THE GEORGE WASHINGTON UNIVERSITY

Linear Regression Analysis for Student Performance Data
STAT 6214

Baiyang.Qi

ABSTRACT

Math is always important and becomes necessary course throughout whole eaducation period, from kindergarten to college. Since the particularity of math which is coherent course, students need to establish a strong basis of math to make sure that they can understand advanced knowledge in high level grades. This project analyzed the giving data Student Performance on Math from Portuguese schools to find which features have high influence on students math score. Linear regression model is major method. A sequence of models has been made and the best one came from these several models by evaluation criteria.

INTRODUCTION

Education is a key factor for achieving a long-term economic progress. Throughout whole education stage, math always plays an important part among all courses. Students need to learn the math from basic calculation to advanced analysis. Furthermore, math is the necessary term for the most of admission tests, such SAT, GRE, GMAT and so on. However, as the characteristic of math which has high continuity, students need to focus on every stages of math. As a building, if students fail on some basic knowledge of math at first few grades, it will be terrible when they come to upper grades. Not only the students or teachers, but also parents realize that basic mathematics education plays an significant role from kindergarten to high school. They prefer to paying additional money on tutoring for math. And many mathematical training institutions appears, such as Mathnasium and Russian School of Mathematics in Virginia. However, most students have troubles in math leaning. They spend more money and time but the their math does not become better as their imagine.

There are many factors that would influence the students‘ performance on math. In general, these factors can be divided into two categories. The first one is objective factors which are difficult to be changed in the short time, such as home address, parents‘ education, parents‘ job, family size and so on. The other is subjective factors which can be controlled by students and their parients, including study time, traveling time, extra paid classes, after school tutoring and so on.

However, not every factors have siginificant effect on students‘ performance on math. This project aims to find the factors, such as gender, age, address, family background and so on, which have high effect on the performance of students on math and how the significant factors influence the grades of math.

DATASET

In this research, we will analyze recent real-world data from two Portuguese secondary school. The original datasets[1] contain two core classes which are Mathematics and Portuguese. This project only consider the situation of Mathematics class.

In Portugal, the secondary education consists of 3 years of schooling, preceding 9 years of basic eduation and followed by higher education[2]. Most of the students choose to join the free public education system. And there exists several courses, like Science and Technologies, that share core subjects such as the Portuguese Language and Mathematics. In U.S area, the Portuguese Language is not popular in secondary that is why this project only consider about the students‘ performance on math. Like the most European countries, including France or Venezuela, a 20-point grading scale is used, where 0 is the lowest grade and 20 is the highest score.

The dataset of students‘ performance on math has 395 observations and 33 variables totally. The descriptions on detials for each variable display in Table 1 below.

Table 1: The preprocessed student related variables[1]

Attribute	Description
sex	student‘s sex (binary: female or male)
age	student‘s age (numeric: from 15 to 22)
school	student‘s school name
address	student‘s home address type (binary: urban or rural)
Pstatus	Parent‘s cohabitation status (binary: living together or apart)
Medu	mother‘s education (numeric: from 0 to 4)
Mjob	mother‘s job (nominal)
Fedu	father‘s education (numeric: from 0 to 4)
Fjob	father‘s job (nominal)
guardian	student‘s guardian (nominal: mother, father or other)
famsize	family size (binary: <3 or >3)
famrel	quality of family relationship(numeric: from 1-very bad to 5-excellent)
reason	reason to choose this school (nominal: close to home, school reputation, course preference or other)
traveltime	home to school travel time (numeric:1-<15mins; 2- 15 to 30 mins; 3-5 to 10 hours or 4->10 hours)
studytime	weekly study time (numeric: 1- <2 hours; 2- 2 to 5 hours, 3 - 5 to 10 hours, 4- >10 hours)
failures	number of past class failure
schoolsup	extra educational school support (binary: yes or no)
famsup	family educational support (binary: yes or no)
activities	extra-curricular activities (binary: yes or no)
paidclass	extra paid classes (binary: yes or no)
internet	Internet access at home (binary: yes or no)
nursery	attended nursery school (binary: yes or no)
higher	wants to take higher education (binary: yes or no)
romantic	with a romantic relationship (binary: yes or no)
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)
Walc	weekend alcohol consumption (numeric: from 1-very low to 5-very high)
Dalc	workday 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
G1	first period grade (numeric: 0 to 20)
G2	second period grade (numeric: 0 to 20)
G3	final period grade (numeric: 0 to 20)

Among these 33 variables, we need to pay attention to some special points. First, some of these 33 variables may have high correlations. Such parents‘ education and parents‘ job, some

special jobs always require high educational level, like medical doctor or college professor. Furthermore, there are three variables to describ the students‘ performance on math for 0 to 20. The only difference is that G1 is for the first period, G2 is for the second period and G3 is the final grades.

In original data, the variables *Fjob* and *Mjob* are category variables with "at_home", "health", "service", "teacher" and "other". For calculating the correlation, the category variable *Fjob* and *Mjob* are transformed into dummy variables from 1 to 5.

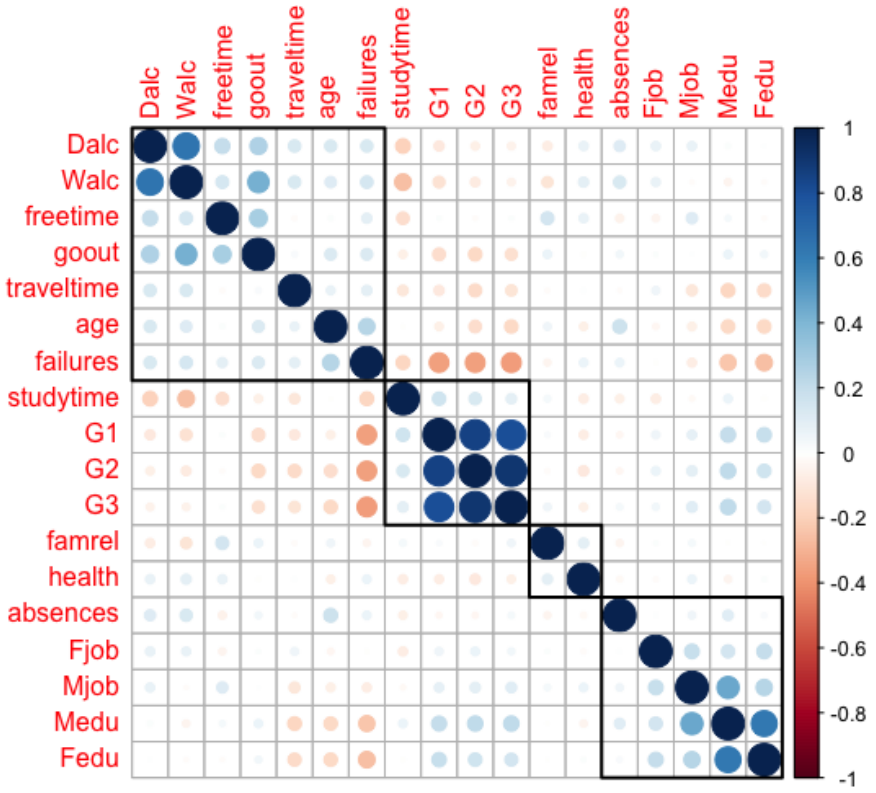


Fig.1: Correlation matrix for numerical variables

The plot Fig.1 above displays the correlations between every numerical variables. As initial guess, parents‘ education and parents‘ job have a little bit high correlation. However, the unexpected point is that the high correlation appears between father‘s job and mother‘ job instead of educational level and job. Another high correlation exists between the variables *Dalc* and *Walc* which are weekend alcohol consumption and workday alcohol consumption. Finally, the highest correlation exists between G1, G2 and G3 which satisfy the preivous analysis. So, this project only considers about the final grades G3 without other two G1 and G2.

METHODOLOGY

The major methods are simple linear regression for whole data set and poisson regression for category variables. For linear regression, several methods like forward and backward of model selection are very useful. Finally, the evaluation criteria is cross validation error for the models we got to get the best one.

LINEAR REGRESSION

The simple linear model is :

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where β_0 is intercept coefficient, $\beta_1 = \frac{\triangle y}{\triangle x}$i s slope coefficient and ϵ is the deviation from points to line. For more than one variable, the definition of linear regession is shown below.
Given a data set $\{y_i, x_{i1}, ..., x_{ip}\}$ of n statistical unit, a linear regression model assumes that the relationship between the dependent variable y_i and the p-vector of regressors $\mathbf{x}_i$ is linear[4].

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \epsilon_i$$

In a linear model the parameters enter linearly, the predictors themselves do not have to be linear[3]. For example:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

is a linear model. But

$$Y = \beta_0 + \beta_1 X_i^{\beta_2} + \epsilon$$

is not.
For convenience, the matrix form is popular to represent the linear regression model.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$; $\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$; $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$: and $\boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$.

Therefore, the problem to fit data a model is reduced to the estimation of parameters in linear regression model. Ordinary least square (OLS) is the simplest and most common estimator. The OLS method minimizes the sum of squared residuals. The expression for estimated unknown parameters $\boldsymbol{\beta}$ is :

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}.$$

POISSON REGRESSION

Poisson regression is a generalized linear model form of regression analysis used to model count data[5]. Poisson regression assumes the response variable Y has a Poisson distribution, and assumes the logarithm of its expected value can be modeled by a linear combination of unknown parameters[5].

The reason to fit the data with Poisson regression model is that there are totally 17 category variables. The Poisson regresion can help us to analyze the influence on each factor. In statistical form, the Poisson regression model are shown below.

If $\boldsymbol{x} \in \mathbb{R}^n$ is a vector of independent variables, then the model takes the form:

$$\log(\boldsymbol{E}(\boldsymbol{Y}|\boldsymbol{x})) = \alpha + \beta' \boldsymbol{x}$$

where $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}^n$. Sometimes this is written more compactly as

$$\log(\boldsymbol{E}(\boldsymbol{Y}|\boldsymbol{x})) = \boldsymbol{\theta}' \boldsymbol{x}$$

where $\boldsymbol{x}$ becomes an $(n+1)$ dimensional vector consisting of n independent variables column binding to a vector of ones. Here $\boldsymbol{\theta}$ is a simply transformation from $\boldsymbol{\beta}$.
Thus, with a Poisson regression model $\boldsymbol{\theta}$ and input vector $\boldsymbol{x}$, the predicted expectation of the associated Poisson distribution is given by :

$$\boldsymbol{E}(\boldsymbol{Y}|\boldsymbol{x}) = e^{\theta' \boldsymbol{x}}.$$

Model Selection

At the beginning, the common idea is to fit a full model which contains every variables with original data. Obviously, the full model always is not good enough. For among 33 variables, not everyone has significant influence on the response, here is the variable *G3*. Model selection becomes more important here, which can reduce the complexity of the model effectively by removing some useless variables. There are many model selection methods and here the stepwise selection will be applied in this project.

In statistics, stepwise selection is a method of fitting regression models in which the choice of predictive variables is carried out by an automatic procedure. Each variable will be considered to add or subtract from the set of explanatory variables based on some presepecified criterion. There are three types of stepwise selection, including forward selection, backward selection and bidirectional selection[3].

Backward Selection Backward selection is the simplest of all variable selection procedures, which involves starting with all variables, testing the deletion of each variable using a chosen model fit criterion, deleting the variables whose loss gives the most statistically insignificant deterioration of the model fit, and repeating this process until no further variables can be deleted without a statistically significant loss of fit[6]

Forward Selection Forward selection just reverses the backward method. It starts with no variables in the model and then for all predictors not in the model. Next step is to check their criteria if they are added to the model and then choose the one with best criteria. The above steps will be repeated until no new predictors can be added[3].

BIDIRECTIONAL SELECTION Bidirectional selection is a combination of the above, testing at each step for variables to be included or excluded[6].

CROSS VALIDATION ERROR

After model selection, there are several models that can be obtained, such as full model, reduced model from forward selection or some models by other modification methods. However, the next problem is how to compare them. The criteria to evaluate models is called cross-validation error to be applied in this project.

CROSS VALIDATION Cross-Validation is a technique used in model selection to better estimate the test error of a predictive model. The idea behind cross-validation is to create a number of partitions of original data set, known as the validation sets. After fitting a model on to the data, its performance is measured against each validation set and then averaged, gaining a better assessment of how the model will perform when asked to predict for new observations. The number of partitions to construct depends on the number of observations in the orginal data set as well as the decision made regarding the bias-variance trade-off, with more partitions leading to a smaller bias but a higher variance[7]. As a exmaple K-fold cross validation, it is the most common use of cross-validation. The original data set is split into K partitions, the model is trained on $K-1$ partitions, and the test error is predicted on the left out partition k. The process is repeated for $k = 1, 2, ..., K$ and the result is averaged.

RESULTS

SIMPLE LINEAR MODEL At the beginnning, the common first step is to fit a full model with whole variables. As introduction above, among three students' performance variables *G1*, *G2* and *G3*, only *G3* is considered in this model. The following plots Fig.2 provide the diagnostic plots of full model. The top-left of Fig.2, Residuals v.s Fitted values plot, shows that the points, looked like parallel lines, are not random around the zero line. Two-tailed of the points on QQ-plot are both under the QQ-line. It is obviout that the full model is not a good option. Next, we can try some transformation on response. After several attempts, such as log-transform, exponent transform or square root transform, the square transformation on response G3 can modify the error term more normalized. Checking the diagnostic plots, Fig.3, of modified model, the Residual and Scale-Location plots look better than the full model. And the QQ plot of Fig.3 displays a better normalization. Then, we need to consider the model selection to reduce the full model.

VARIABLE SELECTION Following is the results after stepwise model selection.
After forward stepwise selection, the formula becomes:
$G3^2$ ~failures + Mjob + schoolsup + Medu + goout + freetime + romantic + famsup + internet + studytime + sex + famsize

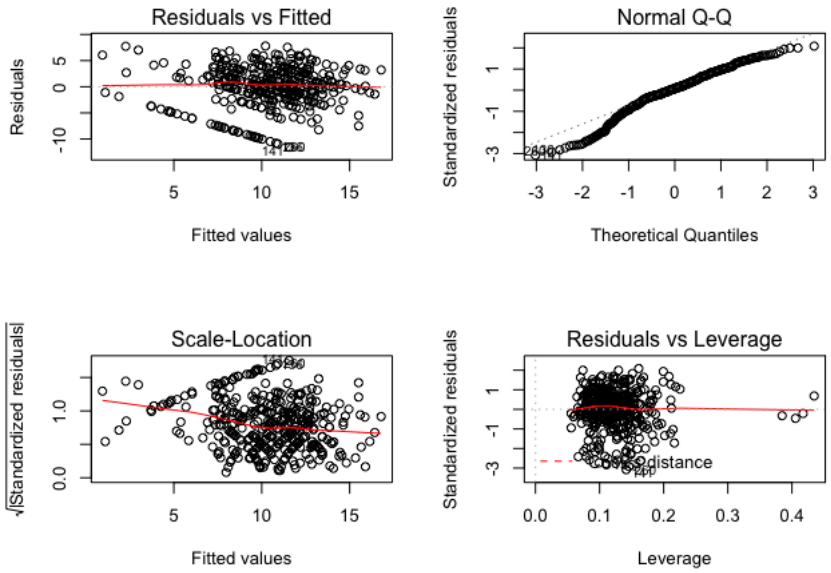


Fig.2: Diagnostic plots of full model

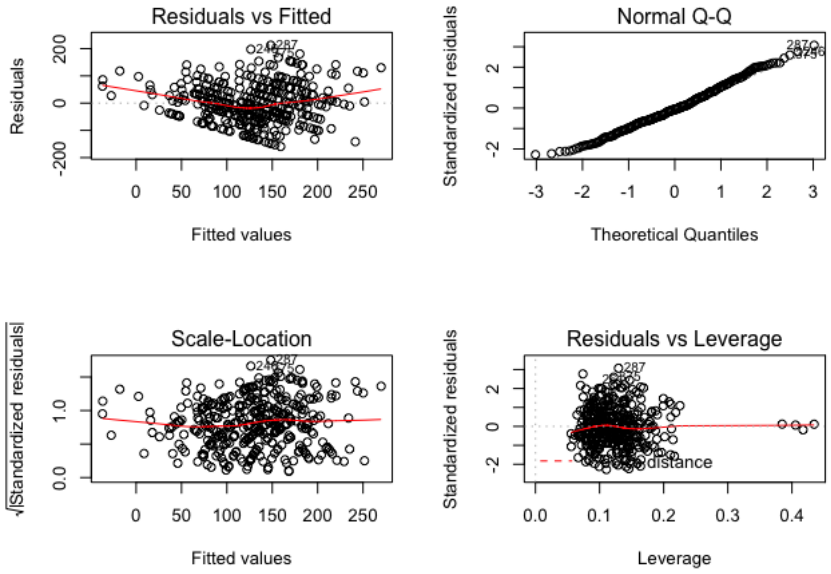


Fig.3: Diagnostic plots of $G3^2$ model

After backward stepwise selection, the formula becomes:
$G3^2$ ~ sex + age + famsize + Medu + Mjob + studytime + failures + schoolsup + famsup + internet + romantic + freetime + goout + health + absences
After bidirection stepwise selection, the formula becomes:
$G3^2$ ~sex + age + famsize + Medu + Mjob + studytime + failures + schoolsup + famsup + internet + romantic + freetime + goout + health

Table.2: Results of model selection

Attribute	Forward	Backward	Bidirection
sex	✓	✓	✓
age		✓	✓
famsize	✓	✓	✓
Medu	✓	✓	✓
Mjob	✓	✓	✓
studytime	✓	✓	✓
failures	✓	✓	✓
schoolsup	✓	✓	✓
famsup	✓	✓	✓
internet	✓	✓	✓
romantic	✓	✓	✓
freetime	✓	✓	✓
goout	✓	✓	✓
health		✓	✓
absences		✓	

For convenience and clarity, the table above, Table.2, provides the results from model selection. It is clear to find the variables that each models contains in detials. Including the full model, there are totaly four model that we had. Which is better becomes next step‘s aim.

Table.3: Cross-Validation Errors

	Full model	Forward Model	Backward Model	Bidirectional Model
CV-Error	6116.473	5936.976	5910.143	5802.723

The table above, Table.3, shows the results of cross validation errors of each model. We can find the model that comes from bidirectional selection has the smallest cross-validation error among these four models. Therefore, until now, the final model we choose is bidirectional selection model whose formula is:
$G3^2$ ~sex + age + famsize + Medu + Mjob + studytime + failures + schoolsup + famsup + internet + romantic + freetime + goout + health

DISCUSSION AND CONCLUSION

According to the analysis above, the final model we choose is the reduced model from bidirectional stepwise selection. The final model contains 14 predictors within 10 category variables.

```
Call:
lm(formula = G3^2 ~ sex + age + famsize + Medu + Mjob + studytime +
    failures + schoolsup + famsup + internet + romantic + freetime +
    goout + health, data = math3)

Residuals:
     Min       1Q   Median       3Q      Max
-174.127  -51.213   -6.051   43.538  236.257

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   282.697     77.640   3.641  0.00031 ***
sexM           15.327      8.431   1.818  0.06987 .
age            -5.978      3.275  -1.825  0.06880 .
famsizeLE3     12.041      8.415   1.431  0.15330
Medu1         -78.391     44.132  -1.776  0.07651 .
Medu2         -74.779     44.271  -1.689  0.09204 .
Medu3         -60.632     44.529  -1.362  0.17414
Medu4         -33.689     45.620  -0.738  0.46071
Mjobhealth      9.927     19.260   0.515  0.60656
Mjobother      -9.236     12.311  -0.750  0.45357
Mjobservices   17.700     13.958   1.268  0.20556
Mjobteacher   -34.826     18.452  -1.887  0.05989 .
studytime      10.565      4.853   2.177  0.03011 *
failures      -30.428      5.571  -5.462 8.69e-08 ***
schoolsupyes  -37.555     11.911  -3.153  0.00175 **
famsupyes     -15.743      8.108  -1.942  0.05296 .
internetyes    15.625     10.716   1.458  0.14566
romanticyes   -18.406      8.270  -2.226  0.02664 *
freetime2      16.991     19.838   0.857  0.39227
freetime3      -6.146     18.700  -0.329  0.74261
freetime4       8.658     19.265   0.449  0.65342
freetime5      41.682     21.850   1.908  0.05722 .
goout2         21.446     17.588   1.219  0.22347
goout3          9.040     17.410   0.519  0.60391
goout4        -12.497     17.956  -0.696  0.48688
goout5        -18.165     19.050  -0.954  0.34095
health         -5.025      2.793  -1.799  0.07281 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 73.5 on 368 degrees of freedom
Multiple R-squared:  0.3025,    Adjusted R-squared:  0.2532
F-statistic: 6.139 on 26 and 368 DF,  p-value: < 2.2e-16
```

Fig.4: Statistical Summary of Final Model

Fig.4 shows the statistical summary of the final model. From this summary, every numerical variables, inlcuding *age*, *studytime*, *failures* and *health*, are significant. For the category variables, at least one factor of these variables is statistical significant for the response. Then, for the estimated parameters, there are few interesting things here. For the variable of *mother's education*, all the factors have a negative influence on the students' performance. More *studytime* meaning more score on math is reasonable. But for the variable *goout*, which means going out with friends, it is positve influence at beginning but becomes nagetive with increassing. The explanation may be that making friends is good for students get better math score but if students spend lots of time on friends, they do not have enough time on study and the score will become worse.

As introduction part said, the variables can be divided into two parts, objective variable and subjective variable. So, the final model has been divided into two models. One contains the objective variables only. And the other only has subjective variables.

```
Call:
lm(formula = G3^2 ~ sex + age + Medu + Mjob + famsize + health,
    data = math3)

Residuals:
     Min       1Q   Median       3Q      Max
-165.046  -54.070   -8.216   49.666  262.390

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   347.633     74.809   4.647 4.64e-06 ***
sexM           18.606      8.434   2.206   0.0280 *
age            -8.863      3.251  -2.726   0.0067 **
Medu1         -92.619     47.875  -1.935   0.0538 .
Medu2         -76.099     47.703  -1.595   0.1115
Medu3         -68.032     48.022  -1.417   0.1574
Medu4         -38.315     48.704  -0.787   0.4320
Mjobhealth     21.994     20.185   1.090   0.2766
Mjobother      -1.944     13.165  -0.148   0.8827
Mjobservices   20.467     14.684   1.394   0.1642
Mjobteacher   -13.666     19.425  -0.704   0.4822
famsizeLE3     12.688      9.120   1.391   0.1650
health         -6.700      3.007  -2.228   0.0264 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 80.65 on 382 degrees of freedom
Multiple R-squared:  0.1282,    Adjusted R-squared:  0.1008
F-statistic:  4.68 on 12 and 382 DF,  p-value: 3.99e-07
```

Fig 5.1: Summary of Objective Model

```
Call:
lm(formula = G3^2 ~ studytime + failures + schoolsup + famsup +
    internet + romantic + freetime + goout, data = math3)

Residuals:
    Min      1Q  Median      3Q     Max
-165.24  -48.77   -8.59   48.70  215.13

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   107.548     25.511   4.216 3.11e-05 ***
studytime       8.089      4.841   1.671  0.09556 .
failures      -34.603      5.418  -6.386 4.97e-10 ***
schoolsupyes  -33.982     11.828  -2.873  0.00429 **
famsupyes     -11.704      8.142  -1.437  0.15143
internetyes    26.753     10.576   2.530  0.01182 *
romanticyes   -21.967      8.319  -2.641  0.00862 **
freetime2      26.598     20.553   1.294  0.19642
freetime3      -4.293     19.298  -0.222  0.82406
freetime4      13.324     19.870   0.671  0.50292
freetime5      49.409     22.285   2.217  0.02721 *
goout2         17.668     17.998   0.982  0.32689
goout3          8.583     17.959   0.478  0.63298
goout4        -13.816     18.488  -0.747  0.45534
goout5        -17.364     19.695  -0.882  0.37852
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 76.7 on 380 degrees of freedom
Multiple R-squared:  0.2157,    Adjusted R-squared:  0.1868
F-statistic: 7.466 on 14 and 380 DF,  p-value: 7.355e-14
```

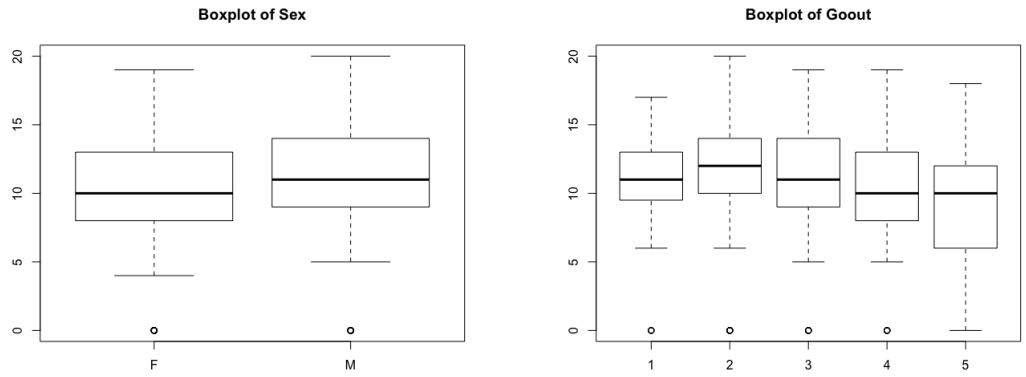
Fig 5.2: Summary of Subjective Model

Fig 5.1 and Fig 5.2 provide the statistical summary of objective model and subjective model. Objective variables are difficult to change in the short time so we focus on the subjective model firstly. From the summary of subjective model, the positive variables are *studytime*,*internet* and *freetime*. As a example the variable *studytime*, more time on math learning implies better grades. However, for the variable *goout*, it is a little bit complicate since some factors are posive and the other are negative. And some social opinions publicize the difference between boys and girls. So, next are the Poisson regression on *sex* and *goout*. According to the summary,Fig 6, and boxplots below, the *sex* are significant and the estimated values are similar. Comparing the boxplot of sex, we can find that even *sex* is statistical significant for the model, the difference

between boys and girls is not huge enough.

```
Call:
glm(formula = G3 ~ sex + goout - 1, family = poisson, data = math3)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.8593  -0.5400   0.1301   0.8917   2.8253

Coefficients:
       Estimate Std. Error z value Pr(>|z|)
sexF    2.24904    0.06763  33.253   <2e-16 ***
sexM    2.34925    0.06885  34.122   <2e-16 ***
goout2  0.11938    0.07264   1.643    0.100
goout3  0.09785    0.07150   1.369    0.171
goout4 -0.02982    0.07494  -0.398    0.691
goout5 -0.10746    0.08081  -1.330    0.184
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 13001.5  on 395  degrees of freedom
Residual deviance:  1123.6  on 389  degrees of freedom
AIC: 2655.1

Number of Fisher Scoring iterations: 5
```

Fig.6: Statistical Summary of Poisson Regression



To conclude, from the results of final models, we have some strategies for parents and students separately. For parents, we suggest that parents can provide Internet for their children. Students can find some tutoring or solution online. For students, we suggest that students can spend more time on math instead of with friend or raomantic relationship. The total time is

limitation. If other activities occupy a little bit more time, students will not have enough time to focus on study.

REFERENCES

[1] Student Performance
http://archive.ics.uci.edu/ml/machine- learning-databases/00320/student.zip

[2] Paulo Cortez , Alice Silva *USING DATA MINING TO PREDICT SECONDARY SCHOOL STUDENT PERFORMANCE*

[3] Julian J. Faraway *Linear Models with R, Second Edition*

[4] Linear regression
https://en.wikipedia.org/wiki/Linear_regression

[5] Poisson Regression
https://en.wikipedia.org/wiki/Poisson_regression

[6] Stepwise Selection
https://en.wikipedia.org/wiki/Stepwise_regression

[7] Cross-Validation: Estimating Prediction Error
https://www.r-bloggers.com/cross-validation-estimating-prediction-error/

[8] Hadley Wickham, Garrett Grolemund *R for Data Science*

[9] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani *An introduction to Statistical Learning with Application in R*

APPENDIX

```
math<-read.csv("~/desktop/student/student-mat.csv",sep=";")
math1<-math[,-c(32,33)]
math2<-math[,-c(31,33)]
math3<-math[,-c(31,32)]
math3$goout<-as.factor(math3$goout)
math3$Medu<-as.factor(math3$Medu)
math3$freetime<-as.factor(math3$freetime)
###Begin simple linear regression
lmod<-lm(G3~.,data=math3)
summary(lmod)
par(mfrow=c(2,2))
plot(lmod)
###transformation on G3 boxcox
lmod1<-lm(G3^2~.,data=math3)
par(mfrow=c(2,2))
plot(lmod1)
summary(lmod1)
###valid variables
####forward stepwise
full<-lm(G3^2~.,data=math3)
null<-lm(G3^2~1,data=math3)
forward<-stepAIC(null,scope=formula(full),direction="forward",trace=T)
summary(forward)
backward<-stepAIC(full,scope=formula(null),direction="backward",trace=T)
summary(backward)
both<-stepAIC(full,direction="both")
summary(both)
plot(forward)
plot(backward)
formula(forward)
formula(backward)
###compare
library(boot)
mod1<-glm(G3^2~.,data=math3)
for_mod<-glm(formula(forward),data=math3)
back_mod<-glm(formula(backward),data=math3)
both_mod<-glm(formula(both),data=math3)
cv.glm(data=math3,glmfit=mod1,K=10)$delta[2]
cv.glm(data=math3,glmfit=for_mod,K=10)$delta[2]
cv.glm(data=math3,glmfit=back_mod,K=10)$delta[2]
cv.glm(data=math3,glmfit=both_mod,K=10)$delta[2]
```

```
formula(forward)
formula(backward)
formula(both)
###forward performance best
###Riger regression
library(glmnet)
attach(math3)
X<-model.matrix(~ failures + Mjob + schoolsup + goout + Medu + sex +
    romantic +
health + studytime + famsup + age + Fjob + famsize + internet-1,data=
    math3)
Y<-as.matrix((math3[c("G3")])^2)
ridge <- cv.glmnet(x = X, y = Y, alpha=0, lambda = seq(0,10,length.out =
     100))
ridge$lambda.min
ridge$lambda.1se



lasso<-cv.glmnet(x = X, y = Y, alpha=1, lambda = seq(0,10,length.out =
    100))
lasso$lambda.min
lasso$lambda.1se
summary(lasso)
lasso$glmnet.fit

par(mfrow=c(1,1))
plot(sex,G3)
poi<-glm(G3~sex-1,data=math3,family = poisson)
summary(poi)
beta<-c(poi$coefficients[1],sum(poi$coefficients))
beta
exp(beta)

new_model<-glm(G3^2 ~ failures + Mjob + schoolsup + goout + Medu + sex +
     romantic +
health + studytime + famsup + age + Fjob + famsize + internet,data=math3
    )
summary(new_model)

plot(Mjob,G3)
plot(schoolsup,G3)
plot(math3$Medu,math3$G3)
poi2<-glm(G3~Medu-1,data=math3,family = poisson)
```

```
summary(poi2)
exp(poi2$coefficients)



plot(math3$famsup,G3)
poi3<-glm(G3~famsup-1,data=math3,family = poisson)
summary(poi3)



plot(Fjob,G3)
plot(famsize,G3)
plot(math3$freetime,G3)


library(corrplot)
math$Mjob <- as.numeric(math$Mjob)
math$Fjob<-as.numeric(math$Fjob)
num<-math[,c(3,7,8,9,10,13,14,15,24,25,26,27,28,29,30,31,32,33)]
M<-cor(num)
corrplot(M,order = "hclust",addrect = 3)
help("stepAIC")

sex_lm<-lm(G3^2~sex-1,data=math3)
summary(sex_lm)

objective<-lm(G3^2~sex+age+Medu+Mjob+famsize+health,data=math3)
subjective<-lm(G3^2~studytime+failures+schoolsup+famsup+internet+
    romantic+freetime+goout,data=math3)
summary(objective)
plot(objective)
summary(subjective)


poi1<-glm(G3~sex-1,data=math3,family = poisson)
poi2<-glm(G3~goout-1,data=math3,family = poisson)
poi3<-glm(G3~sex+goout-1,data=math3,family = poisson)
summary(poi1)
summary(poi2)
summary(poi3)
plot(poi1)

poi4<-glm(G3~Medu+sex+Mjob-1,data=math3,family = poisson)
```

```
summary(poi4)
summary(poi1)
par(mfrow=c(1,1))
plot(math3$sex,math3$G3,main="Boxplot␣of␣Sex")
plot(math3$goout,math3$G3,main="Boxplot␣of␣Goout")
```