Louise Dietrich - Alberto Monari - Kevin Da Silva

Econometrics and Statistical Models Project

Predictive model & linear regression

Objective

Our dataset is made of 395 observations of students from two Portuguese schools.

Our main objective is to create a predictive model able to predict the final grade of a student, based on the different variables recorded.

In order to do that we will perform a multiple linear regression on our data to build our model.

Road Map

Business question

Which factors have an influence on the final grade of the students and how can we maximize it ?

Context

The Portuguese Ministry of Education has decided to collect students data from 2 Portuguese schools to try to better understand the factors influencing students' final grades in Portugal.

Stakeholders

With the study of this dataset, education professionals and student parents will be able to influence the students results by minimizing factors having a negative impact, and strengthening the ones improving the final grade.

Analysis goals

The purpose of this analysis is to model the attribution of final grades to the students, in order to predict their results or to know what factors could improve their final year grade.

Cleaning data

There is no missing value in the data set. We have to change the format of some variables as factor to start our study.

Methodology

We will perform multiple linear regressions to determine the influent factors on the student's final grade.

Data

The original dataset presents 395 observations of students with 33 different attributes, including grades, some of their habits and social-economic situation characteristics.

Intuition -> Hypothesis

We sorted each variable to the corresponding impact we think it might have on students' performance:

Positive Impact	No Significant Impact	Negative Impact
Medu, Fedu, studytime, schoolsup, famsup, paid, higher, internet, health	School, sex, age, famsize, Pstatus, Mjob, Fjob, reason, guardian, activities, nursery,romantic, famrel, freetime	Traveltime, failures, gout, Dalc, Walc, absences

Name of the variables explicited in Data Description

Findings -> Insights

We removed every attribute that was not relevant in the prediction of the final grade.

We are left with 5 different variables,

2 with a positive correlation to the final grade:

- The student gender if it is male
- The mother's level of education

3 with a negative correlation:

- The number of past class failures
- If the student is in a romantic relationship
- If the student goes out with friends

Those 5 variables explain 17,02% of our prediction model, in a confidence interval of 95%.

Recommendations

The education professionals should look into the gender gap in the final grades (10,91/20 on average for male and 9,97/20 for female students) to ensure equality at school.

They could also prevent the negative correlation of the social life of the students on its results (having a romantic relationship or going out with friends) by giving insights on how to combine studies and personal life.

Data source

Provided by the UCI Machine Learning Repository

https://archive.ics.uci.edu/ml/datasets/Student+Performance

Data collected by P. Cortez & A. Silva, 2005-2006, Porto (Portugal), EUROSIS

Data Description

The dataset presents information of 395 students from two Portuguese schools.

It gathers observations of :
- different habits, such as their alcohol consumption or their number of hours studied
- socioeconomic factors, such as their parents' education, profession,…
- their grades for the first & second period, and grades of the final exam

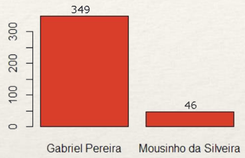
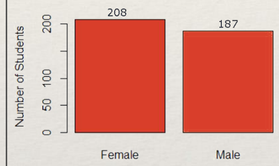
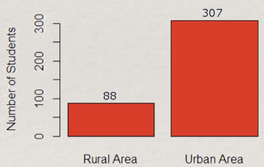
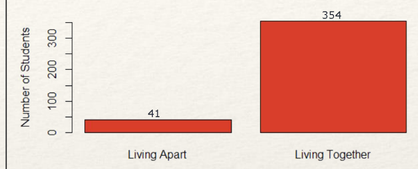
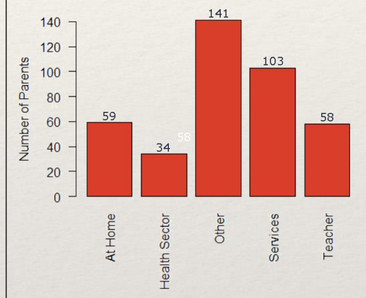
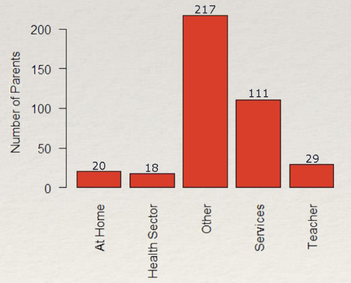
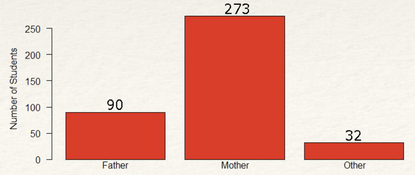
We have chosen 30 independant variables (all the variables except G1, G2, G3) as predictors to start our model : 17 categorial variables (factors), and 13 quantitative variables.

The dependant variable is G3.

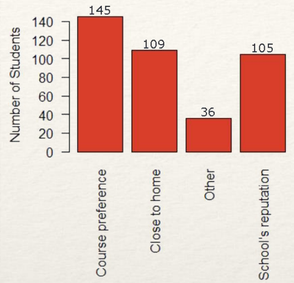
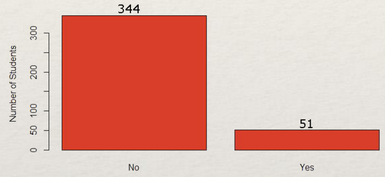
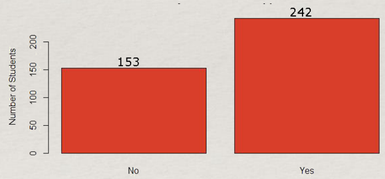
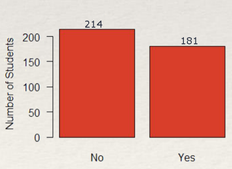
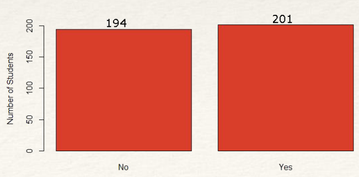
1 school - student's school (binary: "GP" - Gabriel Pereira or "MS" - Mousinho da Silveira)
2 sex - student's sex (binary: "F" - female or "M" - male)
3 age - student's age (numeric: from 15 to 22)
4 address - student's home address type (binary: "U" - urban or "R" - rural)
5 famsize - family size (binary: "LE3" - less or equal to 3 or "GT3" - greater than 3)
6 Pstatus - parent's cohabitation status (binary: "T" - living together or "A" - apart)
7 Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
9 Mjob - mother's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")
10 Fjob - father's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")
11 reason - reason to choose this school (nominal: close to "home", school "reputation", "course" preference or "other")
12 guardian - student's guardian (nominal: "mother", "father" or "other")
13 traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14 studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15 failures - number of past class failures (numeric: n if 1<=n<3, else 4)
16 schoolsup - extra educational support (binary: yes or no)
17 famsup - family educational support (binary: yes or no)
18 paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19 activities - extra-curricular activities (binary: yes or no)
20 nursery - attended nursery school (binary: yes or no)
21 higher - wants to take higher education (binary: yes or no)
22 internet - Internet access at home (binary: yes or no)
23 romantic - with a romantic relationship (binary: yes or no)
24 famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25 freetime - free time after school (numeric: from 1 - very low to 5 - very high)
26 goout - going out with friends (numeric: from 1 - very low to 5 - very high)
27 Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28 Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29 health - current health status (numeric: from 1 - very bad to 5 - very good)
30 absences - number of school absences (numeric: from 0 to 93)
31 G1 - first period grade (numeric: from 0 to 20)
31 G2 - second period grade (numeric: from 0 to 20)
32 G3 - final grade (numeric: from 0 to 20, output target)

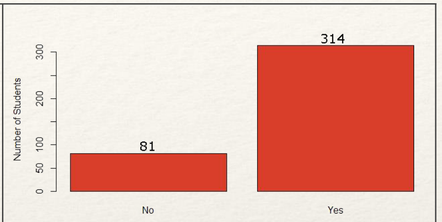
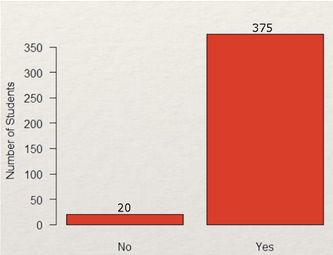
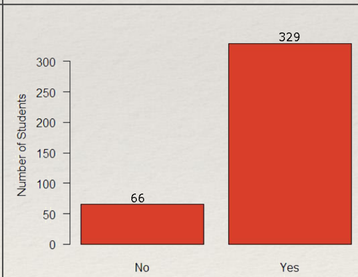
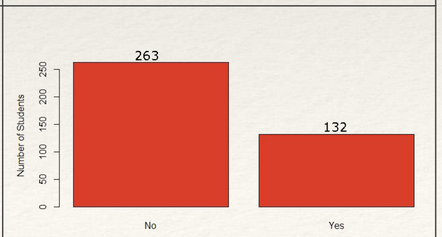
Exploratory analysis

Categorial Variables

school *factor*	School attended 2 levels : Gabriel Pereira or Mousinho da Silveira	
sex *factor*	Gender 2 levels: M or F	
address *factor*	Home address type 2 levels: rural of urban	
famsize *factor*	Family's size 2 levels: more or less than 3 members	
Pstatus *factor*	Parents cohabitation status 2 levels: living apart or together	
Mjob *factor*	Mother's professional occupation 5 levels: at home, health sector, services, teacher, other	
Fjob *factor*	Father's professional occupation 5 levels: at home, health sector, services, teacher, other)	
guardian *factor*	Guardian 3 levels: father, mother, other	

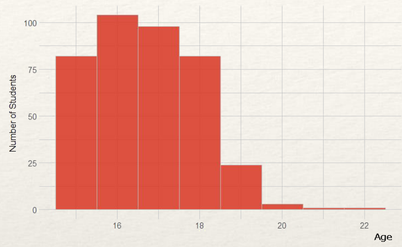
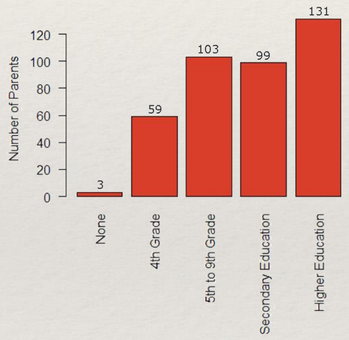
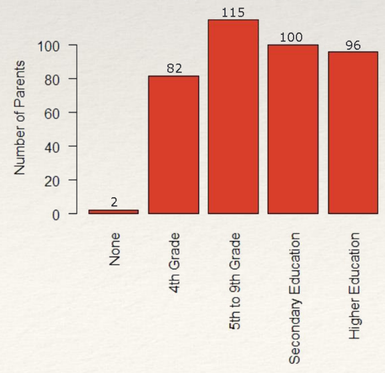
Exploratory analysis

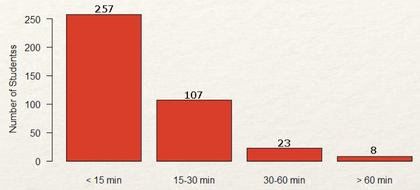
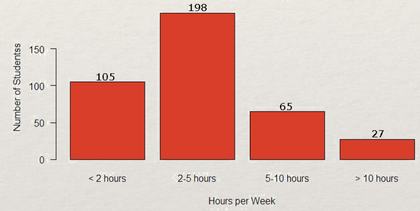
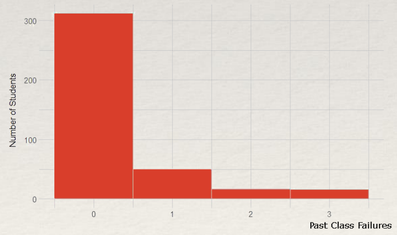
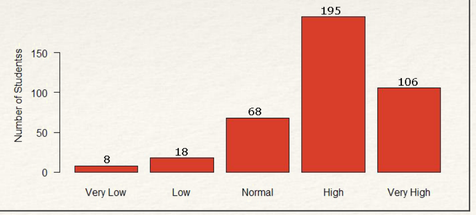
reason *factor*	Reason to chose that school 4 levels: Course preference, close to home, school's reputation, other	
schoolsup *factor*	Extra educational support 2 levels: yes / no	
Famsup *factor*	Family educational support 2 levels: yes/no	
paid *factor*	Following of extra paid classes 2 levels: yes/no	
activities *factor*	Extra-curricular activities 2 levels: yes/no	

nursery *factor*	Attended the nursery school 2 levels: yes/no	
higher *factor*	Intention to take higher education 2 levels: yes/no	
internet *factor*	Internet connection at home 2 levels: yes/no	
romantic *factor*	In romantic relationship 2 levels: yes/no	

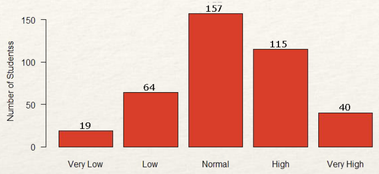
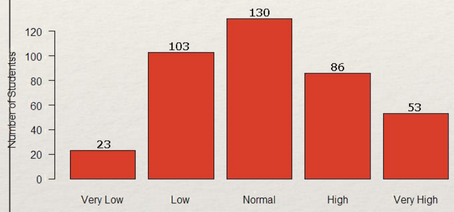
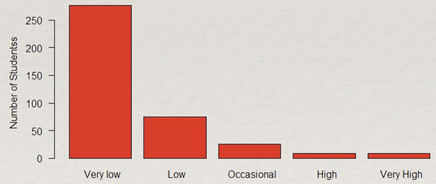
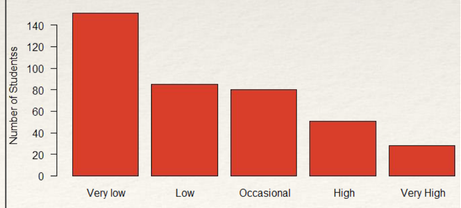
Exploratory analysis

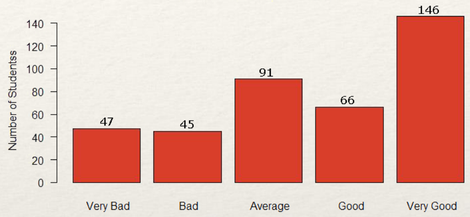
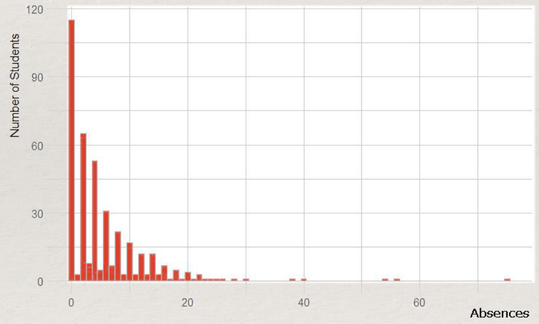
Quantitative Variables

age *interger*	Age	
Medu *integer*	Mother's education Values: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education 4 - higher education	
Fedu *integer*	Father's education Values: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education 4 - higher education	

traveltime *integer*	Home to school time travel Values: 1 - <15 min 2- 15-30 min 3- 30-60 min 4- >60 min	
studytime *integer*	Weekly study time Values: 1 - <2 hours 2- 2 to 5 hours 3- 5 to 10 hours 4- >10 hours	
failures *integer*	Number of past class failures	
famrel *integer*	Quality of the family relationship From 1 – very bad to 5 – excellent	

Exploratory analysis

freetime *Interger*	Free time after school From 1– very low to 5– very high	
goout *integer*	Amount of time spent going out with friends From 1– very low to 5– very high	
Dalc *Integer*	Workday alcohol consumption From 1– very low to 5– very high	
Walc *integer*	Weekend alcohol consumption From 1– very low to 5– very high	

health *integer*	Current health status From 1- very bad to 5- very good	
absences *integer*	Number of school absences	 > summary(student_data$absences) Min. 1st Qu. Median Mean 3rd Qu. Max. 0.000 0.000 4.000 5.709 8.000 75.000

Data Cleaning

Firstly we checked the structure and we had to change the format of different variables from character to factor to explicitly express the levels represented by those variables.

```
> str(student_data)
'data.frame':	395 obs. of  33 variables:
 $ school    : chr  "GP" "GP" "GP" "GP" ...
 $ sex       : chr  "F" "F" "F" "F" ...
 $ age       : int  18 17 15 15 16 16 16 17 15 15 ...
 $ address   : chr  "U" "U" "U" "U" ...
 $ famsize   : chr  "GT3" "GT3" "LE3" "GT3" ...
 $ Pstatus   : chr  "A" "T" "T" "T" ...
 $ Medu      : int  4 1 1 4 3 4 2 4 3 3 ...
 $ Fedu      : int  4 1 1 2 3 3 2 4 2 4 ...
 $ Mjob      : chr  "at_home" "at_home" "at_home" "health" ...
```

```
> str(student_data)
'data.frame':	395 obs. of  33 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 ...
 $ famsize   : Factor w/ 2 levels "GT3","LE3": 1 1 2 1 1 2 2 1 2 1 ...
 $ Pstatus   : Factor w/ 2 levels "A","T": 1 2 2 2 2 2 2 1 1 2 ...
 $ Medu      : int  4 1 1 4 3 4 2 4 3 3 ...
 $ Fedu      : int  4 1 1 2 3 3 2 4 2 4 ...
 $ Mjob      : Factor w/ 5 levels "at_home","health",..: 1 1 1 2 3 4 3 3 4 3 ...
```

Then we summarized the data to check the values. As all our dataset is made of variables contained in intervals, we could see that there was no problem with outliers values, and there was no missing values too.

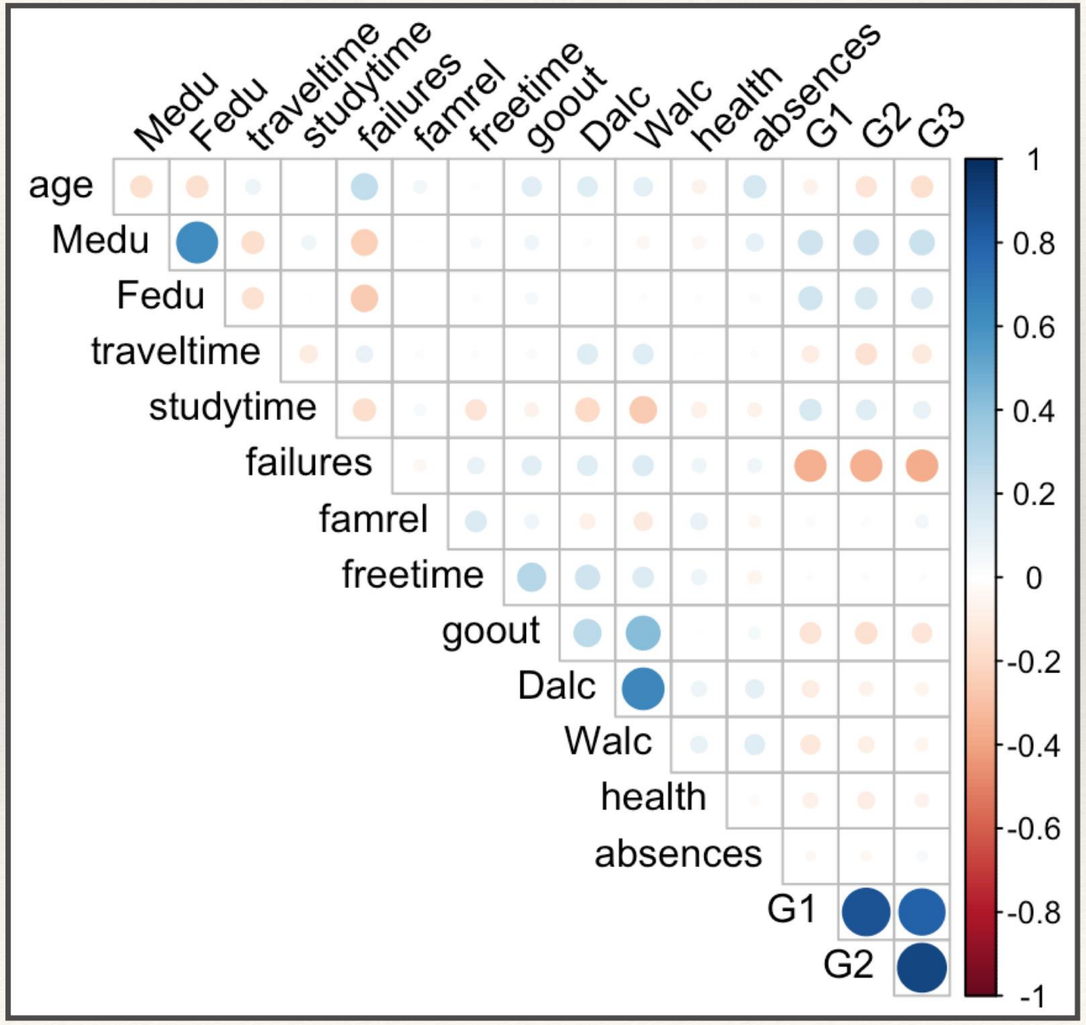
```
> summary(student_data)
 school   sex         age       address famsize   Pstatus      Medu          Fedu          Mjob
 GP:349   F:208   Min.   :15.0   R: 88   GT3:281   A: 41   Min.   :0.000   Min.   :0.000   at_home : 59
 MS: 46   M:187   1st Qu.:16.0   U:307   LE3:114   T:354   1st Qu.:2.000   1st Qu.:2.000   health  : 34
                  Median :17.0                          Median :3.000   Median :2.000   other   :141
                  Mean   :16.7                          Mean   :2.749   Mean   :2.522   services:103
                  3rd Qu.:18.0                          3rd Qu.:4.000   3rd Qu.:3.000   teacher : 58
                  Max.   :22.0                          Max.   :4.000   Max.   :4.000
     Fjob          reason     guardian     traveltime      studytime        failures      schoolsup famsup
 at_home : 20   course   :145   father: 90   Min.   :1.000   Min.   :1.000   Min.   :0.0000   no :344   no :153
 health  : 18   home     :109   mother:273   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000   yes: 51   yes:242
 other   :217   other    : 36   other : 32   Median :1.000   Median :2.000   Median :0.0000
 services:111   reputation:105               Mean   :1.448   Mean   :2.035   Mean   :0.3342
 teacher : 29                                3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:0.0000
                                             Max.   :4.000   Max.   :4.000   Max.   :3.0000
```

Data cleaning

We studied the correlation among quantitative variables by creating a scatterplot.
Since G1 and G2 are the grades for the first and second semester respectively, they are highly correlated with the final grade of the year G3. For this reason we have decided to remove G1 and G2.

We observed a high correlation between Medu/Fedu and decided to only keep mother education because seems to be more commonly to have a biggest impact.

For the correlation between Dalc and Walc (with is the consumption of alcohol) we solve it by computing the mean as an evaluation of the alcohol consumption on the whole week.



Methodology used on the Data

We will perform a multiple linear regression to build a model estimating the performance of the students (in terms of the final grade), based on its correlation (or non-correlation) with all other variables

Linear regression

We used the backward elimination method and removed the variables one by one, removing the less significant variable each time (with the highest p-value greater than 0.05).

The goal is to have a model with only significant variables (p-value lower than 0.05).

```
linreg <- lm(G3~factor(sex) + Medu + failures
        + factor(romantic) + goout, data=student_data)
summary(linreg)
```

Anova test:
p-value < 5%, reject H0
The regression is globally significant

The final model:

```
> summary(linreg)

Call:
lm(formula = G3 ~ factor(sex) + Medu + failures + factor(romantic) +
    goout, data = student_data)

Residuals:
    Min      1Q  Median      3Q     Max
-13.1385  -2.0956   0.4001   2.8171   8.6257

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        10.6342     0.8244  12.900  < 2e-16 ***
factor(sex)M        0.9557     0.4263   2.242  0.02554 *
Medu                0.6157     0.1999   3.080  0.00221 **
failures           -1.8943     0.2964  -6.391 4.72e-10 ***
factor(romantic)yes -0.9279     0.4509  -2.058  0.04028 *
goout              -0.4571     0.1916  -2.386  0.01753 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.173 on 389 degrees of freedom
Multiple R-squared:  0.1807,          Adjusted R-squared:  0.1702
F-statistic: 17.16 on 5 and 389 DF,  p-value: 2.371e-15
```

Linear regression

The final equation to predict G3 (the final grade year) is:

G3 = 10.63 + 0.96 sex(M) + 0.62 Medu - 1.89 failures - 0.93 romantic(yes) - 0.46 goout

The intercept is 10.63.

Meaning of the coefficients of the significant variables (all other conditions being equal):

- surprisingly, males have on average higher grades of almost one unit higher than females;
- the level of education of the mother (and also that of the father, as they are correlated) has a slightly positive impact on the final grade;
- the most important impact on the grade is the number of examinations failed in the past: each failed exam has a negative impact of almost two units on the final grade;
- being in a romantic relationship has a negative impact on the final grade of almost one unit;
- the fact of going out often has a slight negative impact on the final grade.

Quality of the model:

We found a R-squared of 17,02% wich is low but can be explained by the fact that there are probably many other variables which could explain the final grade of the students.

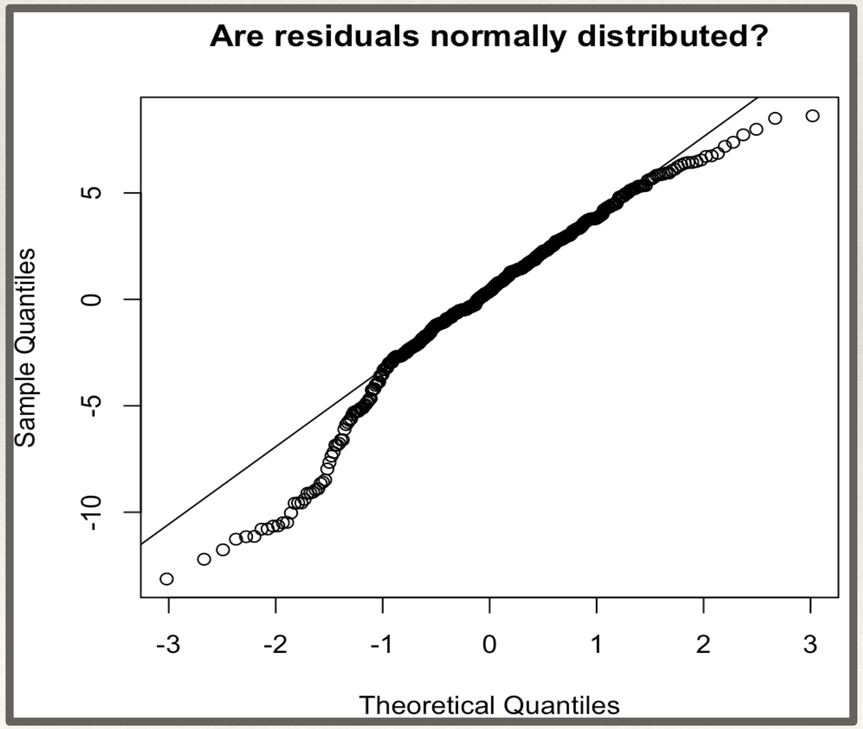
We also have a relatively small dataset. Having more observations would give us a greater power to detect patterns or differences and thus increase our predictive power.

Verifying the hypotheses

```
res <- linreg$residuals
summary(res)
```

```
> summary(res)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-13.1385  -2.0956   0.4001   0.0000   2.8171   8.6257
```

The mean of the residuals is equal to zero



The normality

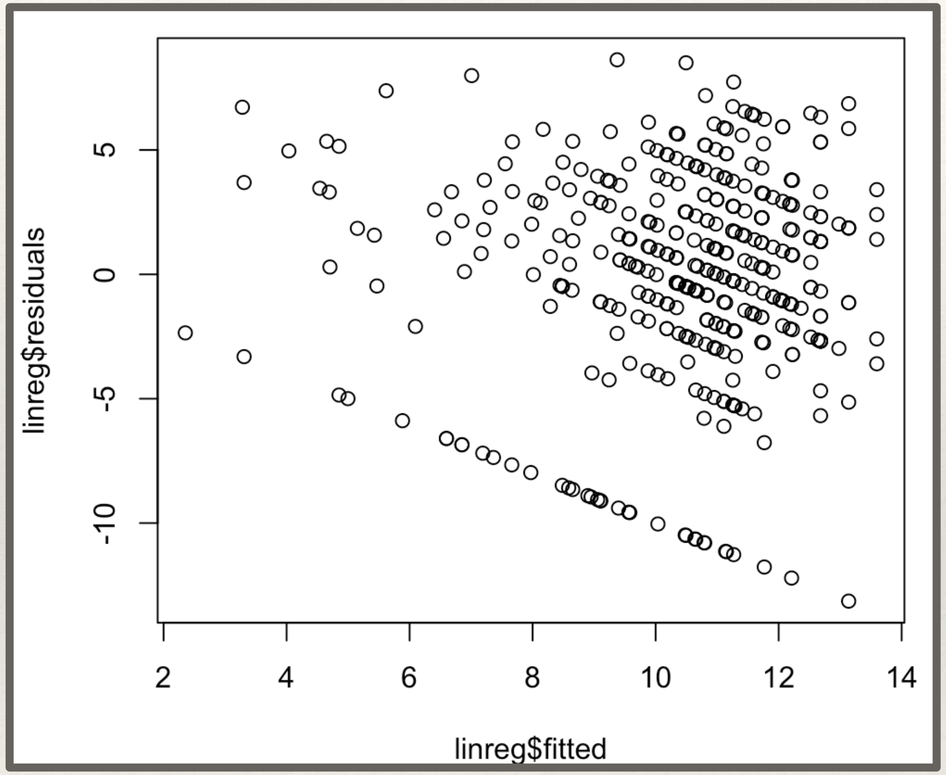
```
        Shapiro-Wilk normality test

data:  linreg$residuals
W = 0.96027, p-value = 7.477e-09
```

The p-value is < 5%
We reject H0 so the residuals are not normally distributed.

The assumption of the normality of the residuals is not the most invalidating. The sample size is large enough (> 30 observations) so the results of the model remain valid, even if the assumption of normality is not verified.

Verifying the hypotheses




```
> bptest(linreg,studentize=FALSE, data=data)

        Breusch-Pagan test

data:  linreg
BP = 5.0352, df = 5, p-value = 0.4116
```

P-value > 5%, we accept H0
Residuals are homoskedastic.

```
> durbinWatsonTest (linreg,max.lag=1)
 lag Autocorrelation D-W Statistic p-value
   1    -0.02873169      2.05287    0.62
 Alternative hypothesis: rho != 0
```

Regarding the durbinWatson test there is no autocorrelation in the residuals.
P-value >5%, we accept H0

```
> vif(linreg)
   factor(sex)          Medu      failures factor(romantic)         goout
      1.027498      1.083203      1.099091      1.026047      1.029474
```

All the VIFs are < 5, so there is no multicollinearity.

Forecasting

We used our predictive model to forecast the final grade of two different student profiles:

1- A male student, who's mother education is equivalent to primary education, who failed 3 classes in the past, is engaged in a romantic relationship, and who often goes out with his friends [sex=M, Medu=1, failures=3, romantic=yes, goout=4]

2- A female student, who's mother education is equivalent to secondary education, who never failed a class in the past, is not engaged in a relationship, and does not got out often with her friends [sex=F, Medu=3, failures=0, romantic=no, goout=1]

```
> predict(linreg, predictors1, type="response",interval = "prediction", level = 0.95)
      fit       lwr     upr
1 3.766192 -4.615018 12.1474
> predictors1 <- data.frame(sex='F', Medu=3, failures=0, romantic ='no',
+                           goout=1)
> predict(linreg, predictors1, type="response",interval = "prediction", level = 0.95)
      fit      lwr      upr
1 12.02425 3.756387 20.29211
```

For the student 1 the predicted grade is 3.77/20, associated with the 95% confidence interval [-4.61 , 12.15]

For the student 2 the predicted grade is 12.02/20, associated with the 95% confidence interval [3.76 , 20.29]

Conclusions

As the quality of the model is low the confidence interval is quite wide and takes values outside of [0,20].
This is mainly due because of the numerous other factors we have to consider for the prediction of the performance at a final exam, such as the anxiety of the student or his/her knowledge level on the specifically evaluated subject or the luck component he/she may have during the final examination.

Still our model gives good insights on the chance of success or failure of the students regarding the different significant variables we identified.

Some of our initial assumptions have been confirmed:
- the mother's education (and in general that of both parents) has a positive impact on the student's final grade, it is likely that these parents have more chances to help their children with their homework and to motivate them to study more.
- the fact of going out a lot during the week takes time away from the study and the fact of having failed previously indicates a poor attitude to study so these two factors have a negative impact on the final grade.

Two initial assumptions have not been confirmed:
- surprisingly gender, in favour of males, is a significant factor in the student's final grade
- the fact of having a romantic relationship is a negative factor in the final grade

In addition, there are many other variables that we expected to have a positive or negative influence on the final grade, but which in the end are not in the final model, which only takes into account the most significant variables.