STAT 306 Group project

December 7th, 2020

Group 20

Henry Tian, Julie Wu, Shijie Li, Tanay Kumar

1 Introduction

1.1 Background

Education plays a large role in a community's economic progress. Subjects such as Mathematics and English form the fundamentals for other school subjects, like chemistry or history. The grades received in these courses in secondary school may be very influential for someone's future. Secondary school grades are evaluated using standardized testing issued by their teacher. Most, if not all, countries use this form of evaluating children.

However, there may be external factors that affect a student's performance in school, such as the amount of time spent studying, or if a tutor was hired to assist the student in learning. Other factors that are unrelated to a student's study habits may also play a part, such as their age, sex, or their parent's education.

Statistically modelling student performance may be important for both teachers and students, as it may help teachers adapt to students' needs and provide them more useful support, and allow the students to realize what needs to change. In this paper, we will explore different variables and which ones may affect the grades of secondary school students using the data set provided (P. Cortez, 2008). We will predict a model for student performance using these variables.

1.2 Data

In Portugal, secondary school education is three years long, given nine years of basic education, and potentially followed by higher education. Most of the students attend public schools where many courses require proficiency in subjects such as the Portuguese language and mathematics. A 20-point grading scale is used, where 0 is the lowest and 20 is the best grade achievable. A score between 0 and 9 is considered a failing grade.

The data used in this study were collected in the 2005- 2006 school year from two public schools in Portugal, Gabriel Pereira (GP) and Moushinho da Silveira (MS). For Mathematics, 395 student responses were compiled, and 649 responses were compiled from the Portuguese language classes. The data was collected using questionaries and school reports and includes 33 different variables:

Table 1. The student-related variables

Variable	**Description**	**Value**
school(x_1)	School attended	GP (Gabriel Pereira), MS (Moushinoho da Silveira)
sex(x_2)	Student's sex	M (male), F (female)
age(x_3)	Student's age	15 - 22
address(x_4)	Student's home address	U (urban), R (rural)
famsize(x_5)	Family size	LE3 (less than or equal to 3), GT3 (greater than 3)
Pstatus(x_6)	Parent's cohabitation status	T (living together), A (living apart)
Medu(x_7)	Mother's education	0 (none), 1 (primary education; 4th grade), 2 (5th - 9th grade), 3 (secondary education) 4 (higher education)
Fedu(x_8)	Father's education	0 (none), 1 (primary education; 4th grade), 2 (5th - 9th grade), 3 (secondary education) 4 (higher education)
Mjob(x_9)	Mother's occupation	Teacher, healthcare-related, civil services, at home, other
Fjob(x_{10})	Father's occupation	Teacher, healthcare-related, civil services, at home, other
reason(x_{11})	Reason to attend school	Close to home, School reputation, course preference, or other
guardian(x_{12})	Student's guardian	Mother, Father, Other
traveltime(x_{13})	home to school travel time	1 (< 15 mins), 2 (15 - 30 mins), 3 (30 mins - 1 hour), 4 (>1 hour)
studytime(x_{14})	Weekly study time	1 (< 2 hours), 2 (2 - 5 hours), 3 (5 - 10 hours), 4 (> 10 hours)
failures(x_{15})	Number of past class failures	n if $1 \leq n < 3$; 4 otherwise
schoolsup(x_{16})	Extra educations school support	Yes, no
famsup(x_{17})	Family educational support	Yes, no
paid(x_{18})	Extra paid classes/tutor	Yes, no
activities(x_{19})	Extra-curricular activities	Yes. no

nursery(x_{20})	Attended nursery school	Yes, no
higher(x_{21})	Desire to take higher education	Yes, no
internet(x_{22})	Internet access at home	Yes, no
romantic(x_{23})	In a romantic relationship	Yes, no
famrel(x_{24})	Quality of family relationships	1 (very bad) - 5 (very good)
freetime(x_{25})	Free time after school	1 (very low) - 5 (very high)
goout(x_{26})	Going out with friends	1 (very low) - 5 (very high)
Dalc(x_{27})	Workday alcohol consumption	1 (very low) - 5 (very high)
Walc(x_{28})	Weekend alcohol consumption	1 (very low) - 5 (very high)
health(x_{29})	Current health status	1 (very bad) - 5 (very good)
absences(x_{30})	Number of school absences	0 - 93
G1(y_1)	First-period grade	0 - 20
G2(y_1)	Second-period grade	0 - 20
G3(y_1)	Final grade (outcome)	0 - 20

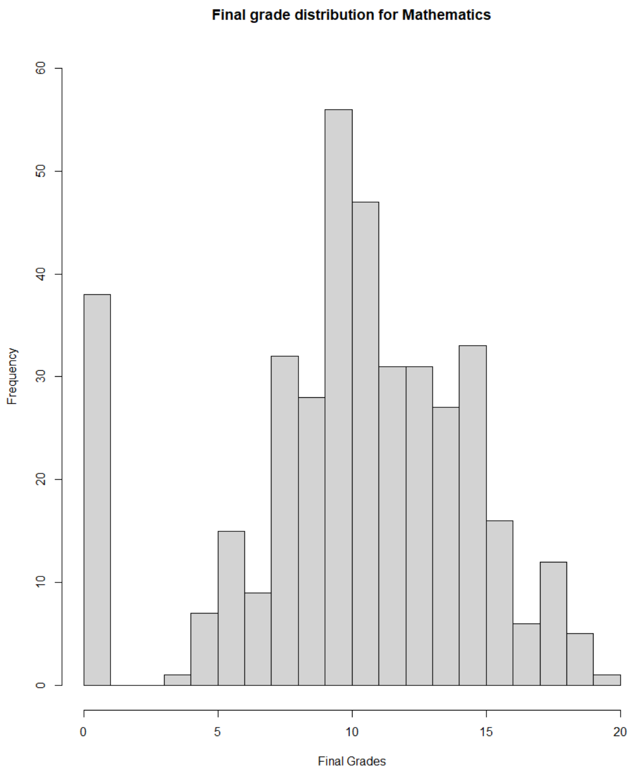


Figure 1. Histogram distribution of grades in Mathematics. A score of 10 and above is a passing grade.

Our goal is to predict a student's grades and to identify the significant variables that affect success in secondary schools. An analysis will be performed over the best models in order to identify the most important variables. We will be focusing on the data provided for students in the mathematics class, a total of 395 student responses.

2 Analysis

2.1 Preliminary Analysis

As our data contains 30 explanatory variables and the main response variable G3. Two possible predictor variables, G1 and G2, are also used. We first needed to discover which variables significantly impact our responses. We used R to compute our data, which is a programming language and free software environment for statistical computing. To begin, we must convert several variables into numeric values. The variables changed are listed as such:

Table 2. Updated numeric values for some student-related variables

Variable	**Value**	**Changed Value**
school(x_1)	GP (Gabriel Pereira), MS (Moushinoho da Silveira)	0, 1
sex(x_2)	M (male), F (female)	0, 1
address(x_3)	U (urban), R (rural)	0, 1
famsize(x_4)	LE3 (less than or equal to 3), GT3 (greater than 3)	0, 1
Pstatus(x_6)	T (living together), A (living apart)	0, 1
Mjob(x_9)	Teacher, healthcare-related, civil services, at home, other	0, 1, 2, 3, 4
Fjob(x_{10})	Teacher, healthcare-related, civil services, at home, other	0, 1, 2, 3, 4
reason(x_{11})	Close to home, School reputation, course preference, or other	0, 1, 2, 3
guardian(x_{12})	Mother, Father, Other	0, 1, 2
schoolsup(x_{16})	Yes, no	1, 0
famsup(x_{17})	Yes, no	1, 0
paid(x_{18})	Yes, no	1, 0
activities(x_{19})	Yes. no	1, 0
nursery(x_{20})	Yes, no	1, 0

higher(x_{21})	Yes, no	1, 0
internet(x_{22})	Yes, no	1, 0
romantic(x_{23})	Yes, no	1, 0

All other variables were already numeric and are unchanged in modelling.

2.2 Models

We begin with a full model to analyze our first predictor variable, G1 (appendix 1). This model shows us eight variables that are significant. The model had an R^2 value of 0.287, meaning 28.7 percent of the G1 grades is explained by the explanatory variables, with an adjusted R^2 of 0.2303. We will first find the most significant variables using an exhaustive search. Using R, we use the *regsubsets* function provided by the *leaps* library. By subsetting the 30 variables (excluding G1, G2, and G3) in relation to G1, we get the data shown in the table below:

Table 3: Summary of exhaustive search for G1

Size	**Variables**	**Adjusted R^2**	**Mallow's Statistic (C_p)**
1	failures	0.1236002	56.499320
2	failures, schoolsup	0.1668630	35.326400
3	Fjob, failures, schoolsup	0.1925002	23.219908
4	Fjob, failures, schoolsup, goout	0.2021774	19.267216
5	sex, Fjob, failures, studytime, schoolsup	0.2131173	14.701445
6	sex, Fjob, failures, studytime, schoolsup, goout	0.2234439	10.473302
7	Medu, Fjob, failures, studytime, schoolsup, famsup, goout	0.2303718	7.980893
8	sex, Medu, Fjob, failures, studytime, schoolsup, famsup, goout	0.2383655	4.971976
	...	...	...
11	sex, Mjob, Fjob, studytime. failures, schoolsup, famsup, higher, freetime, goout, health	0.2505448	1.942617
14	sex, famsize, Fedu, Mjob, Fjob, studytime, failures, schoolsup, famsup, higher, freetime, goout, Dalc, health	0.2537462	3.440824

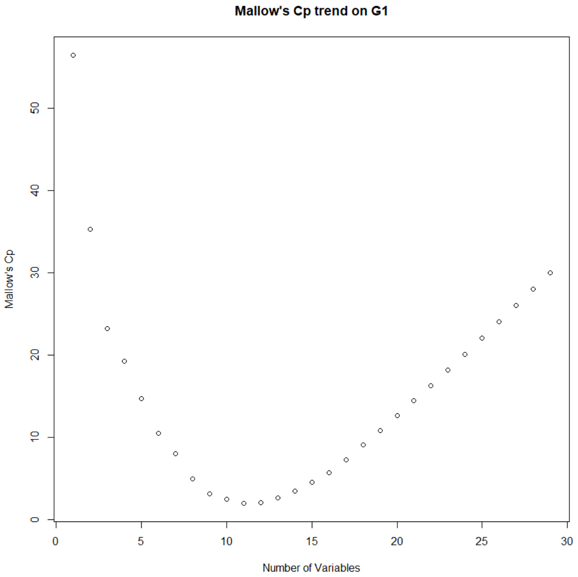


Figure 2: Mallow's C_p trend for G1 models

Using table 3, we find that the model with the highest adjusted-R^2 has 14 variables, and the model with the lowest Mallow's statistic has 11 variables. We decide to see the significance between both the 11-variable model and the 14-variable model (appendix 2 and appendix 3).

Several variables in both models are not significant at the 0.05 level, including *Mjob, freetime*, and *health* for the model with 11 variables, and *famsize, Fedu*, and *Dalc* with the three former variables for the 14-variable model. Still using our exhaustive search data, we find that the model where all variables used have the significance of at least the 0.05 level and has a high adjusted R^2 and small C_p, is the 8-variable model (appendix 5). Although there is a decrease in the R^2 from the full model to the reduced model, there is an increase in the adjusted R^2. The estimated model given by the exhaustive search is:

$$y_1 = 5.9408+0.7057x_2+0.3275x_7 - 0.3602x_{10}+0.5845x_{14} -1.3324x_{15}+2.0050x_{16}+0.7794x_{17} - 0.3615x_{26}$$

Repeating the same steps we performed on G1 for G2, we get the full model (see appendix 5). This model shows nine variables that are significant. It has an R^2 of 0.2653, meaning 26.53 percent of G2 grades are explained by all 30 explanatory variables, with an adjusted R^2 of 0.2069. Using *regsubsets* again, we get the data shown below:

Table 4: Summary of exhaustive search for G2

Size	**Variables**	**Adjusted R^2**	**Mallow's Statistic (C_p)**
1	failures	0.1244395	42.86089

2	Medu, failures	0.1405601	35.78911
3	Medu, failures, goout	0.1560765	29.05583
4	Medu, failures, schoolsup, goout	0.1677986	24.22749
5	Medu, failures, schoolsup, romantic, goout	0.1755284	21.38690
6	Medu, traveltime, failures, schoolsup, romantic, goout	0.1828195	18.78037
7	sex, address, Medu, studytime, failures, schoolsup, goout	0.1909643	15.77573
8	sex, address, Medu, studytime, failures, schoolsup, romatic, goout	0.1976288	13.51204
9	sex, address, Medu, studytime, failures, schoolsup, famsup, romantic, goout	0.2042350	11.293430
	...	...	...
14	sex, famsize, Medu, Fjob, traveltime, failures, schoolsup, famsup, higher, internet, romantic, goout, health	0.2242386	6.692254
16	sex, age, famsize, Medu, Fjob, guardian, traveltime, failures, schoolsup, famsup, higher, internet, romantic, goout, health	0.2261659	7.817435

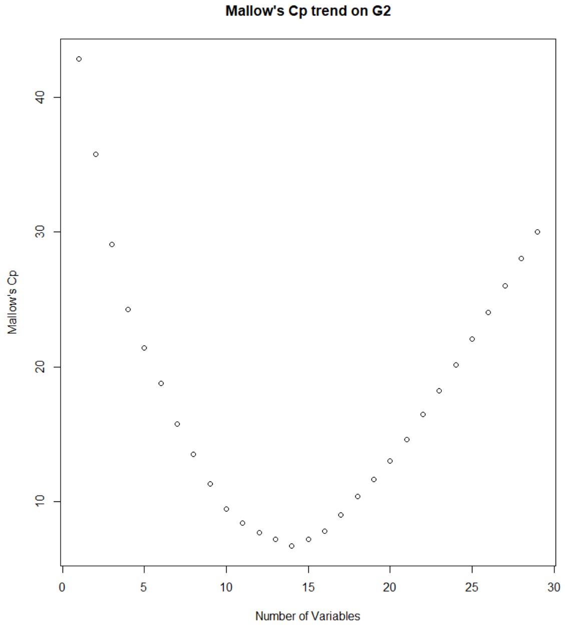


Figure 3: Mallow's C_p trend for G2 models

Using table 4, we find that the model with the highest adjusted R^2 has 16 variables, and the model with the lowest Mallow's statistic has 14 variables. We decide to see the significance between both the 14-variable model and the 16-variable model (appendix 6 and appendix 7).

Several variables in both models are not significant at the 0.05 level. The ones that are shared and significant are Medu, Fjob, traveltime, failures, schoolsup, and goout. We decide to check the model where all variables are significant at a 0.05 level. This turned out to be a 9-variable model (appendix 8). Although there is a decrease in the R^2 from the full model and the highest R^2 model, there is an increase in the adjusted R^2 compared to the full model. The estimated model given by the exhaustive search is:

$$y_2 = 5.6390+0.7963x_2-0.9537x_3 + 0.5041x_7+0.5694x_{14} -1.3827x_{15}+1.2129x_{16}+0.7398x_{17} +0.7568x_{23} - 0.5055x_{26}$$

We can see that first-period grade and the second-period grade both shared *sex, medu, studytime, failures, schoolsup, famsup* and *goout*. The only variable that is present in prediction of G1 and not in G2 is *Fjob*. And similarly, *address* and *romantic* are present in the prediction of G2 and not in the prediction of G1.

Finally, for G3, we perform a similar analysis. The full model (appendix 9) is shown, but the exhaustive search will be omitted in this paper. Using *regsubsets* again, look for the model with the highest adjusted R^2, with the lowest C_p, and where all variables are significant at a 0.05 level.

The highest adjusted R^2 (0.8349497) belongs to the model with 12 variables, and the lowest C_p (0.3967861) belongs to the model with 8 variables. The last model with all significant variables is the model containing 4 variables: *famrel, absences, G1*, and *G2* (appendix 10). It has an adjusted R^2 of 0.8289151 and a C_p of 7.9987191. The resulting equation is:

$$y_3 = -3.40923+0.34252x_{24}+0.03806x_{30} - 0.14183y_1+0.99953y_2$$

3 Conclusion

We created our models using an exhaustive search function. It gave us the possible combinations of variables without interaction. Using the adjusted R^2, Mallow's statistic values, and the number of variables deemed significant, we determined which model would be the most appropriate for the selected response variable. For the first-period grades, the model is:

$$y_1 = 5.9408+0.7057x_2+0.3275x_7 - 0.3602x_{10}+0.5845x_{14} -1.3324x_{15}+2.0050x_{16}+0.7794x_{17} - 0.3615x_{26}$$

For the second-period grades, the model is:

$$y_2 = 5.6390+0.7963x_2-0.9537x_3 + 0.5041x_7+0.5694x_{14} -1.3827x_{15}+1.2129x_{16}+0.7398x_{17} +0.7568x_{23} - 0.5055x_{26}$$

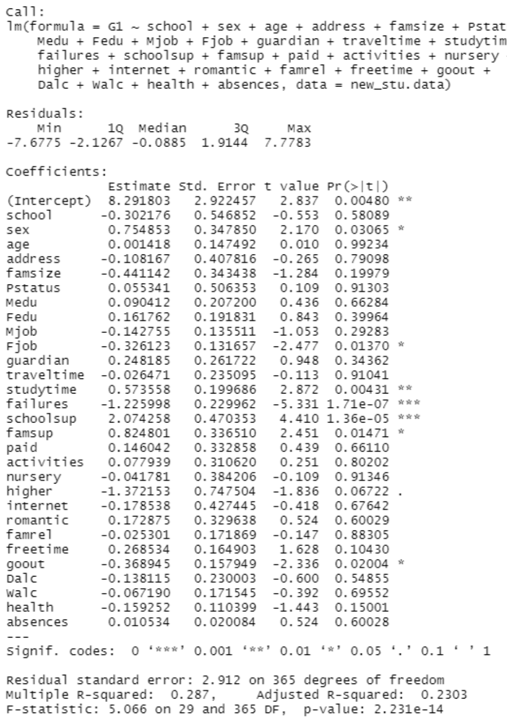
For the final grades, the model is:

$$y_3 = -3.40923+0.34252x_{24}+0.03806x_{30} - 0.14183y_1+0.99953y_2$$

All variables used in these models have a significant effect on a student's final grades. Those variables are *sex, medu, studytime, failures, schoolsup, famsup, goout, Fjob, address, romantic, famrel and absences.* The final model shows that the final grades are also dependent on grades for first-period and second-period.

However, our models are far from perfect. We selected our best model by finding a model where all variables used are significant at a 0.05 level and had a relatively high adjusted R^2 and a relatively low Mallow's statistic. Other models with better numbers than the model we used may exist. Only four students performed this data analysis, and an interaction model was not used as that would require much more time. More research on the interactions between different variables may lead to better-fitted models.

Appendixes



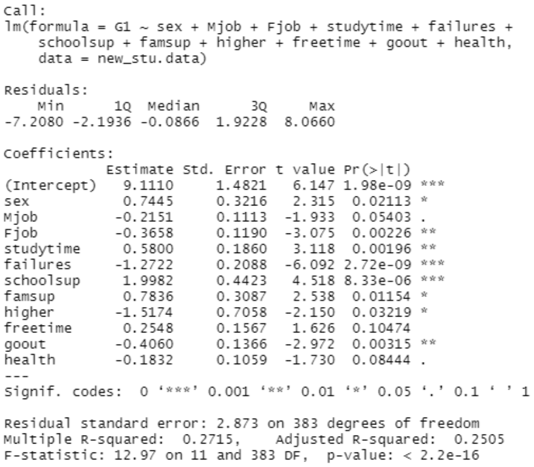
```
Call:
lm(formula = G1 ~ school + sex + age + address + famsize + Pstatus + 
    Medu + Fedu + Mjob + Fjob + guardian + traveltime + studytime + 
    failures + schoolsup + famsup + paid + activities + nursery + 
    higher + internet + romantic + famrel + freetime + goout + 
    Dalc + Walc + health + absences, data = new_stu.data)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.6775 -2.1267 -0.0885  1.9144  7.7783 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.291803   2.922457   2.837  0.00480 ** 
school      -0.302176   0.546852  -0.553  0.58089    
sex          0.754853   0.347850   2.170  0.03065 *  
age          0.001418   0.147492   0.010  0.99234    
address     -0.108167   0.407816  -0.265  0.79098    
famsize     -0.441142   0.343438  -1.284  0.19979    
Pstatus      0.055341   0.506353   0.109  0.91303    
Medu         0.090412   0.207200   0.436  0.66284    
Fedu         0.161762   0.191831   0.843  0.39964    
Mjob        -0.142755   0.135511  -1.053  0.29283    
Fjob        -0.326123   0.131657  -2.477  0.01370 *  
guardian     0.248185   0.261722   0.948  0.34362    
traveltime  -0.026471   0.235095  -0.113  0.91041    
studytime    0.573558   0.199686   2.872  0.00431 ** 
failures    -1.225998   0.229962  -5.331 1.71e-07 ***
schoolsup    2.074258   0.470353   4.410 1.36e-05 ***
famsup       0.824801   0.336510   2.451  0.01471 *  
paid         0.146042   0.332858   0.439  0.66110    
activities   0.077939   0.310620   0.251  0.80202    
nursery     -0.041781   0.384206  -0.109  0.91346    
higher      -1.372153   0.747504  -1.836  0.06722 .  
internet    -0.178538   0.427445  -0.418  0.67642    
romantic     0.172875   0.329638   0.524  0.60029    
famrel      -0.025301   0.171869  -0.147  0.88305    
freetime     0.268534   0.164903   1.628  0.10430    
goout       -0.368945   0.157949  -2.336  0.02004 *  
Dalc        -0.138115   0.230003  -0.600  0.54855    
Walc        -0.067190   0.171545  -0.392  0.69552    
health      -0.159252   0.110399  -1.443  0.15001    
absences     0.010534   0.020084   0.524  0.60028    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.912 on 365 degrees of freedom
Multiple R-squared:  0.287,	Adjusted R-squared:  0.2303 
F-statistic: 5.066 on 29 and 365 DF,  p-value: 2.231e-14
```

Appendix 1: Full model



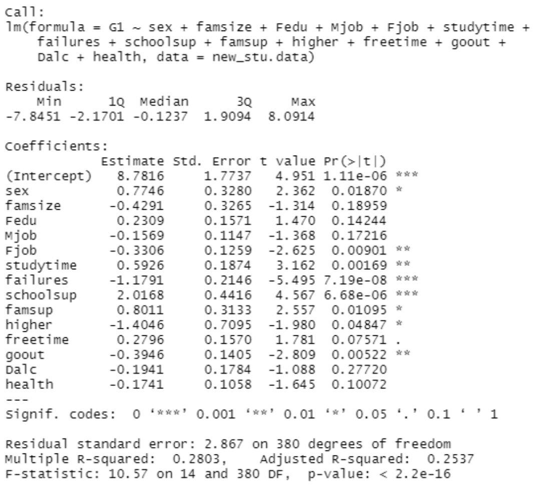
```
Call:
lm(formula = G1 ~ sex + Mjob + Fjob + studytime + failures + 
    schoolsup + famsup + higher + freetime + goout + health, 
    data = new_stu.data)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.2080 -2.1936 -0.0866  1.9228  8.0660 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   9.1110     1.4821   6.147 1.98e-09 ***
sex           0.7445     0.3216   2.315  0.02113 *  
Mjob         -0.2151     0.1113  -1.933  0.05403 .  
Fjob         -0.3658     0.1190  -3.075  0.00226 ** 
studytime     0.5800     0.1860   3.118  0.00196 ** 
failures     -1.2722     0.2088  -6.092 2.72e-09 ***
schoolsup     1.9982     0.4423   4.518 8.33e-06 ***
famsup        0.7836     0.3087   2.538  0.01154 *  
higher       -1.5174     0.7058  -2.150  0.03219 *  
freetime      0.2548     0.1567   1.626  0.10474    
goout        -0.4060     0.1366  -2.972  0.00315 ** 
health       -0.1832     0.1059  -1.730  0.08444 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.873 on 383 degrees of freedom
Multiple R-squared:  0.2715,	Adjusted R-squared:  0.2505 
F-statistic: 12.97 on 11 and 383 DF,  p-value: < 2.2e-16
```

Appendix 2: G1 11-variable model



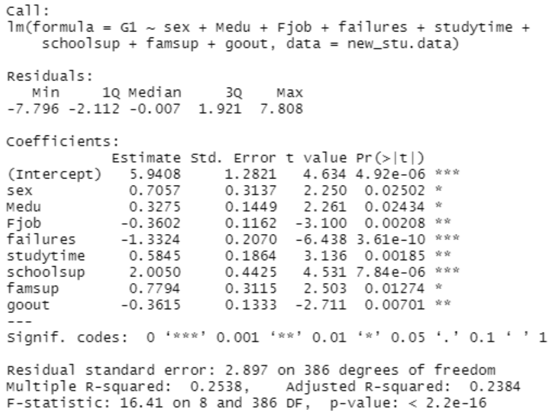
```
Call:
lm(formula = G1 ~ sex + famsize + Fedu + Mjob + Fjob + studytime +
    failures + schoolsup + famsup + higher + freetime + goout +
    Dalc + health, data = new_stu.data)

Residuals:
    Min      1Q  Median      3Q     Max
-7.8451 -2.1701 -0.1237  1.9094  8.0914

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.7816     1.7737   4.951 1.11e-06 ***
sex           0.7746     0.3280   2.362  0.01870 *
famsize      -0.4291     0.3265  -1.314  0.18959
Fedu          0.2309     0.1571   1.470  0.14244
Mjob         -0.1569     0.1147  -1.368  0.17216
Fjob         -0.3306     0.1259  -2.625  0.00901 **
studytime     0.5926     0.1874   3.162  0.00169 **
failures     -1.1791     0.2146  -5.495 7.19e-08 ***
schoolsup     2.0168     0.4416   4.567 6.68e-06 ***
famsup        0.8011     0.3133   2.557  0.01095 *
higher       -1.4046     0.7095  -1.980  0.04847 *
freetime      0.2796     0.1570   1.781  0.07571 .
goout        -0.3946     0.1405  -2.809  0.00522 **
Dalc         -0.1941     0.1784  -1.088  0.27720
health       -0.1741     0.1058  -1.645  0.10072
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.867 on 380 degrees of freedom
Multiple R-squared:  0.2803,    Adjusted R-squared:  0.2537
F-statistic: 10.57 on 14 and 380 DF,  p-value: < 2.2e-16
```

Appendix 3: G1 14-variable model



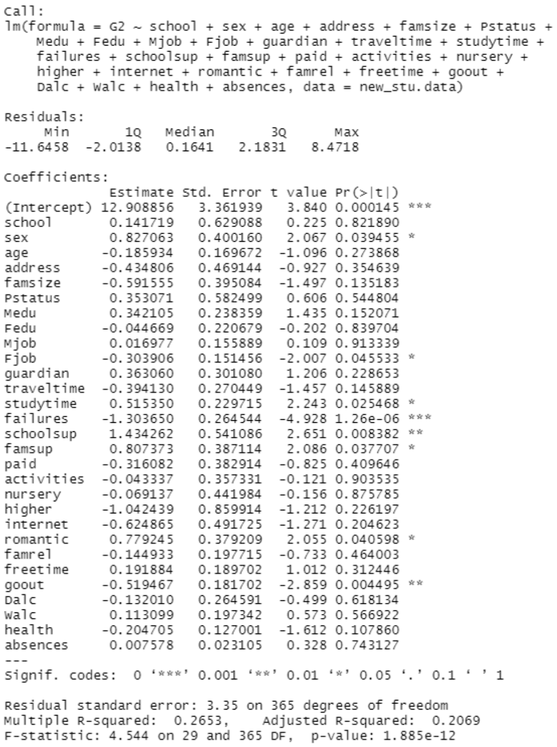
```
Call:
lm(formula = G1 ~ sex + Medu + Fjob + failures + studytime +
    schoolsup + famsup + goout, data = new_stu.data)

Residuals:
   Min     1Q Median     3Q    Max
-7.796 -2.112 -0.007  1.921  7.808

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.9408     1.2821   4.634 4.92e-06 ***
sex           0.7057     0.3137   2.250  0.02502 *
Medu          0.3275     0.1449   2.261  0.02434 *
Fjob         -0.3602     0.1162  -3.100  0.00208 **
failures     -1.3324     0.2070  -6.438 3.61e-10 ***
studytime     0.5845     0.1864   3.136  0.00185 **
schoolsup     2.0050     0.4425   4.531 7.84e-06 ***
famsup        0.7794     0.3115   2.503  0.01274 *
goout        -0.3615     0.1333  -2.711  0.00701 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.897 on 386 degrees of freedom
Multiple R-squared:  0.2538,    Adjusted R-squared:  0.2384
F-statistic: 16.41 on 8 and 386 DF,  p-value: < 2.2e-16
```

Appendix 4: G1 8-variable model



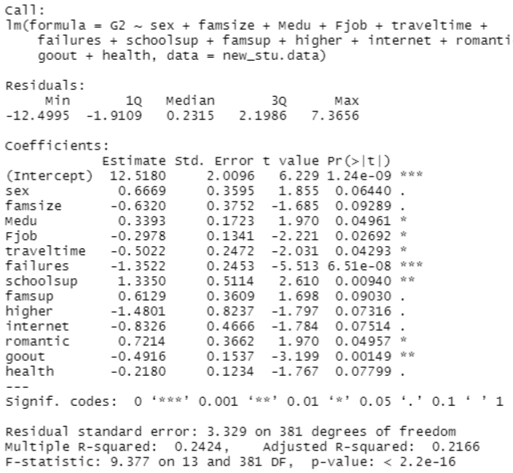
```
Call:
lm(formula = G2 ~ school + sex + age + address + famsize + Pstatus +
    Medu + Fedu + Mjob + Fjob + guardian + traveltime + studytime +
    failures + schoolsup + famsup + paid + activities + nursery +
    higher + internet + romantic + famrel + freetime + goout +
    Dalc + Walc + health + absences, data = new_stu.data)

Residuals:
     Min       1Q   Median       3Q      Max
-11.6458  -2.0138   0.1641   2.1831   8.4718

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.908856   3.361939   3.840 0.000145 ***
school       0.141719   0.629088   0.225 0.821890
sex          0.827063   0.400160   2.067 0.039455 *
age         -0.185934   0.169672  -1.096 0.273868
address     -0.434806   0.469144  -0.927 0.354639
famsize     -0.591555   0.395084  -1.497 0.135183
Pstatus      0.353071   0.582499   0.606 0.544804
Medu         0.342105   0.238359   1.435 0.152071
Fedu        -0.044669   0.220679  -0.202 0.839704
Mjob         0.016977   0.155889   0.109 0.913339
Fjob        -0.303906   0.151456  -2.007 0.045533 *
guardian     0.363060   0.301080   1.206 0.228653
traveltime  -0.394130   0.270449  -1.457 0.145889
studytime    0.515350   0.229715   2.243 0.025468 *
failures    -1.303650   0.264544  -4.928 1.26e-06 ***
schoolsup    1.434262   0.541086   2.651 0.008382 **
famsup       0.807373   0.387114   2.086 0.037707 *
paid        -0.316082   0.382914  -0.825 0.409646
activities  -0.043337   0.357331  -0.121 0.903535
nursery     -0.069137   0.441984  -0.156 0.875785
higher      -1.042439   0.859914  -1.212 0.226197
internet    -0.624865   0.491725  -1.271 0.204623
romantic     0.779245   0.379209   2.055 0.040598 *
famrel      -0.144933   0.197715  -0.733 0.464003
freetime     0.191884   0.189702   1.012 0.312446
goout       -0.519467   0.181702  -2.859 0.004495 **
Dalc        -0.132010   0.264591  -0.499 0.618134
Walc         0.113099   0.197342   0.573 0.566922
health      -0.204705   0.127001  -1.612 0.107860
absences     0.007578   0.023105   0.328 0.743127
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.35 on 365 degrees of freedom
Multiple R-squared:  0.2653,    Adjusted R-squared:  0.2069
F-statistic: 4.544 on 29 and 365 DF,  p-value: 1.885e-12
```

Appendix 5: G2 full model



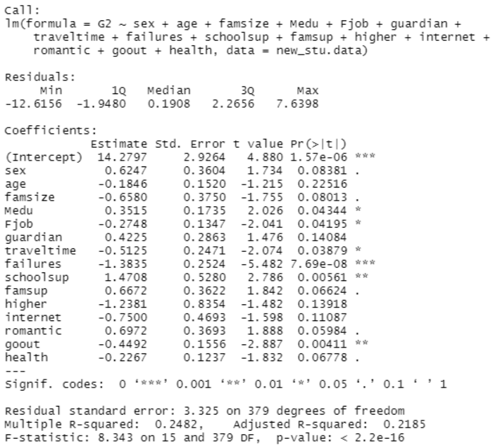
```
Call:
lm(formula = G2 ~ sex + famsize + Medu + Fjob + traveltime +
    failures + schoolsup + famsup + higher + internet + romantic +
    goout + health, data = new_stu.data)

Residuals:
     Min       1Q   Median       3Q      Max
-12.4995  -1.9109   0.2315   2.1986   7.3656

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.5180     2.0096   6.229 1.24e-09 ***
sex           0.6669     0.3595   1.855  0.06440 .
famsize      -0.6320     0.3752  -1.685  0.09289 .
Medu          0.3393     0.1723   1.970  0.04961 *
Fjob         -0.2978     0.1341  -2.221  0.02692 *
traveltime   -0.5022     0.2472  -2.031  0.04293 *
failures     -1.3522     0.2453  -5.513 6.51e-08 ***
schoolsup     1.3350     0.5114   2.610  0.00940 **
famsup        0.6129     0.3609   1.698  0.09030 .
higher       -1.4801     0.8237  -1.797  0.07316 .
internet     -0.8326     0.4666  -1.784  0.07514 .
romantic      0.7214     0.3662   1.970  0.04957 *
goout        -0.4916     0.1537  -3.199  0.00149 **
health       -0.2180     0.1234  -1.767  0.07799 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.329 on 381 degrees of freedom
Multiple R-squared:  0.2424,    Adjusted R-squared:  0.2166
F-statistic: 9.377 on 13 and 381 DF,  p-value: < 2.2e-16
```

Appendix 6: G2 14-variable model



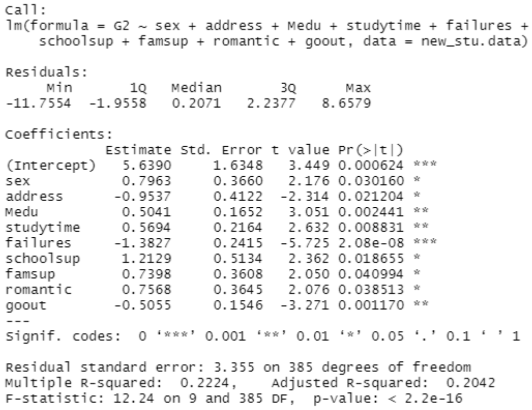
```
Call:
lm(formula = G2 ~ sex + age + famsize + Medu + Fjob + guardian +
    traveltime + failures + schoolsup + famsup + higher + internet +
    romantic + goout + health, data = new_stu.data)

Residuals:
     Min       1Q   Median       3Q      Max
-12.6156  -1.9480   0.1908   2.2656   7.6398

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.2797     2.9264   4.880 1.57e-06 ***
sex           0.6247     0.3604   1.734  0.08381 .
age          -0.1846     0.1520  -1.215  0.22516
famsize      -0.6580     0.3750  -1.755  0.08013 .
Medu          0.3515     0.1735   2.026  0.04344 *
Fjob         -0.2748     0.1347  -2.041  0.04195 *
guardian      0.4225     0.2863   1.476  0.14084
traveltime   -0.5125     0.2471  -2.074  0.03879 *
failures     -1.3835     0.2524  -5.482 7.69e-08 ***
schoolsup     1.4708     0.5280   2.786  0.00561 **
famsup        0.6672     0.3622   1.842  0.06624 .
higher       -1.2381     0.8354  -1.482  0.13918
internet     -0.7500     0.4693  -1.598  0.11087
romantic      0.6972     0.3693   1.888  0.05984 .
goout        -0.4492     0.1556  -2.887  0.00411 **
health       -0.2267     0.1237  -1.832  0.06778 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.325 on 379 degrees of freedom
Multiple R-squared:  0.2482,    Adjusted R-squared:  0.2185
F-statistic: 8.343 on 15 and 379 DF,  p-value: < 2.2e-16
```

Appendix 7: G2 16-variable model



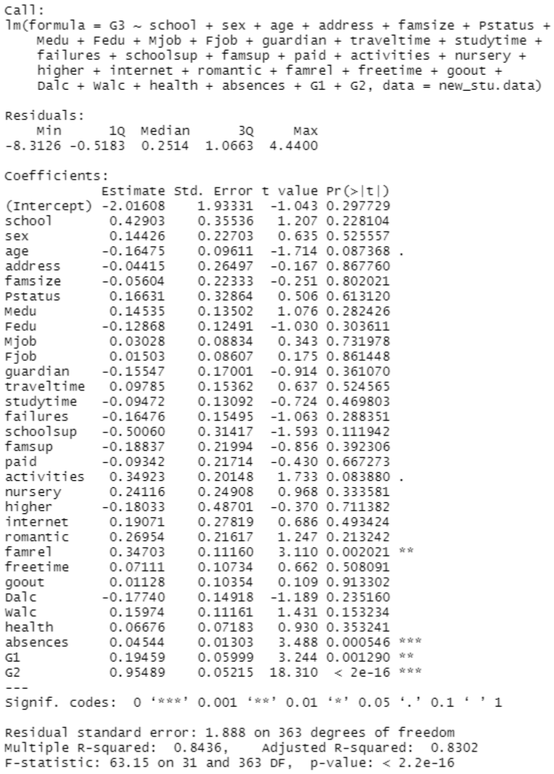
```
Call:
lm(formula = G2 ~ sex + address + Medu + studytime + failures +
    schoolsup + famsup + romantic + goout, data = new_stu.data)

Residuals:
     Min       1Q   Median       3Q      Max
-11.7554  -1.9558   0.2071   2.2377   8.6579

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.6390     1.6348   3.449 0.000624 ***
sex           0.7963     0.3660   2.176 0.030160 *
address      -0.9537     0.4122  -2.314 0.021204 *
Medu          0.5041     0.1652   3.051 0.002441 **
studytime     0.5694     0.2164   2.632 0.008831 **
failures     -1.3827     0.2415  -5.725 2.08e-08 ***
schoolsup     1.2129     0.5134   2.362 0.018655 *
famsup        0.7398     0.3608   2.050 0.040994 *
romantic      0.7568     0.3645   2.076 0.038513 *
goout        -0.5055     0.1546  -3.271 0.001170 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.355 on 385 degrees of freedom
Multiple R-squared:  0.2224,    Adjusted R-squared:  0.2042
F-statistic: 12.24 on 9 and 385 DF,  p-value: < 2.2e-16
```

Appendix 8: G2 9-variable model.



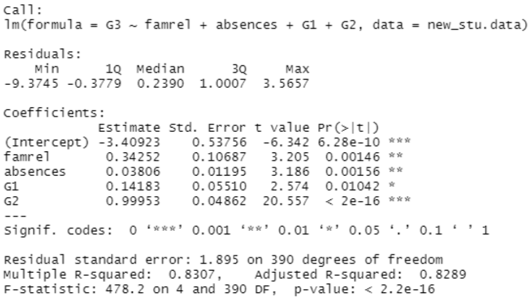
```
Call:
lm(formula = G3 ~ school + sex + age + address + famsize + Pstatus +
    Medu + Fedu + Mjob + Fjob + guardian + traveltime + studytime +
    failures + schoolsup + famsup + paid + activities + nursery +
    higher + internet + romantic + famrel + freetime + goout +
    Dalc + Walc + health + absences + G1 + G2, data = new_stu.data)

Residuals:
    Min      1Q  Median      3Q     Max
-8.3126 -0.5183  0.2514  1.0663  4.4400

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.01608    1.93331  -1.043 0.297729
school       0.42903    0.35536   1.207 0.228104
sex          0.14426    0.22703   0.635 0.525557
age         -0.16475    0.09611  -1.714 0.087368 .
address     -0.04415    0.26497  -0.167 0.867760
famsize     -0.05604    0.22333  -0.251 0.802021
Pstatus      0.16631    0.32864   0.506 0.613120
Medu         0.14535    0.13502   1.076 0.282426
Fedu        -0.12868    0.12491  -1.030 0.303611
Mjob         0.03028    0.08834   0.343 0.731978
Fjob         0.01503    0.08607   0.175 0.861448
guardian    -0.15547    0.17001  -0.914 0.361070
traveltime   0.09785    0.15362   0.637 0.524565
studytime   -0.09472    0.13092  -0.724 0.469803
failures    -0.16476    0.15495  -1.063 0.288351
schoolsup   -0.50060    0.31417  -1.593 0.111942
famsup      -0.18837    0.21994  -0.856 0.392306
paid        -0.09342    0.21714  -0.430 0.667273
activities   0.34923    0.20148   1.733 0.083880 .
nursery      0.24116    0.24908   0.968 0.333581
higher      -0.18033    0.48701  -0.370 0.711382
internet     0.19071    0.27819   0.686 0.493424
romantic     0.26954    0.21617   1.247 0.213242
famrel       0.34703    0.11160   3.110 0.002021 **
freetime     0.07111    0.10734   0.662 0.508091
goout        0.01128    0.10354   0.109 0.913302
Dalc        -0.17740    0.14918  -1.189 0.235160
Walc         0.15974    0.11161   1.431 0.153234
health       0.06676    0.07183   0.930 0.353241
absences     0.04544    0.01303   3.488 0.000546 ***
G1           0.19459    0.05999   3.244 0.001290 **
G2           0.95489    0.05215  18.310  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.888 on 363 degrees of freedom
Multiple R-squared:  0.8436,    Adjusted R-squared:  0.8302
F-statistic: 63.15 on 31 and 363 DF,  p-value: < 2.2e-16
```

Appendix 9: G3 full model



```
Call:
lm(formula = G3 ~ famrel + absences + G1 + G2, data = new_stu.data)

Residuals:
    Min      1Q  Median      3Q     Max
-9.3745 -0.3779  0.2390  1.0007  3.5657

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.40923    0.53756  -6.342 6.28e-10 ***
famrel       0.34252    0.10687   3.205  0.00146 **
absences     0.03806    0.01195   3.186  0.00156 **
G1           0.14183    0.05510   2.574  0.01042 *
G2           0.99953    0.04862  20.557  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.895 on 390 degrees of freedom
Multiple R-squared:  0.8307,    Adjusted R-squared:  0.8289
F-statistic: 478.2 on 4 and 390 DF,  p-value: < 2.2e-16
```

Appendix 10: G3 4-variable model

References

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 978-9077381-39-7.