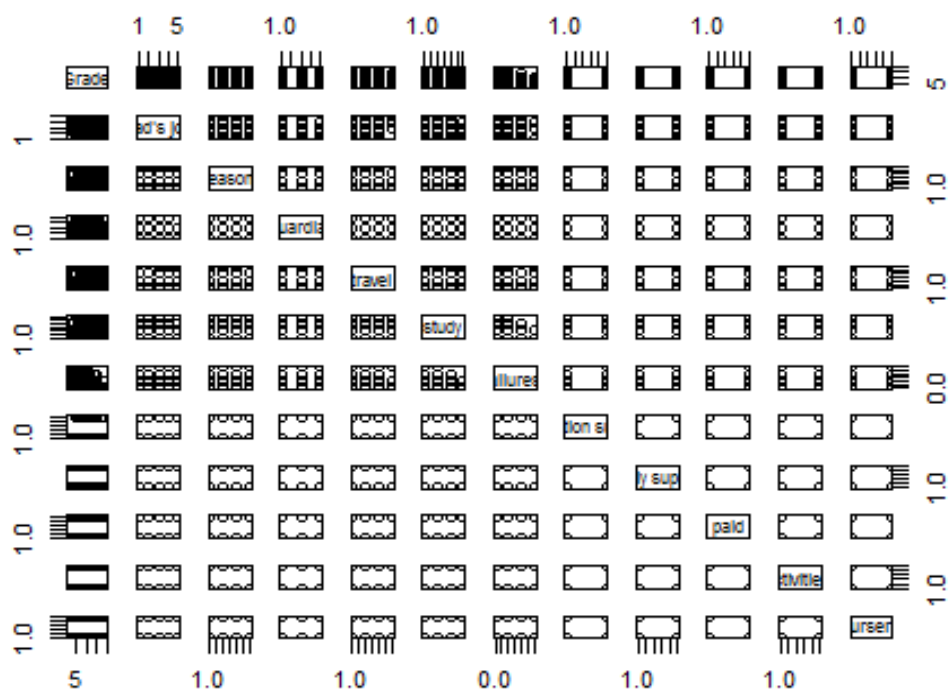


Table 2 : Scatter plot, Grade verses independent predictors 10 through 20.

```
x3 <- e[c(31, 20 : 30)]
pairs(x3)
```

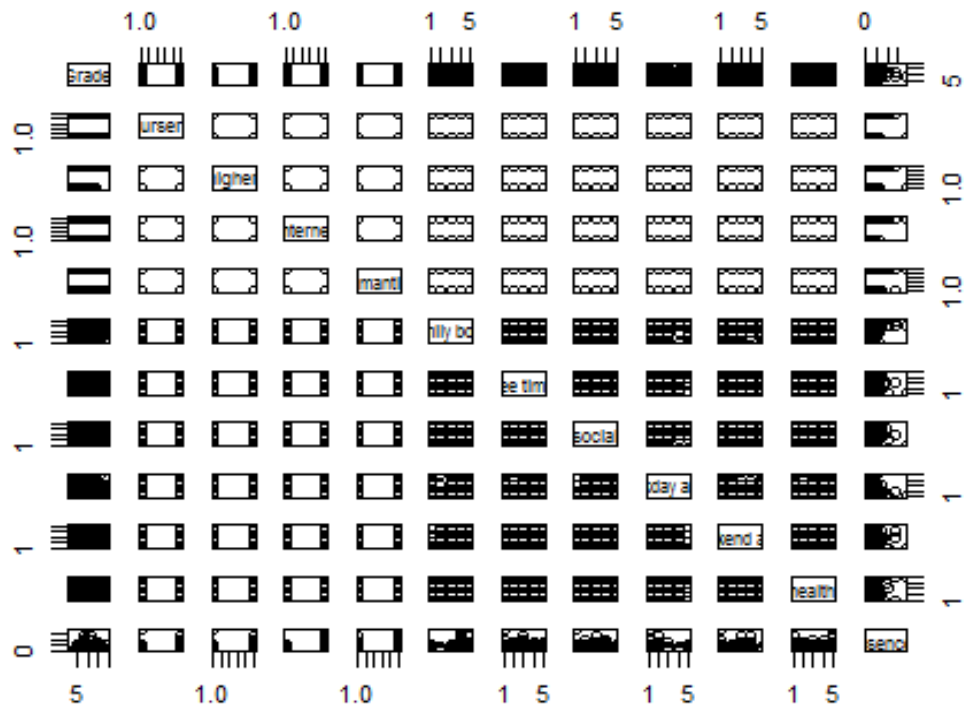


Table 3 : Scatter plot, Grade verses independent predictors 20 through 30.

Model Development

K-fold CV is applied in the fitting of linear models to the training data. Successively, models of lesser complexity are derived (starting with the saturated model), selecting statistically significant predictors that are reported with every model fit. Cross validation indicates a Mean Square Error rate estimate to verify that, in choosing lower complexity models, we are not introducing significant error. Finally, a best-fit model containing significant predictors (showing little difference in MSE from the saturated model), is tested with the clean data to provide an out-of-sample estimate for model performance. The following is a series of progressive model fits performed to find the best possible fit. The saturated model cardinality is 30 variables. 10-fold CV is applied to the MSE estimation of model performance on the test data.

```
names(e)

## [1] "school"          "sex"              "age"
## [4] "address"         "family size"      "parents cohab."
## [7] "mom's education" "dad's education"  "mom's job"
## [10] "dad's job"       "reason"           "guardian"
## [13] "travel"          "study"            "failures"
## [16] "education support" "family support"   "paid"
## [19] "activities"      "nursery"          "higher"
## [22] "internet"        "romantic"         "family bond"
## [25] "free time"       "social"           "workday alch."
## [28] "weekend alch."  "health"           "absences"
## [31] "Grade"

fit <- glm(Grade~., data = e)
MSE1 <- cv.glm(e, fit, K = 10)$delta[1]
summary(fit)

##
## Call:
## glm(formula = Grade ~ ., data = e)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -10.8517  -1.4833   0.1019   1.8281   7.8999
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```

## (Intercept)          9.718585    1.641229    5.922 4.38e-09 ***
## schoolMS            -0.492338    0.235632   -2.089 0.036919 *
## sexM                -0.065729    0.202918   -0.324 0.746068
## age                 0.030970    0.083072    0.373 0.709372
## addressU            0.240670    0.221106    1.088 0.276645
## `family size`LE3    0.369219    0.199709    1.849 0.064783 .
## `parents cohab.`T   0.023677    0.287473    0.082 0.934375
## `mom's education`   0.173160    0.126079    1.373 0.169925
## `dad's education`   0.042871    0.112327    0.382 0.702792
## `mom's job`health   0.934994    0.442614    2.112 0.034896 *
## `mom's job`other   -0.020608    0.262211   -0.079 0.937372
## `mom's job`services 0.524154    0.310235    1.690 0.091426 .
## `mom's job`teacher -0.013337    0.410768   -0.032 0.974105
## `dad's job`health  -0.057531    0.600577   -0.096 0.923704
## `dad's job`other   -0.065647    0.386378   -0.170 0.865120
## `dad's job`services -0.247048    0.404383   -0.611 0.541386
## `dad's job`teacher  1.133663    0.538623    2.105 0.035562 *
## reasonhome          0.133123    0.229150    0.581 0.561410
## reasonother         0.066553    0.311433    0.214 0.830825
## reasonreputation    0.303609    0.239565    1.267 0.205329
## guardianmother     -0.220538    0.219213   -1.006 0.314636
## guardianother       0.217507    0.420273    0.518 0.604896
## travel              -0.094595    0.132621   -0.713 0.475841
## study               0.418159    0.115143    3.632 0.000296 ***
## failures            -1.476144    0.148519   -9.939 < 2e-16 ***
## `education support`yes -1.398765    0.286959   -4.874 1.27e-06 ***
## `family support`yes  -0.273525    0.188290   -1.453 0.146627
## paidyes             -0.768545    0.221702   -3.467 0.000549 ***
## activitiesyes       0.097293    0.181423    0.536 0.591887
## nurseryyes         -0.025260    0.222561   -0.113 0.909661
## higheryes           1.409229    0.341220    4.130 3.93e-05 ***
## internetyes         0.323375    0.233715    1.384 0.166780
## romanticyes        -0.448088    0.188898   -2.372 0.017874 *
## `family bond`       0.102933    0.096962    1.062 0.288680
## `free time`         0.032757    0.093005    0.352 0.724759
## social              -0.218046    0.089048   -2.449 0.014510 *
## `workday alch.`     -0.117328    0.128143   -0.916 0.360092
## `weekend alch.`    -0.008548    0.098360   -0.087 0.930768
## health              -0.156854    0.063972   -2.452 0.014378 *
## absences            -0.016961    0.014979   -1.132 0.257773
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 7.662112)
##
##    Null deviance: 10806.2  on 1043  degrees of freedom
## Residual deviance:  7692.8  on 1004  degrees of freedom
## AIC: 5129.8
##
## Number of Fisher Scoring iterations: 2

```

A report on statistical significance of saturated model coefficients indicates significant ($p < 0.01$) predictors of Grade to be study, failures, education, support, paid, and higher.

MSE for 10-Fold CV of fit of saturated model:

```
MSE1
```

```
## [1] 7.988418
```

A check for multicollinearity by VIF shows negative results, indicating the potential for linear modeling success (conditioned on all GVIF values being less than 5).

```
vif(fit)
```

```
##              GVIF Df GVIF^(1/(2*Df))
## school          1.457487  1      1.207264
## sex             1.378085  1      1.173919
## age             1.444359  1      1.201815
## address         1.322027  1      1.149794
## `family size`   1.125962  1      1.061114
## `parents cohab.` 1.153808  1      1.074155
## `mom's education` 2.738116  1      1.654725
## `dad's education` 2.077963  1      1.441514
## `mom's job`     2.713832  4      1.132916
## `dad's job`     1.890115  4      1.082832
## reason          1.427053  3      1.061060
## guardian        1.472356  2      1.101547
## travel          1.281912  1      1.132215
## study           1.256347  1      1.120869
## failures        1.292695  1      1.136968
## `education support` 1.133132  1      1.064487
## `family support` 1.145957  1      1.070494
## paid            1.113884  1      1.055407
## activities      1.121040  1      1.058792
## nursery         1.080640  1      1.039538
## higher          1.237127  1      1.112262
## internet        1.225436  1      1.106994
## romantic        1.113774  1      1.055355
```

```
## `family bond`      1.115004  1      1.055937
## `free time`        1.252843  1      1.119305
## social              1.433910  1      1.197460
## `workday alch.`    1.857976  1      1.363076
## `weekend alch.`    2.174975  1      1.474780
## health              1.130731  1      1.063359
## absences            1.177898  1      1.085310
```

Second Fit:

A lower complexity model of 5 variables (reported significant), from the saturated model is fitted. All variables included are checked for significance ($p < 0.01$).

```
fit2 <- glm(Grade ~ study + failures + `education support` + paid +
higher, data = e)
MSE2 <- cv.glm(e, fit2, K = 10)$delta[1]
summary(fit2)

##
## Call:
## glm(formula = Grade ~ study + failures + `education support` +
##     paid + higher, data = e)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -10.4849  -1.6574   0.0707   1.9185   7.9790
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.2697     0.3692  25.108  < 2e-16 ***
## study           0.4942     0.1091   4.530 6.57e-06 ***
## failures       -1.6364     0.1415 -11.566  < 2e-16 ***
## `education support`yes -1.2480     0.2796  -4.463 8.96e-06 ***
## paidyes        -0.6331     0.2193  -2.887  0.00397 **
## higheryes       1.8935     0.3368   5.622 2.42e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 8.158794)
##
## Null deviance: 10806.2 on 1043 degrees of freedom
## Residual deviance: 8468.8 on 1038 degrees of freedom
## AIC: 5162.2
##
## Number of Fisher Scoring iterations: 2
```

10-Fold CV MSE estimate of Model II:

MSE2

```
## [1] 8.222127
```

This model shows negligible difference in MSE from the saturated model and contains 26 fewer predictors, thus indicating potential for out-of-sample performance.

```
summary(fit2)
```

```
##
## Call:
## glm(formula = Grade ~ study + failures + `education support` +
##      paid + higher, data = e)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -10.4849  -1.6574   0.0707   1.9185   7.9790
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      9.2697     0.3692  25.108 < 2e-16 ***
## study             0.4942     0.1091   4.530 6.57e-06 ***
## failures         -1.6364     0.1415 -11.566 < 2e-16 ***
## `education support`yes -1.2480     0.2796  -4.463 8.96e-06 ***
## paidyes          -0.6331     0.2193  -2.887 0.00397 **
## higheryes         1.8935     0.3368   5.622 2.42e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 8.158794)
##
##      Null deviance: 10806.2  on 1043  degrees of freedom
## Residual deviance:  8468.8  on 1038  degrees of freedom
## AIC: 5162.2
##
## Number of Fisher Scoring iterations: 2

par(mfrow = c(3, 2))

plot(Grade, study)
plot(Grade, failures)
plot(Grade, `education support`, yaxt='n')
```

```

axis(2, labels = c("false","true"), at = c(1, 2))
plot(Grade, paid, yaxt='n')
axis(2, labels = c("false","true"), at = c(1, 2))
plot(Grade, higher, yaxt='n')
axis(2, labels = c("false","true"), at = c(1, 2))
mtext("Significant Factor Plots", side = 3, line = -3, outer = TRUE)

```

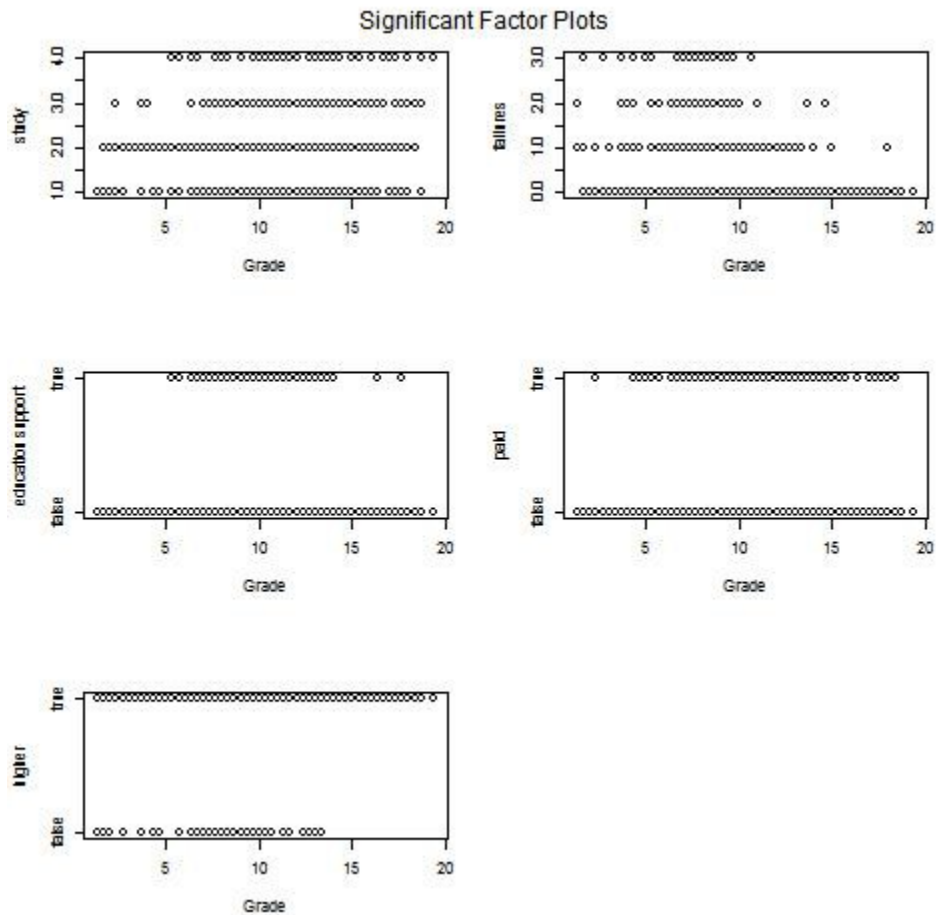


Fig 2: Scatter-plots of Grade verses significant discovered factors.

Results

This results of this work should be seen as a starting point for more advanced studies of success prediction in general education. They may only hold significance for the originating educational department. It appears strictly domain specific, general claims to any general predictive success of any derived models is not generally indicated. This view is surmised from the collected data and available documentation. The intent interpreted is to find a model of specific factors relevant to learning success that are shared between Mathematics and Portuguese, discussed in identical terms: identical variables are chosen for both data sets as collected from student surveys. This indicates an implicit assumption of the study that a uniform learning measure exists between mathematics and language. This assumption appears to be latent in the study variables, as chosen factors are more generally living-condition or non-subject specific. It is apparent that for the chosen causal-factors, the data is not sufficiently diverse in academic types to address the generality of their scope. In addition, this study has a limited data quantity relative to the number of the predictors. All inclusions and exclusions of predictors suggested by this regression model should be seen as restricted from any general claims predictive power (in a broader or differing range or academic subjects).

Conclusion

This work could be seen a good use of resources for determining how best to design future studies, specifically, what questions to exclude from study surveys in any following work. Results indicate a reduction in the set of explanatory variables by a factor of 5 in predictive modeling. Certain difficulties are present in interpreting results from the data documentation, indicating restricted application to the listed education categories.