# Using Demographic and Survey Data to Predict Student Performance

Benjamin Russoniello, David Smith

May 1 2020

**Abstract:** We examine data from a Portuguese secondary school mathematics course and explore the different factors that are meaningful to student performance. Our focus is heavily aligned with creating a practical and interpretable model that can be used by educators and government officials to find students at risk of performing poorly. Specifically, several models are built to predict the first and last grades of the course using both multiple linear regression and logistic regression. The linear regression models show violation of basic assumptions, and attempts at fitting logistic regression models with relaxed assumptions are done. After the models are assessed, we discuss methods that may result in a better inference. We conclude that the most significant variable in predicting future performance is past performance, although we maintain that some of the survey and demographic data is important to student success.

## 1 Introduction

During the 2005-2006 school year, Dr. Paula Cortez and Alice Silva of the University of Minho, Portugal, collected data on 395 secondary school students taking a mathematics course. These data can lend itself to important questions on what factors impact student performance, and to what degree. We explore linear methods to assess the important factors affecting student grades, and what predictive accuracy can be achieved using these methods. Linear methods transfer well to usability, as the coefficients of models' variables are often easy to interpret.

## 2 The Data and Problem Statement

The data set was retrieved from the UCI Machine Learning Repository. The variables in the data set were collected using surveys and school records, and the complete list of variables are shown below in Table 1.

Prior to modelling of the data, cleaning the data is necessary to ensure an adept and meaningful analysis. 38 students received a score of 0 for G3 (the lowest score possible), although they did have positive scores for G1 and G2. The source paper provided no clear explanation as to why nearly 10% of the students received a score of 0. We decided that it was unreasonable to conclude that these were actual scores earned by the students. Rather, they represent students who either did not complete the course, or took a grade of 0 for another academic reason. These observations are not relevant to the question at hand, as we only care about the performance of students who completed a course. We maintain that these observations may be useful for modelling the risk of dropping out, but since this is not the focus of this paper, we decided to conduct our analysis without these observations. No other instances of missing data were observed, so we began to conduct our exploratory data analysis.

While we want to determine a model to predict student grades at the end of a class, we suspect that one of the largest factors in that potential model will be each student's grade at earlier checkpoints during the class. Academic institutions would most likely find more use out of a model that seeks to predict students' final grades without using previous grades in the same course. We still wish to explore both of the above options, so this study aims to answer two main questions:

**1. Primary Question:** How can we use regression to find the variables that impact **G3**, a student's final grade?

**2. Secondary Question:** How can we use regression to find the variables that impact **G1**, a student's first reported grade?

| Variable Name | Description |
|---|---|
| school | student's school (binary: "GP" - Gabriel Pereira or "MS" - Mousinho da Silveira) |
| sex | student's sex (binary: "F" - female or "M" - male) |
| age | student's age (numeric: from 15 to 22) |
| address | student's home address type (binary: "U" - urban or "R" - rural) |
| famsize | family size (binary: "LE3" - less or equal to 3 or "GT3" - greater than 3) |
| Pstatus | parent's cohabitation status (binary: "T" - living together or "A" - apart) |
| Medu | mother's education (numeric: 0(none) - 4(higher education) |
| Fedu | father's education (numeric: 0(none) - 4(higher education) |
| Mjob | mother's job (nominal: "teacher", "health" "services" , "at_home" or "other") |
| Fjob | (nominal: "teacher", "health" "services" , "at_home" or "other") |
| reason | reason to choose this school (nominal: close to "home", school "reputation", "course" preference or "other") |
| guardian | student's guardian (nominal: "mother", "father" or "other") |
| 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) |
| studytime | weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours) |
| failures | number of past class failures (numeric: n if $1 <= n < 3$, else 4) |
| schoolsup | extra educational support (binary: yes or no) |
| famsup | family educational support (binary: yes or no) |
| paid | extra paid classes within the course subject (Math or Portuguese) (binary: yes or no) |
| activities | extra-curricular activities (binary: yes or no) |
| nursery | attended nursery school (binary: yes or no) |
| higher | wants to take higher education (binary: yes or no) |
| internet | Internet access at home (binary: yes or no) |
| romantic | with a romantic relationship (binary: yes or no) |
| famrel | quality of family relationships (numeric: from 1 - very bad to 5 - excellent) |
| freetime | free time after school (numeric: from 1 - very low to 5 - very high) |
| goout | going out with friends (numeric: from 1 - very low to 5 - very high) |
| Dalc | workday alcohol consumption (numeric: from 1 - very low to 5 - very high) |
| Walc | weekend alcohol consumption (numeric: from 1 - very low to 5 - very high) |
| health | current health status (numeric: from 1 - very bad to 5 - very good) |
| absences | number of school absences (numeric: from 0 to 93) |
| G1 | first period grade (numeric: from 0 to 20) |
| G2 | second period grade (numeric: from 0 to 20) |
| G3 | final grade (numeric: from 0 to 20) |

Table 1: Full list of data set variables.

# 3 G3 Model: Preliminary data analysis

In order to begin the data analysis relevant to building a model to address our primary question, we decided to examine the variables that are highly correlated with G3. Since a 33x33 scatterplot matrix would be difficult to interpret, we decided to create a correlation matrix first to find the numeric and factor variables most highly correlated with G3. These correlation values are shown below in table 2.

It is evident, and expected, that G1 and G2 both have strong correlations with G3. Then, any model that we build to make inference about what variables are important for predicting G3 should include one or both of G1 and G2. In order to further analyze these relationships, we proceeded to create a 3x3 scatterplot matrix for these variables:

By inspection of figure 1, we can conclude that a strong mutual linear relationship exists among G1, G2, and G3. In order to prevent multicollinearity, it would be best to not include both G2 and G1 in the model. Thus, we define a new variable, "Improvement", as G2-G1. Improvement showed very little correlation with G1, and hence was appropriate to add to a model containing G1. Additionally, another new variable was defined prior to modelling. AvEdu, or "Average Education Level", is defined as (Medu+Fedu)/2. This represents, roughly, the overall education level of the student's parents.

| Variable Name | Correlation with G3 |
|---|---|
| age | -0.14 |
| Medu | 0.19 |
| Fedu | 0.16 |
| traveltime | -0.10 |
| studytime | 0.13 |
| failures | -0.29 |
| famrel | 0.04 |
| freetime | 0.02 |
| goout | -0.18 |
| Dalc | -0.14 |
| Walc | -0.19 |
| health | -0.08 |
| absences | -0.21 |
| G1 | 0.89 |
| G2 | 0.97 |

Table 2: numeric variables and their correlation with G3

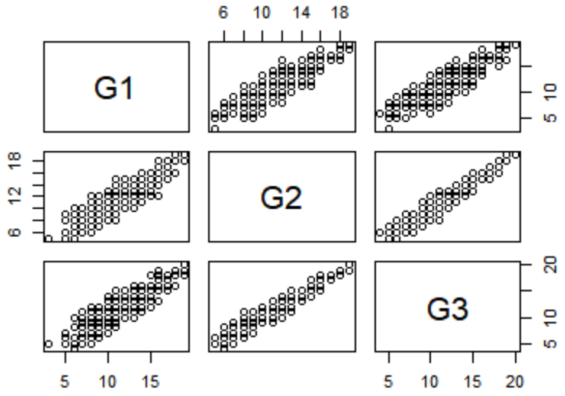


Figure 1: G1,G2,G3 scatterplot matrix

# 4 G3 Model Building: Multiple Linear Regression

In order to model G3 as a function of other covariates, we begin with ordinary least squares regression. Given G3's strong correlations with G1 and Improvement, we start by fitting the following model:

$$E(G3|X) = \beta_0 + \beta_1 G1 + \beta_2 Improvement$$

Table 3 below summarizes the fit of this model:

| Value | Estimate | P-value |
|---|---|---|
| Intercept | .195 | .241 |
| $\beta_1$ | .998 | <.0001 |
| $\beta_2$ | .887 | <.0001 |
| $R^2$ | .9347 | N/A |
| RSE | .8272 | N/A |

Table 3: G3 Multiple Regression - Intercept estimates and p-values for testing $\beta_i = 0$

At first glance, this seems like a promising model, with a high $R^2$ and a low residual standard error. However, it is necessary that we verify our assumptions before declaring this as a valid model. By fitting multiple linear regression, we assumed that the residuals were normally distributed with constant variance. By examining a residuals vs. fitted

plot, we can check the assumption of constant variance. Additionally, we can use a normal quantile plot to verify that the residuals are normally distributed. These two plots are shown below:

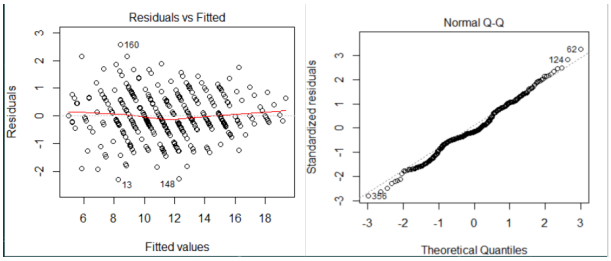


Figure 2: G3 - Fitted vs. residuals and residual quantile plot

The main issue with the figures above is the lack of constant variance shown by the residuals. As the fitted value of G3 increases, we can see that the variance of the residuals decreases. Thus, this model should not be used and we should seek a different set of variables to explain G3. If the nonconstant variance is still not fixed after adding more variables, we would then try variable transformations and apply weighted least squares.

In order to find other variables in the dataset that can explain variability in G3, we used forward stepwise regression, and we chose to stop adding variables after we had reached a minimum in AIC. By using this technique, we fit the following model:

$$E(G3|X) = \beta_0 + \beta_1 G1 + \beta_2 Improvement + \beta_3 famrel + \beta_4 goout + \beta_5 health + \beta_6 absences + \beta_7 paid$$

A summary of the fit of this model is given below in Table 4. Although stepwise regression was able to find

| Value | Estimate | P-value |
|---|---|---|
| Intercept | .320 | .314 |
| $\beta_1$ | .987 | <.0001 |
| $\beta_2$ | .876 | <.0001 |
| $\beta_3$ | .158 | .0013 |
| $\beta_4$ | -.082 | .0398 |
| $\beta_5$ | -.067 | .0311 |
| $\beta_6$ | -.010 | .0519 |
| $\beta_7(level = "yes")$ | -.124 | .1489 |
| $R^2$ | .9389 | N/A |
| RSE | .8057 | N/A |

Table 4: G3 stepwise Multiple Regression - Intercept estimates and p-values for testing $\beta_i = 0$

additional significant variables, the residuals plot remained roughly the same as it was previously. Thus, we began to check if variable transformations or weighted least squares could aid us in our fit.

A likelihood ratio test was performed on G3 to see if a power transformation of it would be helpful in the fit. More specifically, if lambda is the transformation parameter, then we performed the test:

$$H_0 : \lambda = 1$$

$$H_1 : \lambda \neq 1$$

When this test was conducted, we received a p-value of .67, meaning that we could not find sufficient evidence that a transformation of G3 would aid in our fit. Additionally, most of the independent variables had very small ranges, so transformations of these variables were not likely to be helpful. We tried several more fits by applying shifted power transformations to the variables (due to absences having a minimum value of 0), but none of these transformations resulted in a residual plot that met our assumptions. Due to the nonconstant variance of the residuals we observed, we began fitting weighted least squares to the model to see if that would improve the fit. Unfortunately, we did not have a theoretical basis for a choice of weights. We were not aware of any literature that discussed the factors

that affected the variance of student test scores. Nonetheless, we began to experiment with using functions of our variables for choices of weights. None of these resulted in a residual plot that fit our assumptions. Additionally, we worried about the interpretability of a model that used weighted least squares. We hoped to fit a simple model that was easily reproducible and understood by laypersons in government and education. We were concerned that using weighted least squares would add unnecessary complexity, so we dropped this venture entirely in favor of an alternative.

# 5 Model Building - Logistic Regression

For educators and academic institutions, a more practical model could be one that predicts the risk of a student failing a course rather than one that attempts to predict an exact numerical grade. According to our data's source paper, students in the class needed to earn at least a G3 score of 10 in order to pass the class. To examine this, we set out to create a model using logistic regression that aims to predict the probability of a student failing the course, i.e. $P(G3 < 10)$:

$$P(G3 < 10) = \frac{e^{\beta_0 + \sum_{i=1}^{p} \beta_i X_i}}{1 + e^{\beta_0 + \sum_{i=1}^{p} \beta_i X_i}}$$

We fit the logistic regression model using the variables found using forward stepwise regression from before. Our results in table 5 show that while the majority of the significant predictors from our linear model remained important in this logistic model, a couple of them were revealed as unimportant in determining the probability of a student failing the course.

| Value | Estimate | P-value |
|---|---|---|
| $\beta_0$ | 17.569 | <.0001 |
| $\beta_1$ | -1.836 | <.0001 |
| $\beta_2$ | -1.482 | <.0001 |
| $\beta_3$ | -.914 | .0191 |
| $\beta_4$ | .582 | .0346 |
| $\beta_5$ | .051 | .8296 |
| $\beta_6$ | .052 | .0995 |
| $\beta_7(level = "yes")$ | .083 | .8894 |

Table 5: G3 Logistic Regression - Intercept estimates and p-values for testing $\beta_i = 0$

In order to assess the predictive accuracy of this model, the model was trained with 70% of the data. We held out the remaining 30% in a test set. We used our predicted values to produce the ROC plot shown above in figure 3.

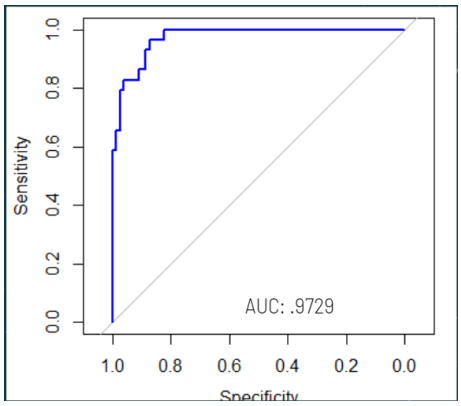


Figure 3: ROC curve and AUC for P(G3<10) model

Once again, the results of our model fit show that G1 and improvement have a very meaningful impact on class performance. Family relationships, social gatherings, and absences also have roles in affecting the chances of passing or failing the course. Health and extra paid classes, however, are not significant in our new logistic model. The

resulting ROC curve from our logistic regression model proves shows that we can use the model above to achieve both high sensitivity and specificity. Although our initial idea of a multiple linear regression model was deemed unusable because of diagnostic issues, the new logistic model that we pivoted to seems to do very well.

# 6 Exploring the Secondary Question

Our secondary question arose out of the desire of a pragmatic model for end users. Our model that was established in the previous section relies on data that is not available until we are about two-thirds into the school year. We desired a model that could take simple survey and demographic data, and help identify the students who would be at risk of failing the math course. Since G1 explains roughly 80% of the variability in G3, we figured that obtaining a reasonable estimate for G1 would provide a good proxy for the G3 score. Thus, we wanted to build a multiple regression model to predict the G1 score, in order to provide educators with a tool that would help them target low-performing students to provide additional educational resources to.

Prior to building the model, we had to first identify variables that would not be appropriate to include. First, we could not use the scores obtained for G2 or G3, since those would not have been obtained yet by the student. We also chose not to include the absences variable, as those are accumulated throughout the school year. It is possible to estimate the absences that would be obtained by the student by dividing absences by three, but that would require the assumption that the absences accumulated by the student are roughly independent of the time of year. We did believe this to be a reasonable assumption, so we chose not to use any variation of the absences variable in the model.

# 7 G1 Model Building: Multiple Linear Regression

From observing a correlation matrix, G1 did not have a strong correlation with any variables in the dataset, so we did not have a specific model to begin with. Thus, we fit the null model and used forward stepwise regression to find the significant variables to explain variability in G1. Using this approach, choosing the model with the lowest AIC resulted in fitting the following model:

$$E(G1|X) = \beta_0 + \beta_1 failures + \beta_2 schoolsup + \beta_3 Fjob + \beta_4 Walc + \beta_5 Mjob + \beta_6 famsup + \beta_7 studytime$$
$$+\beta_8 sex + \beta_9 goout + \beta_{10} health + \beta_{11} freetime$$

A summary of the fit of this model is summarized table 6 below.

This is not a very usable model, since these variables only explain about 30% of the variation in G1. Additionally, figure 4 shows a plot of the residuals against the fitted values, which indicates nonconstant variance in the model's residuals. Additionally, the quantile-quantile plot indicates some nonnormality in the residuals.

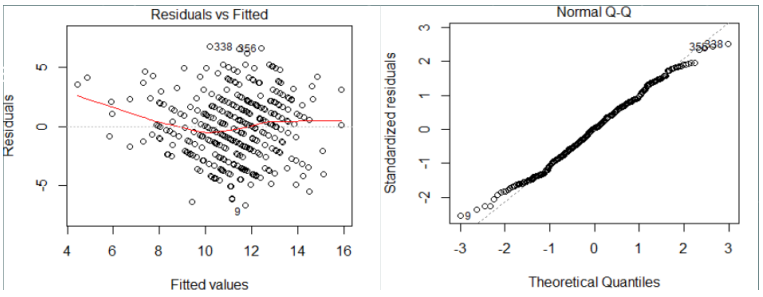


Figure 4: G1 - Fitted vs. residuals and residual quantile plot

While a Box-Cox power transformation indicated a transformation of the response could be useful, the transformations we attempted had little to no effect on the nonconstant variance of the residuals. Most of our predictor variables had very small ranges, so transformations of them would not be useful. Thus, we began to seek alternative modelling strategies to make a useful model for G1.

| Value | Estimate | P-value |
|---|---|---|
| Intercept | 13.138 | <.0001 |
| $\beta_1$ | -1.222 | <.0001 |
| $\beta_2(level = "yes")$ | -2.358 | <.0001 |
| $\beta_3(level = "health")$ | -.766 | .433 |
| $\beta_3(level = "other")$ | -1.171 | .104 |
| $\beta_3(level = "services")$ | -1.126 | .134 |
| $\beta_3(level = "teacher")$ | 1.222 | .171 |
| $\beta_4$ | -.209 | .134 |
| $\beta_5(level = "health")$ | -1.459 | .028 |
| $\beta_5(level = "other")$ | -.511 | .287 |
| $\beta_5(level = "services")$ | -.916 | .071 |
| $\beta_5(level = "teacher")$ | -.223 | .703 |
| $\beta_6(level = "yes")$ | -.9016 | .005 |
| $\beta_7$ | .629 | .001 |
| $\beta_8(level = "M")$ | .763 | .022 |
| $\beta_9$ | -.361 | .024 |
| $\beta_{10}$ | -.193 | .076 |
| $\beta_{11}$ | .226 | .158 |
| $R^2$ | .305 | N/A |
| RSE | 2.769 | N/A |

Table 6: G1 Multiple Regression - Intercept estimates and p-values for testing $\beta_i = 0$

# 8 G1 Model Building: Logistic Regression

Due to the issues with the multiple regression model built for G1, we decided to once again look towards logistic regression for building a useful model. One issue we encountered was the decision of what score would constitute our cutoff for "failing" G1. A student could not "fail" in the first period, they could only fail the entire class if their G3 score was less than 10. We examined the G1 and G3 scores and found that when a student scored less than 10 in G1, they also failed the course about two-thirds of the time. Thus, we felt comfortable using the same cutoff for G1 as we did for G3, G1<10.

We used stepwise regression to find significant variables for this logistic regression model. This resulted in the following model:

$$log\frac{P(G1 < 10)}{1 - P(G < 10)} = \beta_0 + \beta_1 AvEdu + \beta_2 failures + \beta_3 schoolsup + \beta_4 famsup + \beta_5 romantic + \beta_6 Dalc + \beta_7 Walc + \beta_8 absences$$

A summary of the model fit can be found in table 7 below.

| Value | Estimate | P-value |
|---|---|---|
| Intercept | -.228 | .666 |
| $\beta_1$ | -.508 | .002 |
| $\beta_2$ | -.943 | .003 |
| $\beta_3(level = "yes")$ | 1.67 | <.0001 |
| $\beta_4(level = "yes")$ | .386 | .222 |
| $\beta_5(level = "yes")$ | -.449 | .201 |
| $\beta_6$ | -.2607 | .257 |
| $\beta_7$ | -.308 | .046 |
| $\beta_8$ | -.0001 | .997 |

Table 7: G1 Logistic Regression - Intercept estimates and p-values for testing $\beta_i = 0$

Once again, using a 70/30 training/test split to investigate the predictive accuracy of the model, we produced the ROC curve in figure 5.

Indicated by the ROC curve and the low estimate of AUC, the logistic regression model does not do well in predicting which students will do poorly, either. We cannot achieve high sensitivity without having poor specificity, and vice versa. Thus, we conclude that logistic regression is not appropriate to accurately predict what score a student will receive for G1. This is unfortunate, but expected since most of the variables only range from one to five. If the variables contained more information, we might be able to do a better job at predicting G1.

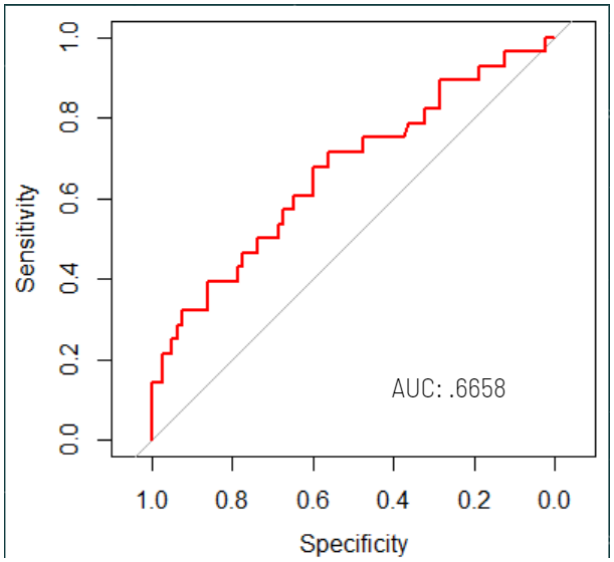


Figure 5: G1 - ROC Curve and AUC for P(G1<10) model

# 9 Conclusions

The results of this study have highlighted key factors in student performance and have provided us with inference into potential methods for further studies on this topic. Although our initial vision of the models for G1 and G3 were through multiple linear regression, our residuals and diagnostics of those models showed evidence of nonconstant variance. Thus, we concluded that multiple linear regression was not the correct method to approach the problem at hand. After finding the criteria for passing the course, we sought to instead build a logistic regression model that could predict the chance of a given student failing the course. We found success in the logistic regression model to predict G3; however, a majority of its predictive power came from the variables of previous performance in the same class. The predictive power in our logistic regression model for G1 was much weaker. This reinforced a common theme throughout the study: past academic performance has a much stronger effect on future academic performance than any other demographic or survey-based data.

Although our models without past grades were weak in predicting future grades, they still succeeded in finding other variables that have an influence on academic performance. In every model that we attempted for both G1 and G3, the amount of a student's absences proved to be significant variables. Even though it was the only non-performance variable shared by the G1 and G3 models, the other significant predictors for G1 and G3 shared common themes such as family situations, additional school support, personal relationships, and social activity.

While our methods and data could not provide us with a strong model that was unrelated to past class performance, there are paths that could lead to better inference. Finding better survey questions and wider ranges for numeric variables could improve potential regression models in this study. Different variables that contain more information could contain better information and be less biased than the survey data. For example, instead of having students rate their own health, we could record some of the student's vitals. In addition to this, a different experimental design could benefit us as well. A longitudinal study that follows each student's habits and grades over time might be more beneficial in finding important variables than the above linear methods used in this paper.

# 10 References

1. 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.

2. S. Weisberg. Applied Linear Regression, 4th ed. John Wiley  Sons, Inc., 2014.

3. X. Zhang. Lecture Notes on Statistics in Applications II. STA 5167. Spring Semester, 2020.

# A R code

```
library(alr4)
library(MASS)
library(pROC)


attach(student.mat.clean)

cor(student.mat.numeric[student.mat.numeric$G3>0,])

pairs(student.mat.numeric[,14:16])


student.mat.clean.original<-student.mat.clean #create data
#backup just in case


lmG3start<- lm(G3~G1+Improvement,data=student.mat.clean)
summary(lmG3start)
plot(lmG3start)
lmG3full <- lm(G3~.,data=student.mat.clean)
summary(lmG3start)
step.modelG3<- step(lmG3start,scope=list(lower=lmG3start, upper=lmG3full),direction="forward")

summary(step.modelG3)

lmstepG3 <- lm(G3 ~ G1 + Improvement+ famrel + goout + health + absences+
                 paid,data=student.mat.clean)
summary(lmstepG3)
plot(lmstepG3)
summary(powerTransform(lmstepG3))

#The diagnostic plots do not look good at all,
#and transformation will not be helpful.
#Thus, let's try a fit with logistic regression.


smp_siz=floor(.7*nrow(student.mat.clean)) #70/30 train/test split

set.seed(5167)
train_ind = sample(seq_len(nrow(student.mat.clean)),size = smp_siz)  # Randomly identifies therows
    equal to sample size ( defined in previous instruction) from  all the rows of Smarket dataset
    and stores the row number in train_ind
train =student.mat.clean[train_ind,] #creates the training dataset with row numbers stored in train
    _ind
test=student.mat.clean[-train_ind,]  # creates the test dataset excluding the row numbers mentioned
     in train_ind

train$G3 <- as.numeric(train$G3 <10) #Score under 10 indicates failure
test$G3 <- as.numeric(test$G3 <10)

logisticG3 <- glm(G3 ~ G1 + Improvement+ famrel + goout + health + absences+
                 paid,data=train,family="binomial")

pred <- predict(logisticG3,newdata=test[,c("G1","Improvement",

```

```
                                                                    "famrel","goout","health"
    ,"absences","paid")],type="response")
roc(test$G3,pred)
plot(roc(test$G3,pred),col="blue")
summary(logisticG3)

#Code for G3 ends here.

lmG1null <- lm(G1~1,data=student.mat.clean)
lmG1full <- lm(G1~.-G3-Improvement-absences,data=student.mat.clean)


# Stepwise regression model
step(lmG1null,scope=list(lower=lmG1null, upper=lmG1full),direction="forward")

lmG1step<- lm(formula = G1 ~ failures + schoolsup + Fjob + Walc + Mjob +
                famsup + studytime + sex + goout + health + freetime, data = student.mat.clean)

summary(lmG1step)
plot(lmG1step)

summary(powerTransform(lmstepG3))
#Let's try transforming G1 to G1^2/3 due to the results from
#the Box-Cox power transform likelihood ratio test.

lmG1stepTransform<- lm(formula = G1^.66 ~ failures + schoolsup + Fjob + Walc + Mjob +
                famsup + studytime + sex + goout + health + freetime, data = student.mat.clean)

summary(lmG1step)
plot(lmG1step)

#The transformation did not help the nonconstant variance
#of the residuals. Thus, we will try logistic regression to
#predict G1.

set.seed(13424)
train_indG1 = sample(seq_len(nrow(student.mat.clean)),size = smp_siz)
trainG1 =student.mat.clean[train_indG1,]
testG1 =student.mat.clean[-train_indG1,]
trainG1$G1 <- as.numeric(trainG1$G1 <10)
testG1$G1 <- as.numeric(testG1$G1 <10)

lmG1logistic <- glm(G1 ~.-Improvement-G3, data = trainG1, family = "binomial",maxit=100)
stepAIC(lmG1logistic,trace = FALSE)

lmG1logistic <- glm(formula = G1 ~ AvEdu + failures + schoolsup + famsup + romantic +
                      Dalc + Walc + absences, family = "binomial", data = trainG1,
                    maxit = 100)

pred <- predict(lmG1logistic,newdata=testG1[,c("AvEdu","failures","schoolsup","famsup","romantic","
    Dalc","Walc","absences")],type="response")

summary(lmG1logistic)

roc(testG1$G1,pred)
plot(roc(testG1$G1,pred),col="red")
```