**Main**

Step1: Exploratory Analysis

(a)Deal with the missing values

> table(is.na(data))

FALSE TRUE

25379 4651

Math9 Math5 Eng5 English9 MP\_ADD\_5 MP\_ADD\_9

188 1 1 1 1 1 1 0

70 1 1 1 1 1 0 1

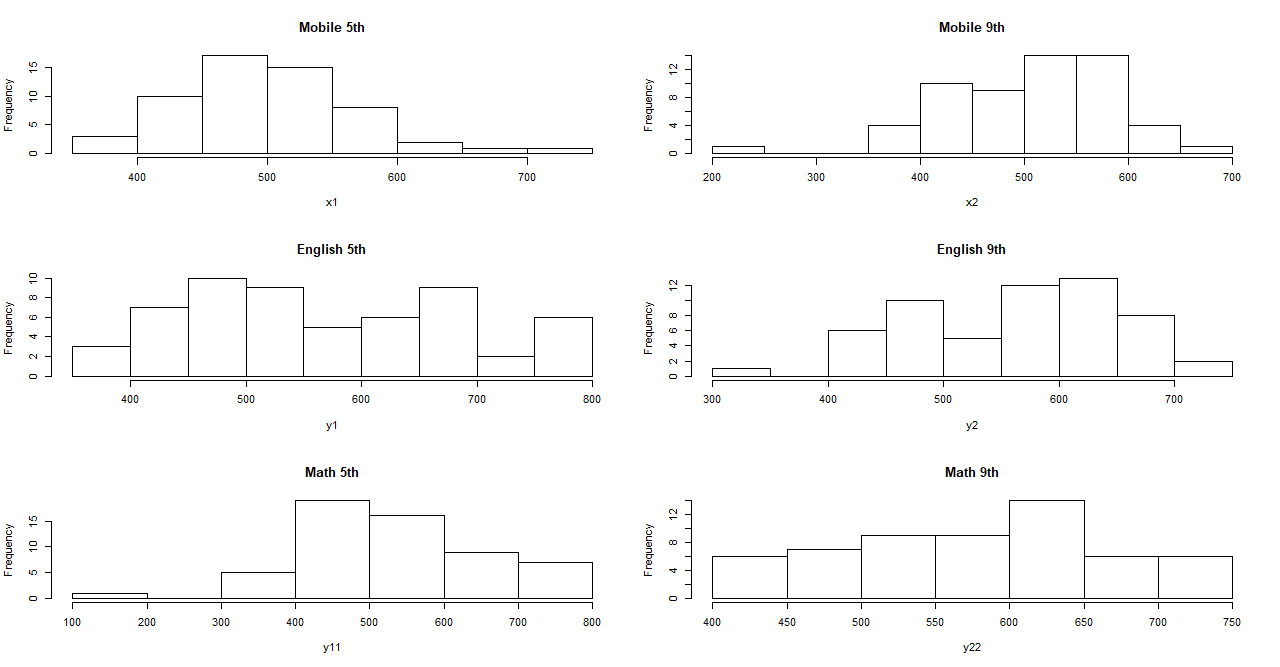
132 1 1 1 1 0 0 2

0 0 0 0 132 202 334

By using the table function, it is obvious that there are 4651 missing values which are supposed to omit.

From above matrix, there are no missing value in the Math or English scores but 132 in 5th mobile addiction students’ scores and 202 in 9th  mobile addiction students’ scores.

(b) Create variables and study the distribution of the test scores variable



It is easily to find out that the Mobile 5th test scores student dataset might follow a normal distribution despite a little right skewness.

For the Mobile 9th test scores student dataset, it does not follow a normal distribution due to high left skewness.

For the English 5th test scores student dataset, it does not follow a normal distribution.

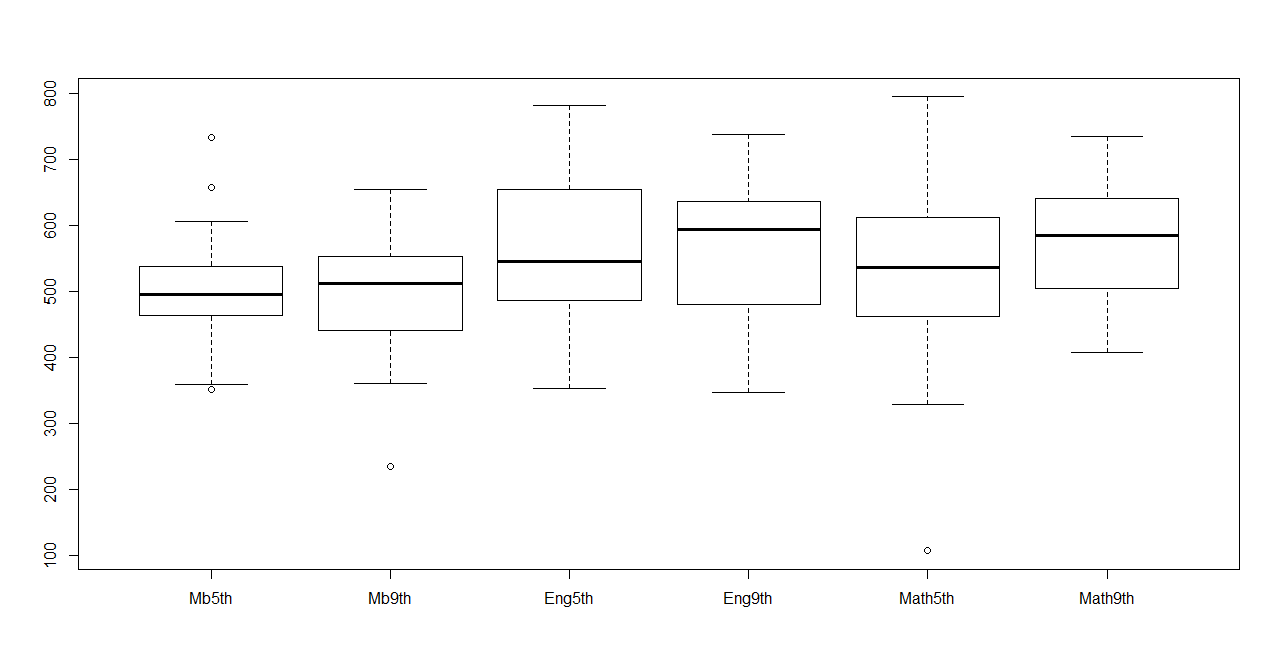
For the English 9th scores student dataset, it does not follow a normal distribution due to little left skewness.

For the Math 5th scores student dataset, it does not follow a normal

distribution due to high left skewness.

For the Math 9th scores dataset, it may follow a normal distribution.

(c)Check Outliers



It is obvious that there are 3 outliers in Mobile 5th dataset, one outlier in Mobile 9th dataset and one outlier in Math 5th dataset. We need to do further survey about the weird data to check the errors.

(d) Additional analysis

The 9th students who are addicted in mobile phones have higher mean scores than 5th students. Then we can assume that the senior students seem to have more self-control to get higher scores.

The 9th Students also have higher mean scores in English and Math than the 5th Students. We can assume that the senior students have better ability to get better performance in their academic studies.

The mean English scores of 5th students have higher differences from the mobile addiction students scores than the mean scores of Math. This phenomenon also happened in 9th students, the 9th students have higher average scores than the 5th students in both subjects. And the mean scores difference between math and mobile addiction, English and mobile addiction keep the same trend as 5th students. Finally, we can assume that the mobile phone may have less influence on Math than English.

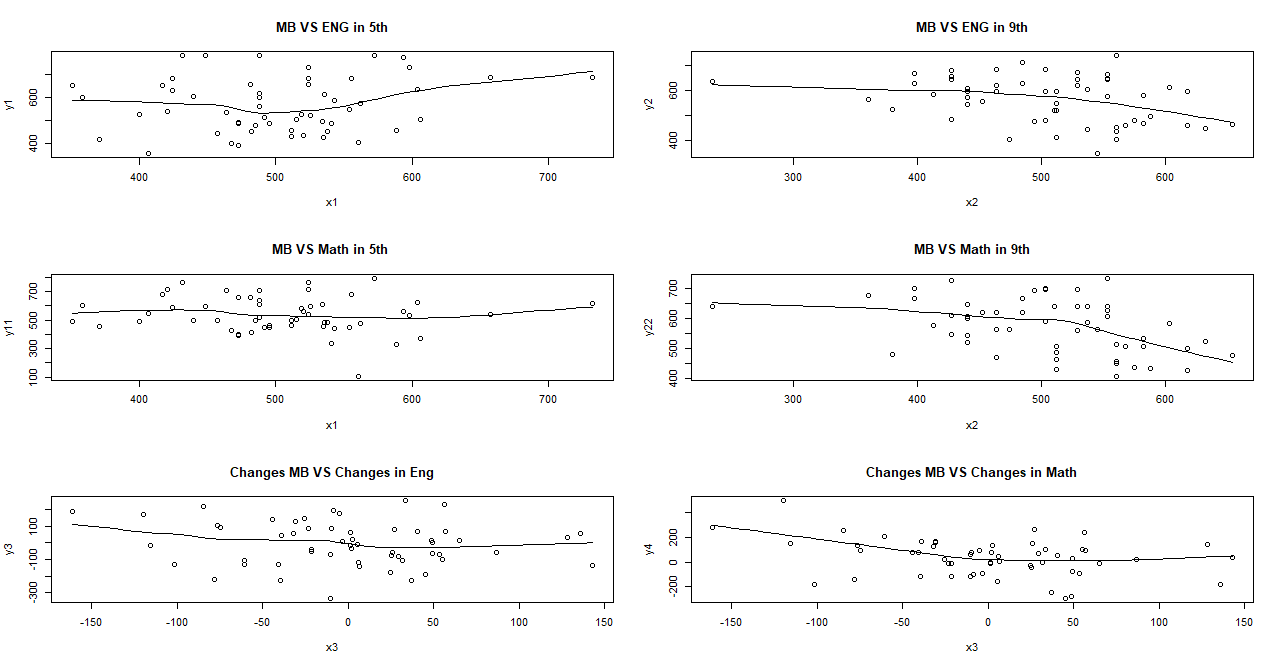
Step2: Regression Analysis

(a) Initial analysis

From step 1, we obtain that the 9th students who are addicted to Mobile phone have higher scores than the 5th students. Then the question arises: Is the Mobile addiction students’ scores have relationship with their ages and the subjects? Is it a positive linear relationship or a negative relationship?

What is the different influence the mobile phone have on English and Math?

What is their relationship?



Assume that there only exists linear relationship. Then 6 general model proposed.

Model 1: English5th~MP\_ADD\_5

Maybe there exists positive linear relationship.

The coefficient of independent variable may be bigger than 0 and smaller than 1

> cor(x1,y1)

[1] 0.1343989

It fits to the previous expectation.

Model 2: English9th~MP\_ADD\_9

Maybe there exists negative linear relationship.

The coefficient of independent variable may be smaller than 0

> cor(x2,y2)

[1] -0.3189546

It fits to the previous expectation.

Model3:Math5th~MP\_ADD\_5

Maybe there exists positive linear relationship

The coefficient of independent variable may be bigger than 0 and smaller than 1.

> cor(x1,y11)

[1] -0.119507

It fits to the previous expectation.

Model4:Math9th~MP\_ADD\_9

Maybe there exists negative linear relationship.

The coefficient of independent variable may be smaller than 0

> cor(x2,y22)

[1] -0.3994082

It fits to the previous expectation.

Model5:(English9th-English5th) ~(MP\_ADD\_9-MP\_ADD\_5)

Maybe there exists negative linear relationship.

The coefficient of independent variable may be smaller than 0

> cor(x3,y3)

[1] -0.1418862

It fits to the previous expectation.

Model6:(Math9th-Math5th) ~(MP\_ADD\_9-MP\_ADD\_5)

Maybe there exists negative linear relationship.

The coefficient of independent variable may be smaller than 0

> cor(x3,y4)

[1] -0.3119246

It fits to the previous expectation.

(b)Model Estimated

x1=data$MP\_ADD\_5

x2=data$MP\_ADD\_9

y1=data$ENG5

y2=data$ENG9

y11=data$MATH5

y22=data$MATH9

y3=y2-y1

x3=x2-x1

y4=y22-y11

Model 1:

y1= 0.2237x1+ 456.6742

The P value of estimated x1 is 0.318890 which is far bigger than 0.05, then we can conclude that x1 is not statistic important to y1. In other words, there is no obvious influence of mobile addiction on the 5th students’ English scores.

Adjusted R-squared: 0.0002097

It means that nearly no variation in English Scores of 5th students are explained by the Mobile phone addiction students’ scores. We need to seek other variables to fit the model.

Model 2:

y2= -0.3788x2+ 755.5361

The P value of estimated x2 is 0.0156 which is less than 0.05, then we can conclude that x2 is statistic important to y2. In other words, there is a negative linear relationship between mobile addiction students’ scores and English scores of 9th students.

Adjusted R-squared: 0.0854

It means a little variation in English scores of 9th students are explained. But we still need to seek extra variables to fit the model.

Model3:

y11= -0.2081x1+ 643.2812

The P value of estimated x1 is 0.376 which is far more than 0.05, we can conclude that x1 is not statistic important to y11. In other words, there is no obvious influence of mobile addiction on the 5th students’ Math scores.

Adjusted R-squared: -0.00364

It means that nearly no variation in Math Scores of 5th students are explained by the Mobile phone addiction students’ scores. We need to seek other variables to fit the model.

Model 4:

y22= -0.4770x2+ 818.1573

The P value of estimated x2 is 0.00208 which is far less than 0.05, then we can conclude that x2 is statistic important to y22. In other words, there is a negative linear relationship between mobile addiction students’ scores and Math scores of 9th students.

Adjusted R-squared: 0.1442

It means a part variation in Math scores of 9th students are explained. But we still need to seek extra variables to fit the model.

Model 5:

y3= -0.2924x3+ -3.2874

The P value of the estimated x3 is 0.292 which is far more than 0.05, then we can conclude that x3 is not statistic significant to y33. In other words, there is no relationship between the change of Mobile phone addiction and change of English scores in 5th and 9th.

Adjusted R-squared: 0.002316

It means that nearly no variation in changes of English scores in 5th and 9th are explained by the change of Mobile phone addiction students’ scores. We need to seek other variables to fit the model.

Model 6:

y4= -0.7371x3+ 40.4144

The P value of the estimated x3 is 0.0182 which is less than 0.05, then we can conclude that x3 is statistic significant to y33. In other words, there is negative linear relationship between the change of Mobile phone addiction and change of Math scores in 5th and 9th.

Adjusted R-squared: 0.08088

It means that a little variation in changes of Math scores in 5th and 9th are explained by the change of Mobile phone addiction students’ scores. We need to seek extra variables to fit the model.

(c)Summary and problem analysis

Summary:

From part (b) we can easily get that the mobile addiction influences the performance of the 9th students a lot while has no influence on 5th students.

Besides, it can conclude that the mobile addiction has no influence on English scores but very obvious influence on Math scores.

Problem:

The common problem is that the models are not good enough. All the models are supposed to have more variables to explain the variation of the associated dependent variables. In the further research, it requires to consider more factors that influence the scores of the students. What’s more, there are too many missing values and with deleting of this missing value, there is not enough observation in our model. Therefore, we need to do further survey to have more detailed data.