Stat 402 Project 1

Prakul Asthana

About the Dataset

This data approach student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features) and it was collected by using school reports and questionnaires. Two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por). In [Cortez and Silva, 2008], the two datasets were modeled under binary/five-level classification and regression tasks. Important note: the target attribute G3 has a strong correlation with attributes G2 and G1. This occurs because G3 is the final year grade (issued at the 3rd period), while G1 and G2 correspond to the 1st and 2nd period grades. It is more difficult to predict G3 without G2 and G1, but such prediction is much more useful (see paper source for more details).

Codebook

Attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course) datasets:

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 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education)

8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ 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)

these grades are related with the course subject, Math or Portuguese:

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)

Source

UCI - https://archive.ics.uci.edu/ml/datasets/Student+Performance Paper - P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

Loading dataset

```
math_data = read.table("student-mat.csv",sep=";",header=TRUE)
lang_data = read.table("student-por.csv",sep=";",header=TRUE)
```

Some ordinal variables are being read as numerical, changing them to factors

```
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------------- tidyverse 1.2.1 --
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
## -- Conflicts ---------------------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
ordinal_var = c("Medu","Fedu","traveltime","studytime","famrel","freetime","goout","Dalc","Walc","healt
lang_data <- lang_data %>% mutate_if(colnames(lang_data) %in% ordinal_var, as.factor)
math_data <- math_data %>% mutate_if(colnames(math_data) %in% ordinal_var, as.factor)
```

Exploring Math data

```
summary(math_data)
```

```
##  school   sex          age       address famsize   Pstatus Medu    Fedu
##  GP:349   F:208   Min.   :15.0   R: 88   GT3:281   A: 41   0:  3   0:  2
##  MS: 46   M:187   1st Qu.:16.0   U:307   LE3:114   T:354   1: 59   1: 82
##                   Median :17.0                             2:103   2:115
##                   Mean   :16.7                             3: 99   3:100
##                   3rd Qu.:18.0                             4:131   4: 96
##                   Max.   :22.0
##        Mjob           Fjob            reason      guardian   traveltime
##  at_home : 59   at_home : 20   course     :145   father: 90   1:257
##  health  : 34   health  : 18   home       :109   mother:273   2:107
##  other   :141   other   :217   other      : 36   other : 32   3: 23
##  services:103   services:111   reputation:105                4:  8
##  teacher : 58   teacher : 29
##
##  studytime failures schoolsup famsup     paid     activities nursery
##  1:105     0:312    no :344   no :153   no :214   no :194    no : 81
##  2:198     1: 50    yes: 51   yes:242   yes:181   yes:201    yes:314
##  3: 65     2: 17
##  4: 27     3: 16
##
##
##  higher    internet  romantic  famrel  freetime goout   Dalc    Walc
##  no : 20   no : 66   no :263   1:  8   1: 19    1: 23   1:276   1:151
##  yes:375   yes:329   yes:132   2: 18   2: 64    2:103   2: 75   2: 85
##                                3: 68   3:157    3:130   3: 26   3: 80
##                                4:195   4:115    4: 86   4:  9   4: 51
##                                5:106   5: 40    5: 53   5:  9   5: 28
##
##  health     absences            G1              G2              G3
##  1: 47   Min.   : 0.000   Min.   : 3.00   Min.   : 0.00   Min.   : 0.00
##  2: 45   1st Qu.: 0.000   1st Qu.: 8.00   1st Qu.: 9.00   1st Qu.: 8.00
##  3: 91   Median : 4.000   Median :11.00   Median :11.00   Median :11.00
##  4: 66   Mean   : 5.709   Mean   :10.91   Mean   :10.71   Mean   :10.42
##  5:146   3rd Qu.: 8.000   3rd Qu.:13.00   3rd Qu.:13.00   3rd Qu.:14.00
##          Max.   :75.000   Max.   :19.00   Max.   :19.00   Max.   :20.00
```

Exploring Language data

```
summary(lang_data)
```

```
##  school   sex          age        address famsize   Pstatus Medu
##  GP:423   F:383   Min.   :15.00   R:197   GT3:457   A: 80   0:  6
##  MS:226   M:266   1st Qu.:16.00   U:452   LE3:192   T:569   1:143
##                   Median :17.00                             2:186
##                   Mean   :16.74                             3:139
##                   3rd Qu.:18.00                             4:175
##                   Max.   :22.00
##  Fedu          Mjob           Fjob            reason      guardian
##  0:  7   at_home :135   at_home : 42   course     :285   father:153
##  1:174   health  : 48   health  : 23   home       :149   mother:455
```

```
##  2:209   other   :258   other   :367   other     : 72   other : 41
##  3:131   services:136   services:181   reputation:143
##  4:128   teacher : 72   teacher : 36
##
##  traveltime studytime failures schoolsup famsup     paid     activities
##  1:366      1:212     0:549    no :581   no :251   no :610   no :334
##  2:213      2:305     1: 70    yes: 68   yes:398   yes: 39   yes:315
##  3: 54      3: 97     2: 16
##  4: 16      4: 35     3: 14
##
##
##  nursery   higher    internet  romantic  famrel  freetime goout   Dalc
##  no :128   no : 69   no :151   no :410   1: 22   1: 45    1: 48   1:451
##  yes:521   yes:580   yes:498   yes:239   2: 29   2:107    2:145   2:121
##                                          3:101   3:251    3:205   3: 43
##                                          4:317   4:178    4:141   4: 17
##                                          5:180   5: 68    5:110   5: 17
##
##  Walc    health     absences            G1             G2
##  1:247   1: 90   Min.   : 0.000   Min.   : 0.0   Min.   : 0.00
##  2:150   2: 78   1st Qu.: 0.000   1st Qu.:10.0   1st Qu.:10.00
##  3:120   3:124   Median : 2.000   Median :11.0   Median :11.00
##  4: 87   4:108   Mean   : 3.659   Mean   :11.4   Mean   :11.57
##  5: 45   5:249   3rd Qu.: 6.000   3rd Qu.:13.0   3rd Qu.:13.00
##                  Max.   :32.000   Max.   :19.0   Max.   :19.00
##        G3
##  Min.   : 0.00
##  1st Qu.:10.00
##  Median :12.00
##  Mean   :11.91
##  3rd Qu.:14.00
##  Max.   :19.00
```

Cheecking datatypes

```
data.frame(unlist(lapply(lang_data, class)))
```

```
##            unlist.lapply.lang_data..class..
## school                                factor
## sex                                   factor
## age                                  integer
## address                               factor
## famsize                               factor
## Pstatus                               factor
## Medu                                  factor
## Fedu                                  factor
## Mjob                                  factor
## Fjob                                  factor
## reason                                factor
## guardian                              factor
## traveltime                            factor
## studytime                             factor
## failures                              factor
## schoolsup                             factor
```

```
## famsup                                 factor
## paid                                   factor
## activities                             factor
## nursery                                factor
## higher                                 factor
## internet                               factor
## romantic                               factor
## famrel                                 factor
## freetime                               factor
## goout                                  factor
## Dalc                                   factor
## Walc                                   factor
## health                                 factor
## absences                              integer
## G1                                    integer
## G2                                    integer
## G3                                    integer
```

```
data.frame(unlist(lapply(math_data, class)))
```

```
##            unlist.lapply.math_data..class..
## school                              factor
## sex                                 factor
## age                                integer
## address                             factor
## famsize                             factor
## Pstatus                             factor
## Medu                                factor
## Fedu                                factor
## Mjob                                factor
## Fjob                                factor
## reason                              factor
## guardian                            factor
## traveltime                          factor
## studytime                           factor
## failures                            factor
## schoolsup                           factor
## famsup                              factor
## paid                                factor
## activities                          factor
## nursery                             factor
## higher                              factor
## internet                            factor
## romantic                            factor
## famrel                              factor
## freetime                            factor
## goout                               factor
## Dalc                                factor
## Walc                                factor
## health                              factor
## absences                           integer
## G1                                 integer
## G2                                 integer
## G3                                 integer
```

Checing correlation of G1,G2,G3

```
print("Maths Scores")
```

```
## [1] "Maths Scores"
```

```
cor(math_data[c("G1","G2","G3")])
```

```
##           G1        G2        G3
## G1 1.0000000 0.8521181 0.8014679
## G2 0.8521181 1.0000000 0.9048680
## G3 0.8014679 0.9048680 1.0000000
```

```
print("Language Scores")
```

```
## [1] "Language Scores"
```

```
cor(lang_data[c("G1","G2","G3")])
```

```
##           G1        G2        G3
## G1 1.0000000 0.8649816 0.8263871
## G2 0.8649816 1.0000000 0.9185480
## G3 0.8263871 0.9185480 1.0000000
```

```
print("As G3 is final grade and G1, G2 are term one and two grades")
```

```
## [1] "As G3 is final grade and G1, G2 are term one and two grades"
```

```
print("We calculate finalgrade = (G1 + G2 + 2*G3)/4")
```

```
## [1] "We calculate finalgrade = (G1 + G2 + 2*G3)/4"
```

```
math_data$outcome <- (math_data$G1 + math_data$G2 + 2*math_data$G3)/4
lang_data$outcome <- (lang_data$G1 + lang_data$G2 + 2*lang_data$G3)/4

# Remove G1,G2,G3 data
math_data <- subset(math_data, select = -c(G1,G2,G3))
lang_data <- subset(lang_data, select = -c(G1,G2,G3))
```

Checking distributions variable wise

Gender vs School

```
print("Math Data distribution")
```

```
## [1] "Math Data distribution"
```

```
table(math_data$sex,math_data$school)
```

```
##
##      GP  MS
##   F 183  25
##   M 166  21
```

```
print("Language data distribution")
```

```
## [1] "Language data distribution"
```

```
table(lang_data$sex,lang_data$school)
```

```
##
##      GP  MS
##   F 237 146
##   M 186  80
```

Checking distrubtions of numerical variables

Age

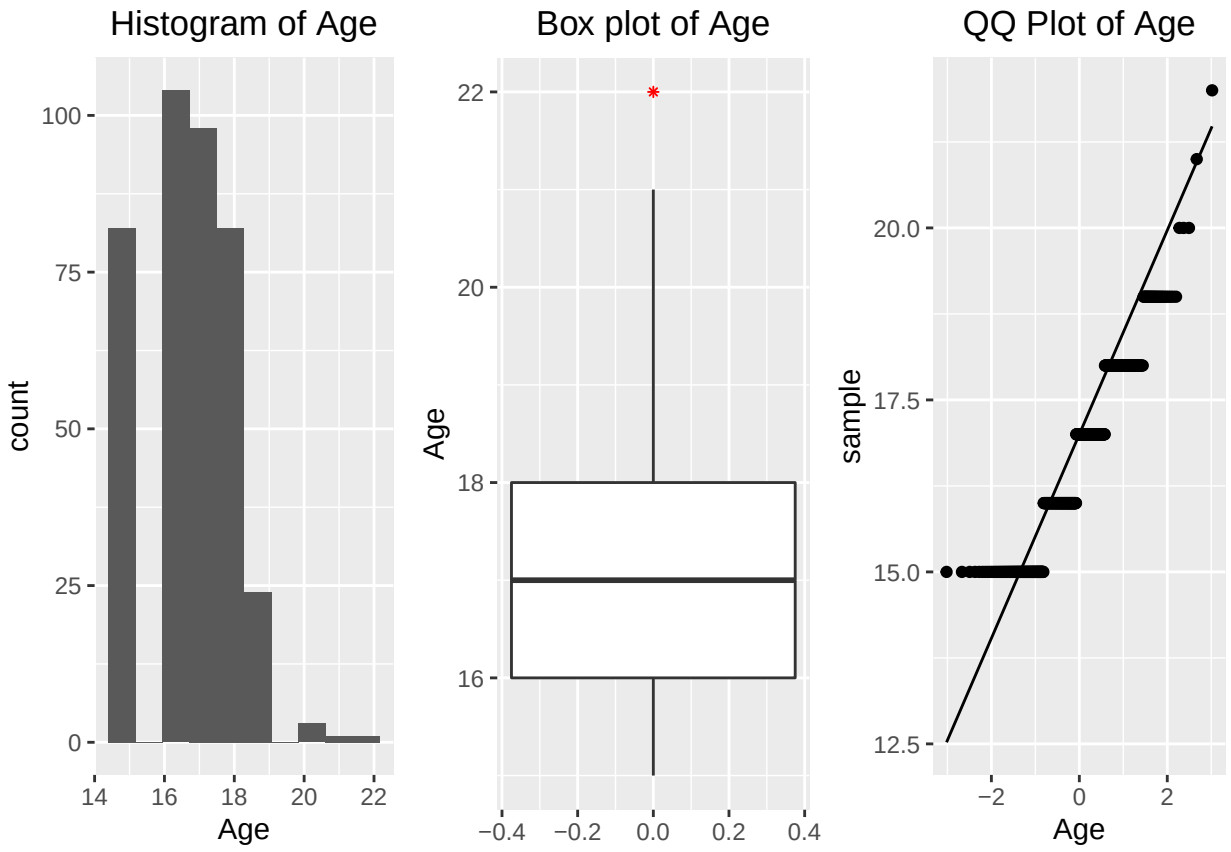
Math

```
library(ggplot2)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
histogram <- ggplot(math_data, aes(x=age))  + geom_histogram(bins = 10) + labs(title="Histogram of Age"

box_plot <- ggplot(math_data, aes(y=age)) +
  geom_boxplot(outlier.colour="red", outlier.shape=8,
               outlier.size=1) + labs(title="Box plot of Age",y="Age") + theme(plot.title = element_te

qq_plot <- ggplot(math_data, aes(sample=age)) + geom_qq() + labs(title="QQ Plot of Age",x="Age") + theme

grid.arrange(histogram, box_plot, qq_plot, nrow = 1)
```

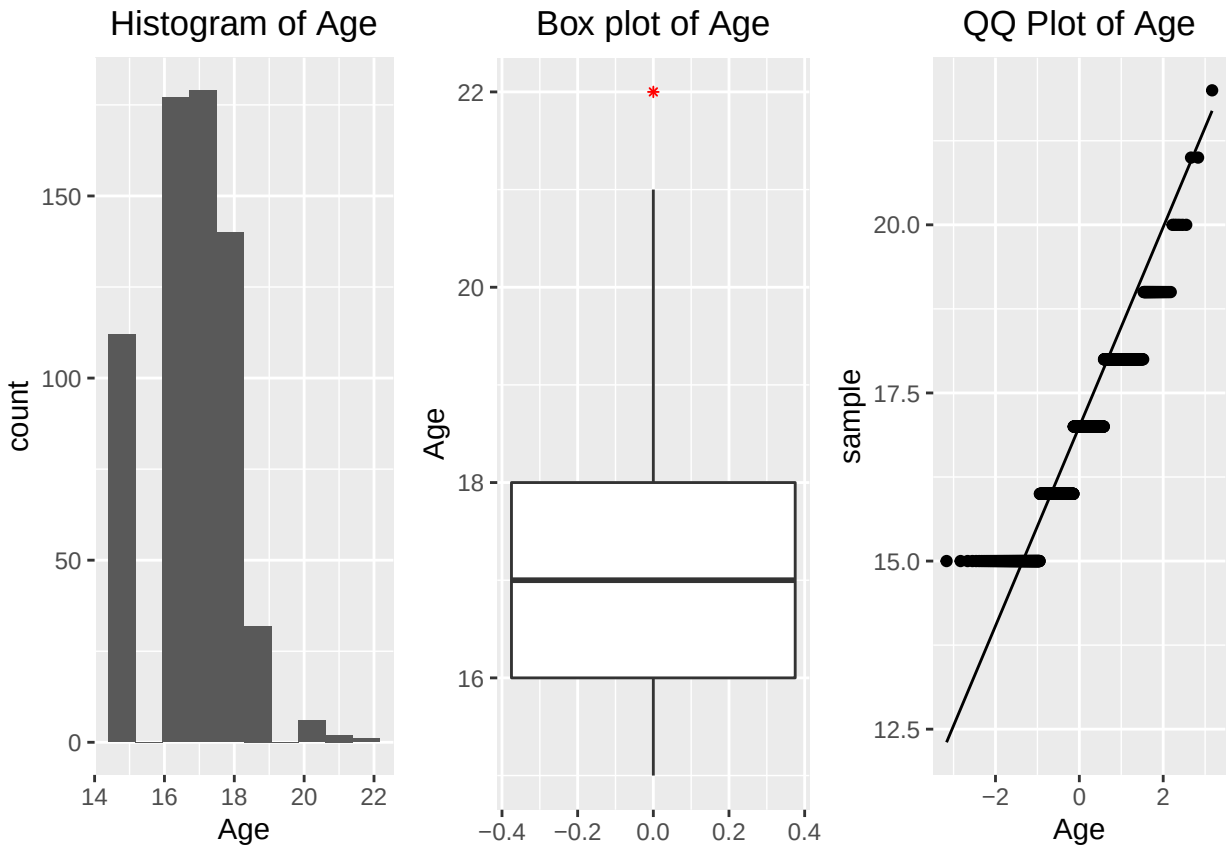


We need to drop the one observation of 22 year old student as it is an outlier.

```
math_data <- subset(math_data, age != 22)
```

Language

```
histogram <- ggplot(lang_data, aes(x=age))  + geom_histogram(bins = 10) + labs(title="Histogram of Age"

box_plot <- ggplot(lang_data, aes(y=age)) +
  geom_boxplot(outlier.colour="red", outlier.shape=8,
                outlier.size=1) + labs(title="Box plot of Age",y="Age") + theme(plot.title = element_t

qq_plot <- ggplot(lang_data, aes(sample=age)) + geom_qq() + labs(title="QQ Plot of Age",x="Age") + theme

grid.arrange(histogram, box_plot, qq_plot, nrow = 1)
```



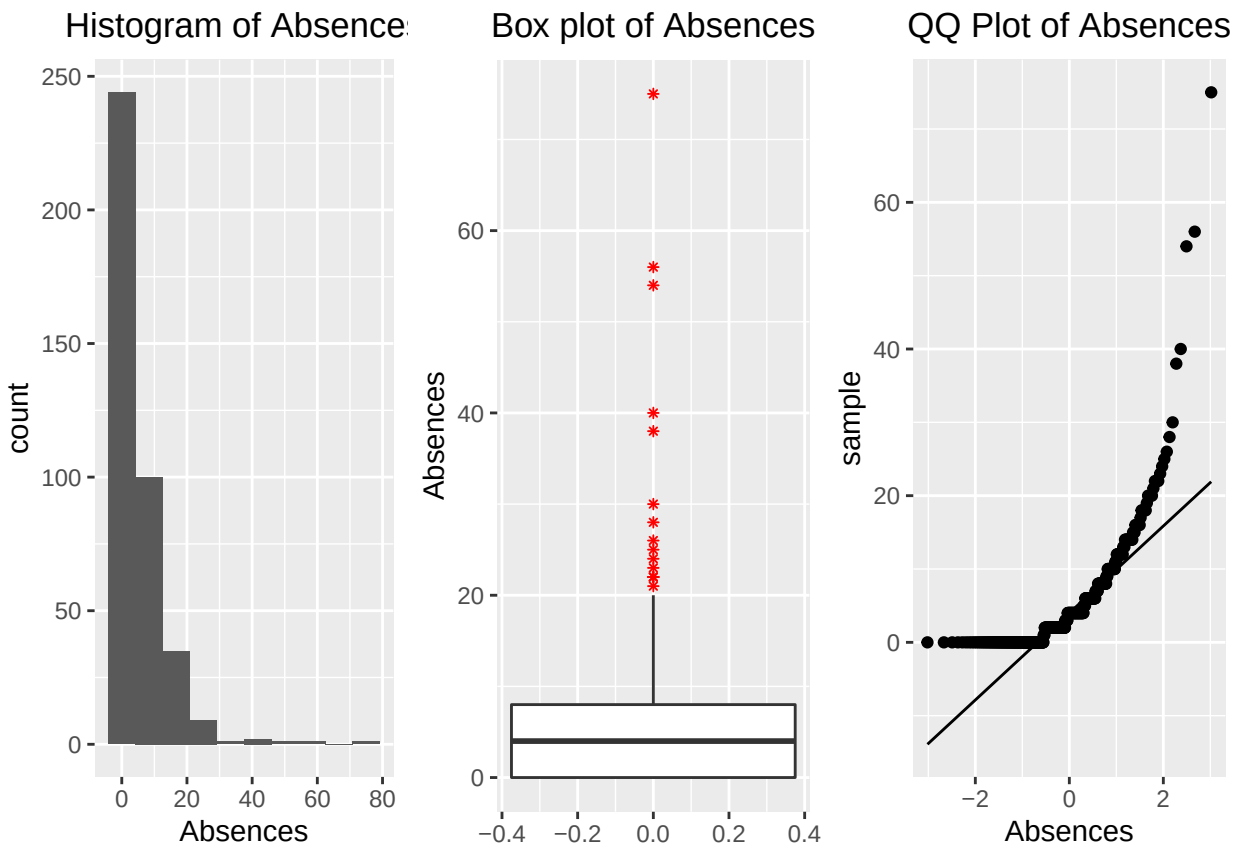
Again we see there is an outlier of 22 year old student which we can remove.

```
lang_data <- subset(lang_data, age != 22)
```

Absences

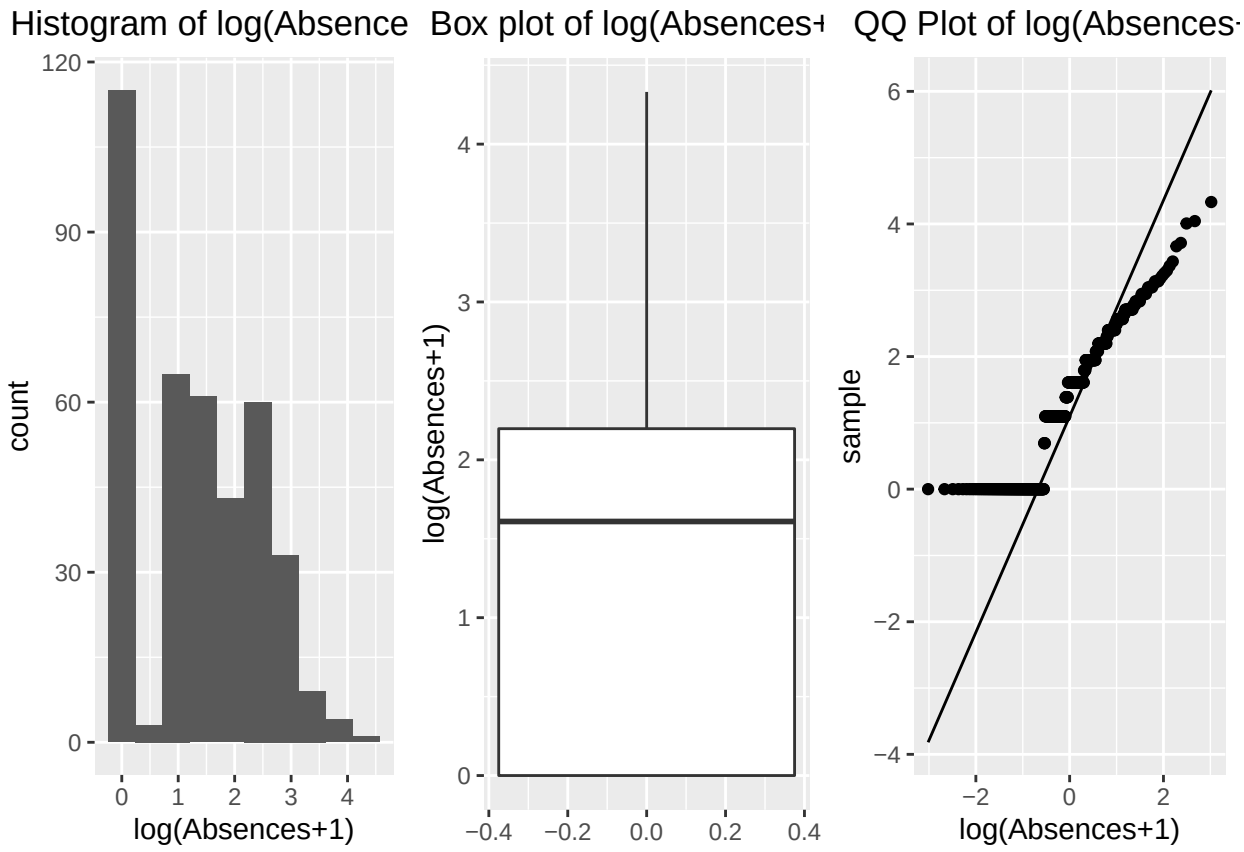
Math

```
histogram <- ggplot(math_data, aes(x=absences))  + geom_histogram(bins = 10) + labs(title="Histogram of

box_plot <- ggplot(math_data, aes(y=absences)) +
  geom_boxplot(outlier.colour="red", outlier.shape=8,
               outlier.size=1) + labs(title="Box plot of Absences",y="Absences") + theme(plot.title = 

qq_plot <- ggplot(math_data, aes(sample=absences)) + geom_qq() + labs(title="QQ Plot of Absences",x="Abs

grid.arrange(histogram, box_plot, qq_plot, nrow = 1)
```



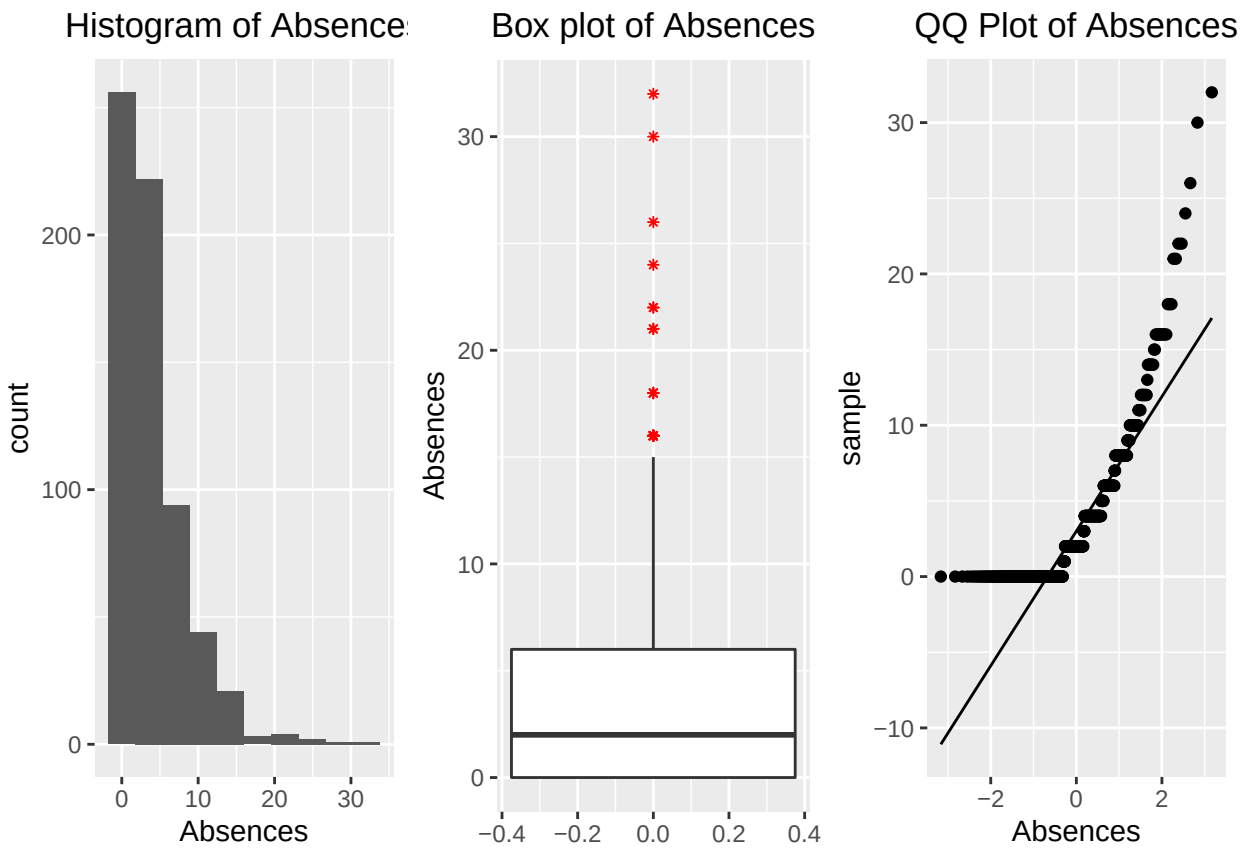
Absences data is skewed, we can log transform and check again log_Absences = log(Absences + 1)

```
math_data$absences <- math_data$absences + 1
math_data$log_absences <- log(math_data$absences)
math_data <- subset(math_data, select = -c(absences))

histogram <- ggplot(math_data, aes(x=log_absences))  + geom_histogram(bins = 10) + labs(title="Histogra

box_plot <- ggplot(math_data, aes(y=log_absences)) +
  geom_boxplot(outlier.colour="red", outlier.shape=8,
               outlier.size=1) + labs(title="Box plot of log(Absences+1)",y="log(Absences+1)") + theme

qq_plot <- ggplot(math_data, aes(sample=log_absences)) + geom_qq() + labs(title="QQ Plot of log(Absence

grid.arrange(histogram, box_plot, qq_plot, nrow = 1)
```



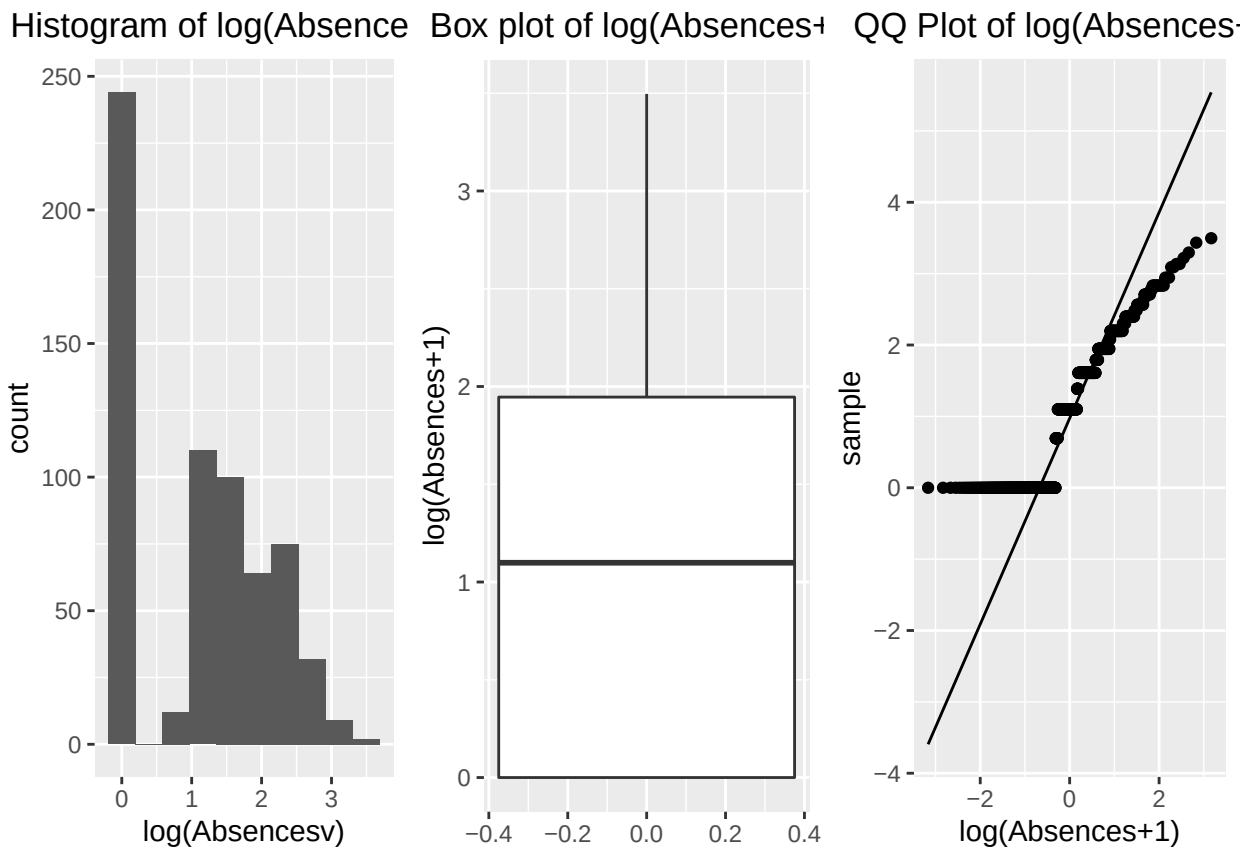
Language

```
histogram <- ggplot(lang_data, aes(x=absences))  + geom_histogram(bins = 10) + labs(title="Histogram of

box_plot <- ggplot(lang_data, aes(y=absences)) +
  geom_boxplot(outlier.colour="red", outlier.shape=8,
               outlier.size=1) + labs(title="Box plot of Absences",y="Absences") + theme(plot.title = e

qq_plot <- ggplot(lang_data, aes(sample=absences)) + geom_qq() + labs(title="QQ Plot of Absences",x="Abs

grid.arrange(histogram, box_plot, qq_plot, nrow = 1)
```



Absences data is skewed, we can log transform and check again log_Absences = log(Absences + 1)

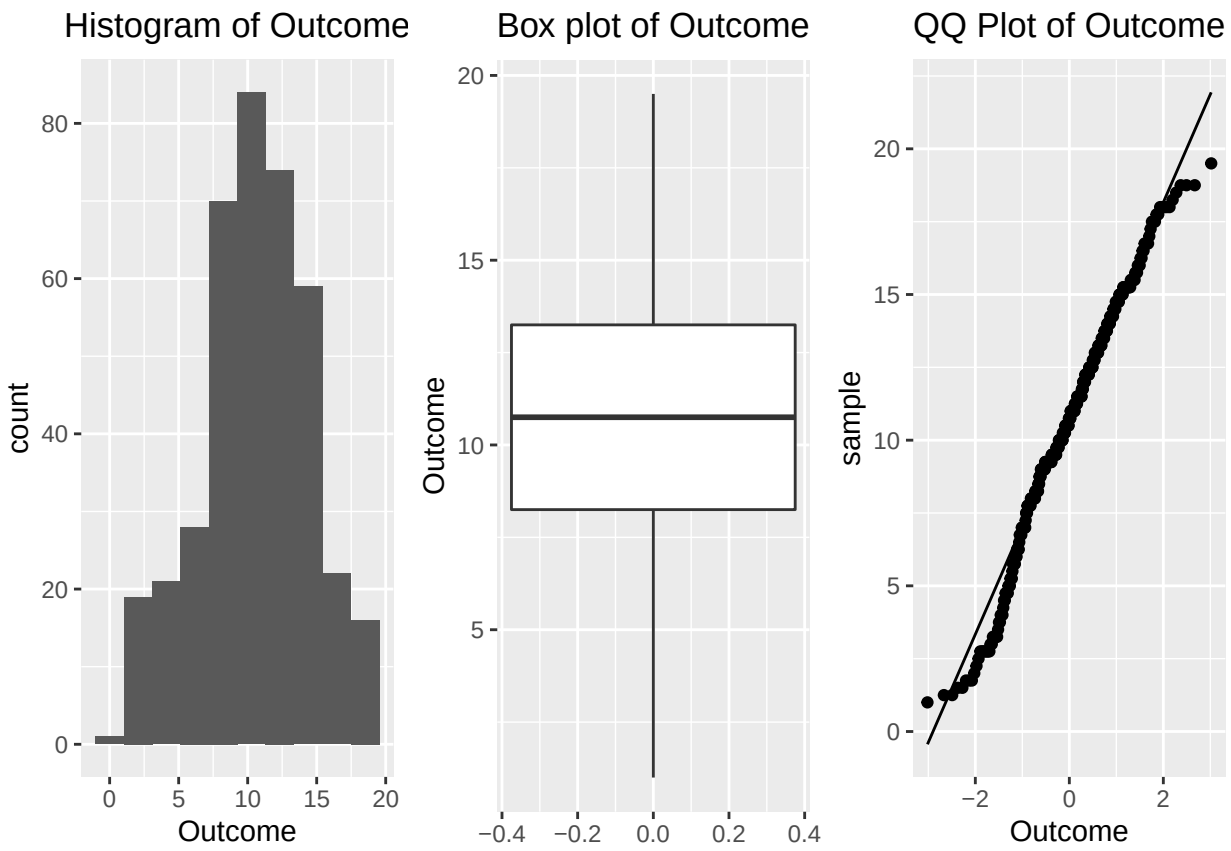
```
lang_data$absences <- lang_data$absences + 1
lang_data$log_absences <- log(lang_data$absences)
lang_data <- subset(lang_data, select = -c(absences))

histogram <- ggplot(lang_data, aes(x=log_absences))  + geom_histogram(bins = 10) + labs(title="Histogra

box_plot <- ggplot(lang_data, aes(y=log_absences)) +
  geom_boxplot(outlier.colour="red", outlier.shape=8,
               outlier.size=1) + labs(title="Box plot of log(Absences+1)",y="log(Absences+1)") + theme

qq_plot <- ggplot(lang_data, aes(sample=log_absences)) + geom_qq() + labs(title="QQ Plot of log(Absence

grid.arrange(histogram, box_plot, qq_plot, nrow = 1)
```



Outcome $= 2(G1 + G2 + 2\mathrm{G3})/4$

Math

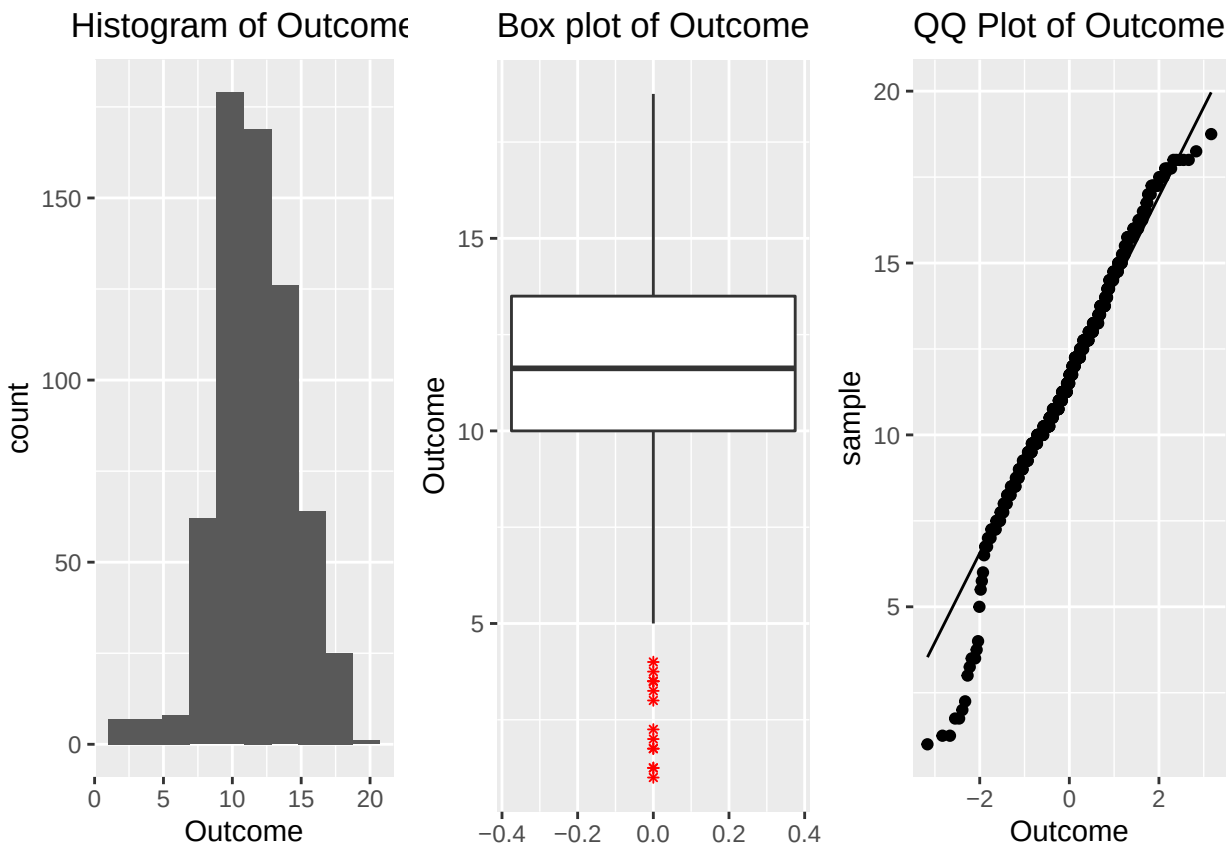
```
histogram <- ggplot(math_data, aes(x=outcome))  + geom_histogram(bins = 10) + labs(title="Histogram of (

box_plot <- ggplot(math_data, aes(y=outcome)) +
  geom_boxplot(outlier.colour="red", outlier.shape=8,
               outlier.size=1) + labs(title="Box plot of Outcome",y="Outcome") + theme(plot.title = el

qq_plot <- ggplot(math_data, aes(sample=outcome)) + geom_qq() + labs(title="QQ Plot of Outcome",x="Outc

grid.arrange(histogram, box_plot, qq_plot, nrow = 1)
```



Looks normally distributed

Language

```
histogram <- ggplot(lang_data, aes(x=outcome))  + geom_histogram(bins = 10) + labs(title="Histogram of 

box_plot <- ggplot(lang_data, aes(y=outcome)) +
  geom_boxplot(outlier.colour="red", outlier.shape=8,
               outlier.size=1) + labs(title="Box plot of Outcome",y="Outcome") + theme(plot.title = el

qq_plot <- ggplot(lang_data, aes(sample=outcome)) + geom_qq() + labs(title="QQ Plot of Outcome",x="Outc

grid.arrange(histogram, box_plot, qq_plot, nrow = 1)
```



There are some outliers on low end,but we can proceed forward for now due to smooth QQ Plot.

Checking frequency distribution of categorical variables and collapsing levels

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
print("Medu : 0+1,2,3,4")
```

```
## [1] "Medu : 0+1,2,3,4"
```

```
math_data$Medu <- recode(math_data$Medu, "c(0, 1)='0+1'")
lang_data$Medu <- recode(lang_data$Medu, "c(0, 1)='0+1'")

print("Fedu : 0+1,2,3,4")
```

```
## [1] "Fedu : 0+1,2,3,4"
```

```r
math_data$Fedu <- recode(math_data$Fedu, "c(0, 1)='0+1'")
lang_data$Fedu <- recode(lang_data$Fedu, "c(0, 1)='0+1'")

print("traveltime : 1,2,3+4")
```

```
## [1] "traveltime : 1,2,3+4"
```

```r
math_data$traveltime <- recode(math_data$traveltime, "c(3, 4)='3+4'")
lang_data$traveltime <- recode(lang_data$traveltime, "c(3, 4)='3+4'")

print("studytime : 1,2,3+4")
```

```
## [1] "studytime : 1,2,3+4"
```

```r
math_data$studytime <- recode(math_data$studytime, "c(3, 4)='3+4'")
lang_data$studytime <- recode(lang_data$studytime, "c(3, 4)='3+4'")

print("failures : 0,1+2+3+4")
```

```
## [1] "failures : 0,1+2+3+4"
```

```r
math_data$failures <- recode(math_data$failures, "c(1, 2, 3, 4)='1+2+3+4'")
lang_data$failures <- recode(lang_data$failures, "c(1, 2, 3, 4)='1+2+3+4'")

print("famrel : 1+2+3,4,5")
```

```
## [1] "famrel : 1+2+3,4,5"
```

```r
math_data$famrel <- recode(math_data$famrel, "c(1, 2, 3)='1+2+3'")
lang_data$famrel <- recode(lang_data$famrel, "c(1, 2, 3)='1+2+3'")

print("freetime : 1+2,3,4+5")
```

```
## [1] "freetime : 1+2,3,4+5"
```

```r
math_data$freetime <- recode(math_data$freetime, "c(1, 2)='1+2';c(4, 5)='4+5'")
lang_data$freetime <- recode(lang_data$freetime, "c(1, 2)='1+2';c(4, 5)='4+5'")

print("goout : 1+2,3,4+5")
```

```
## [1] "goout : 1+2,3,4+5"
```

```r
math_data$goout <- recode(math_data$goout, "c(1, 2)='1+2';c(4, 5)='4+5'")
lang_data$goout <- recode(lang_data$goout, "c(1, 2)='1+2';c(4, 5)='4+5'")

print("Dalc : 1,2,3+4+5")
```

```
## [1] "Dalc : 1,2,3+4+5"
```

```r
math_data$Dalc <- recode(math_data$Dalc, "c(3, 4, 5)='3+4+5'")
lang_data$Dalc <- recode(lang_data$Dalc, "c(3, 4, 5)='3+4+5'")

print("Walc : 1,2+3,4+5")
```

```
## [1] "Walc : 1,2+3,4+5"
```

```r
math_data$Walc <- recode(math_data$Walc, "c(2, 3)='2+3';c(4, 5)='4+5'")
lang_data$Walc <- recode(lang_data$Walc, "c(2, 3)='2+3';c(4, 5)='4+5'")
```

```r
print("health : 1+2+3,4+5")
```

```
## [1] "health : 1+2+3,4+5"
```

```r
math_data$health <- recode(math_data$health, "c(1, 2, 3)='1+2+3';c(4, 5)='4+5'")
lang_data$health <- recode(lang_data$health, "c(1, 2, 3)='1+2+3';c(4, 5)='4+5'")
```

Checking multi-collinearity for numemrical variables

Math

```r
cor(math_data[c("age","log_absences","outcome")],use="complete.obs")
```

```
##                      age log_absences    outcome
## age            1.0000000    0.1369376 -0.1383891
## log_absences   0.1369376    1.0000000  0.1135130
## outcome       -0.1383891    0.1135130  1.0000000
```

Language

```r
cor(lang_data[c("age","log_absences","outcome")],use="complete.obs")
```

```
##                      age log_absences    outcome
## age            1.0000000    0.1230496 -0.1165589
## log_absences   0.1230496    1.0000000 -0.1076935
## outcome       -0.1165589   -0.1076935  1.0000000
```

Final visualization of dataset after EDA

```r
print("Math")
```

```
## [1] "Math"
```

```r
summary(math_data)
```

```
##  school   sex          age        address famsize   Pstatus  Medu
##  GP:348   F:208   Min.   :15.00   R: 88   GT3:280   A: 41   0+1: 62
##  MS: 46   M:186   1st Qu.:16.00   U:306   LE3:114   T:353   2  :103
##                   Median :17.00                             3  : 98
##                   Mean   :16.68                             4  :131
##                   3rd Qu.:18.00
##                   Max.   :21.00
##   Fedu           Mjob           Fjob            reason      guardian
##  0+1: 83   at_home : 59   at_home : 20   course    :145   father: 90
##  2  :115   health  : 34   health  : 18   home      :109   mother:272
##  3  :100   other   :141   other   :217   other     : 35   other : 32
##  4  : 96   services:102   services:110   reputation:105
##            teacher : 58   teacher : 29
##
##  traveltime studytime    failures   schoolsup famsup     paid
##  1  :256    1  :104   0      :312   no :343   no :152   no :213
##  2  :107    2  :198   1+2+3+4: 82   yes: 51   yes:242   yes:181
##  3+4: 31    3+4: 92
##
##
```

```
##
##  activities nursery   higher    internet  romantic    famrel    freetime
##  no :193    no : 80   no : 19   no : 66   no :263   1+2+3: 94   1+2: 83
##  yes:201    yes:314   yes:375   yes:328   yes:131   4    :195   3  :157
##                                                     5    :105   4+5:154
##
##
##
##  goout        Dalc      Walc       health       outcome
##  1+2:126   1    :276   1  :151   1+2+3:182   Min.   : 1.00
##  3  :130   2    : 75   2+3:165   4+5  :212   1st Qu.: 8.25
##  4+5:138   3+4+5: 43   4+5: 78               Median :10.75
##                                              Mean   :10.62
##                                              3rd Qu.:13.25
##                                              Max.   :19.50
##   log_absences
##  Min.   :0.000
##  1st Qu.:0.000
##  Median :1.609
##  Mean   :1.367
##  3rd Qu.:2.197
##  Max.   :4.331
```

```
print("Language")
```

```
## [1] "Language"
```

```
summary(math_data)
```

```
##  school   sex          age        address famsize   Pstatus  Medu
##  GP:348   F:208   Min.   :15.00   R: 88   GT3:280   A: 41   0+1: 62
##  MS: 46   M:186   1st Qu.:16.00   U:306   LE3:114   T:353   2  :103
##                   Median :17.00                             3  : 98
##                   Mean   :16.68                             4  :131
##                   3rd Qu.:18.00
##                   Max.   :21.00
##   Fedu           Mjob           Fjob            reason      guardian
##  0+1: 83   at_home : 59   at_home : 20   course    :145   father: 90
##  2  :115   health  : 34   health  : 18   home      :109   mother:272
##  3  :100   other   :141   other   :217   other     : 35   other : 32
##  4  : 96   services:102   services:110   reputation:105
##            teacher : 58   teacher : 29
##
##  traveltime studytime    failures   schoolsup famsup     paid
##  1  :256    1  :104   0      :312   no :343   no :152   no :213
##  2  :107    2  :198   1+2+3+4: 82   yes: 51   yes:242   yes:181
##  3+4: 31    3+4: 92
##
##
##
##  activities nursery   higher    internet  romantic    famrel    freetime
##  no :193    no : 80   no : 19   no : 66   no :263   1+2+3: 94   1+2: 83
##  yes:201    yes:314   yes:375   yes:328   yes:131   4    :195   3  :157
##                                                     5    :105   4+5:154
##
```

```
##
##
##  goout        Dalc      Walc       health       outcome
##  1+2:126   1    :276   1  :151   1+2+3:182   Min.   : 1.00
##  3  :130   2    : 75   2+3:165   4+5  :212   1st Qu.: 8.25
##  4+5:138   3+4+5: 43   4+5: 78               Median :10.75
##                                              Mean   :10.62
##                                              3rd Qu.:13.25
##                                              Max.   :19.50
##   log_absences
##  Min.   :0.000
##  1st Qu.:0.000
##  Median :1.609
##  Mean   :1.367
##  3rd Qu.:2.197
##  Max.   :4.331
```

Checking for variation in scores across school

```
print("Maths t-test")
```

```
## [1] "Maths t-test"
```

```
t.test(outcome~school, data = math_data)
```

```
##
##  Welch Two Sample t-test
##
## data:  outcome by school
## t = 0.9318, df = 59.203, p-value = 0.3552
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.6233528  1.7099970
## sample estimates:
## mean in group GP mean in group MS
##         10.68463         10.14130
```

```
print("Language t-test")
```

```
## [1] "Language t-test"
```

```
t.test(outcome~school, data = lang_data)
```

```
##
##  Welch Two Sample t-test
##
## data:  outcome by school
## t = 7.1715, df = 350.57, p-value = 4.434e-12
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.314318 2.307626
## sample estimates:
## mean in group GP mean in group MS
##         12.33531         10.52434
```

For Maths, we get a high p-value of 0.39, hence Maths scores are almost similar across both schools.

For Portuguese we get low p-value of almost 0, hence Portuguese scores are different across both schools. So we should slice the Portuguese data schoolwise, and build individual models.

Slicing data set on school basis

```
# gp_math <- subset(math_data, school == "GP")
gp_lang <- subset(lang_data, school == "GP")
gp_lang <- subset(gp_lang, select = -c(school))
# ms_math <- subset(math_data, school == "MS")
ms_lang <- subset(lang_data, school == "MS")
ms_lang <- subset(ms_lang, select = -c(school))
```

Building stepwise models

Math

```
library(olsrr)
```

```
##
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:datasets':
##
##     rivers
```

```
# stepwise regression
print("Both : Forward and Backward selection")
```

```
## [1] "Both : Forward and Backward selection"
```

```
model_math <- lm(outcome ~ ., data = math_data)
both_model_math <- ols_step_both_p(model_math, prem=0.1)
```

```
## Stepwise Selection Method
## ---------------------------
##
## Candidate Terms:
##
## 1. school
## 2. sex
## 3. age
## 4. address
## 5. famsize
## 6. Pstatus
## 7. Medu
## 8. Fedu
## 9. Mjob
## 10. Fjob
## 11. reason
## 12. guardian
## 13. traveltime
## 14. studytime
## 15. failures
## 16. schoolsup
## 17. famsup
## 18. paid
```

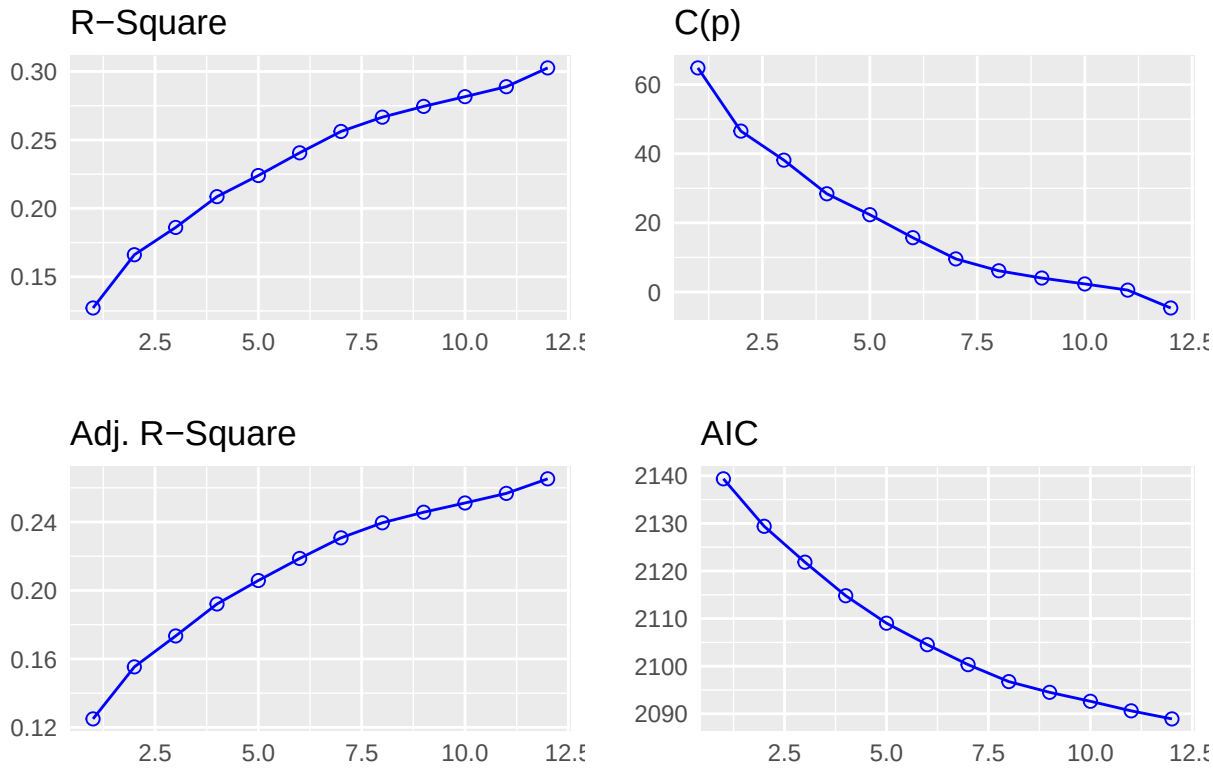
```
## 19. activities
## 20. nursery
## 21. higher
## 22. internet
## 23. romantic
## 24. famrel
## 25. freetime
## 26. goout
## 27. Dalc
## 28. Walc
## 29. health
## 30. log_absences
##
## We are selecting variables based on p value...
##
## Variables Entered/Removed:
##
## - failures added
## - Mjob added
## - log_absences added
## - goout added
## - sex added
## - freetime added
## - studytime added
## - famsup added
## - schoolsup added
## - romantic added
## - address added
## - Medu added
##
## No more variables to be added/removed.
##
##
## Final Model Output
## ------------------
##
##                          Model Summary
## ----------------------------------------------------------------
## R                        0.550       RMSE                 3.332
## R-Squared                0.303       Coef. Var           31.370
## Adj. R-Squared           0.265       MSE                 11.101
## Pred R-Squared           0.222       MAE                  2.577
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## -------------------------------------------------------------------
##                 Sum of
##                Squares         DF    Mean Square      F        Sig.
## -------------------------------------------------------------------
## Regression    1797.086         20         89.854    8.094    0.0000
## Residual      4140.815        373         11.101
```

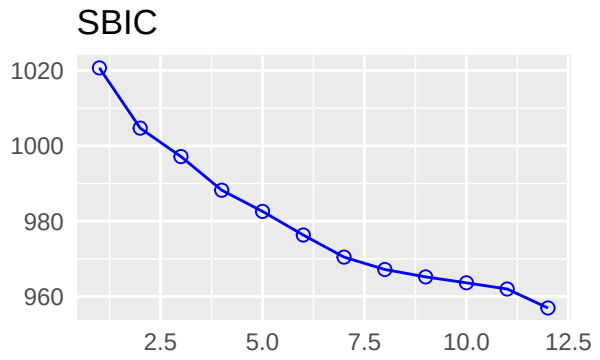
```
## Total          5937.901        393
## -----------------------------------------------------------------
##
##                                  Parameter Estimates
## ----------------------------------------------------------------------------------------------
##           model      Beta     Std. Error    Std. Beta      t        Sig      lower     upper
## ----------------------------------------------------------------------------------------------
##     (Intercept)    10.222          0.824                12.412    0.000     8.603    11.842
## failures1+2+3+4    -3.164          0.452       -0.331    -7.005    0.000    -4.052    -2.276
##      Mjobhealth     0.834          0.850        0.060     0.981    0.327    -0.838     2.505
##       Mjobother    -0.324          0.550       -0.040    -0.589    0.556    -1.405     0.758
##    Mjobservices     0.889          0.615        0.100     1.445    0.149    -0.320     2.098
##     Mjobteacher    -1.146          0.810       -0.105    -1.415    0.158    -2.739     0.447
##    log_absences     0.626          0.166        0.170     3.783    0.000     0.301     0.952
##          goout3    -0.583          0.438       -0.071    -1.333    0.183    -1.443     0.277
##        goout4+5    -1.665          0.440       -0.205    -3.781    0.000    -2.531    -0.799
##            sexM     0.993          0.379        0.128     2.619    0.009     0.247     1.739
##       freetime3    -0.835          0.479       -0.105    -1.743    0.082    -1.777     0.107
##     freetime4+5     0.190          0.485        0.024     0.392    0.695    -0.763     1.144
##      studytime2     0.562          0.428        0.072     1.312    0.190    -0.280     1.403
##    studytime3+4     1.618          0.523        0.176     3.096    0.002     0.591     2.646
##       famsupyes    -0.851          0.364       -0.107    -2.338    0.020    -1.566    -0.135
##    schoolsupyes    -1.252          0.518       -0.108    -2.417    0.016    -2.271    -0.234
##     romanticyes    -0.823          0.369       -0.100    -2.229    0.026    -1.548    -0.097
##        addressU     0.772          0.421        0.083     1.834    0.067    -0.056     1.600
##           Medu2    -0.167          0.568       -0.019    -0.294    0.769    -1.284     0.950
##           Medu3     0.208          0.599        0.023     0.348    0.728    -0.969     1.385
##           Medu4     1.320          0.688        0.160     1.918    0.056    -0.033     2.673
## ----------------------------------------------------------------------------------------------
```

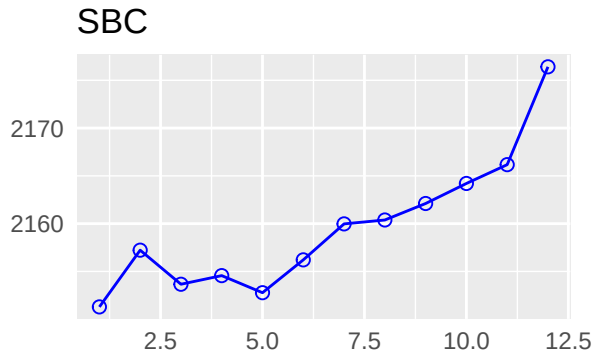
```
both_model_math
```

```
##
##                                Stepwise Selection Summary
## ---------------------------------------------------------------------------------------
##                          Added/                   Adj.
## Step      Variable      Removed     R-Square    R-Square     C(p)          AIC         RMSE
## ---------------------------------------------------------------------------------------
##    1      failures      addition       0.127       0.125    64.8140    2139.3595    3.6361
##    2        Mjob        addition       0.166       0.155    46.5360    2129.3897    3.5724
##    3    log_absences    addition       0.186       0.173    38.1390    2121.8477    3.5340
##    4       goout        addition       0.209       0.192    28.4080    2114.7967    3.4938
##    5         sex        addition       0.224       0.206    22.3620    2109.0335    3.4640
##    6      freetime      addition       0.241       0.219    15.7150    2104.5175    3.4358
##    7     studytime      addition       0.256       0.231     9.5750    2100.3282    3.4092
##    8       famsup       addition       0.267       0.240     6.1380    2096.7612    3.3896
##    9     schoolsup      addition       0.275       0.246     4.0360    2094.5092    3.3759
##   10      romantic      addition       0.282       0.251     2.3320    2092.6294    3.3637
##   11      address       addition       0.289       0.257     0.5410    2090.6198    3.3511
##   12        Medu        addition       0.303       0.265    -4.6220    2088.9285    3.3319
## ---------------------------------------------------------------------------------------
```

```
plot(both_model_math)
```


R−Square
0.30
0.25
0.20
0.15
2.5
5.0
7.5
10.0
C(p)
60
40
20
0
Adj. R−Square
0.24
0.20
0.16
0.12
AIC
2140
2130
2120
2110
2100
2090



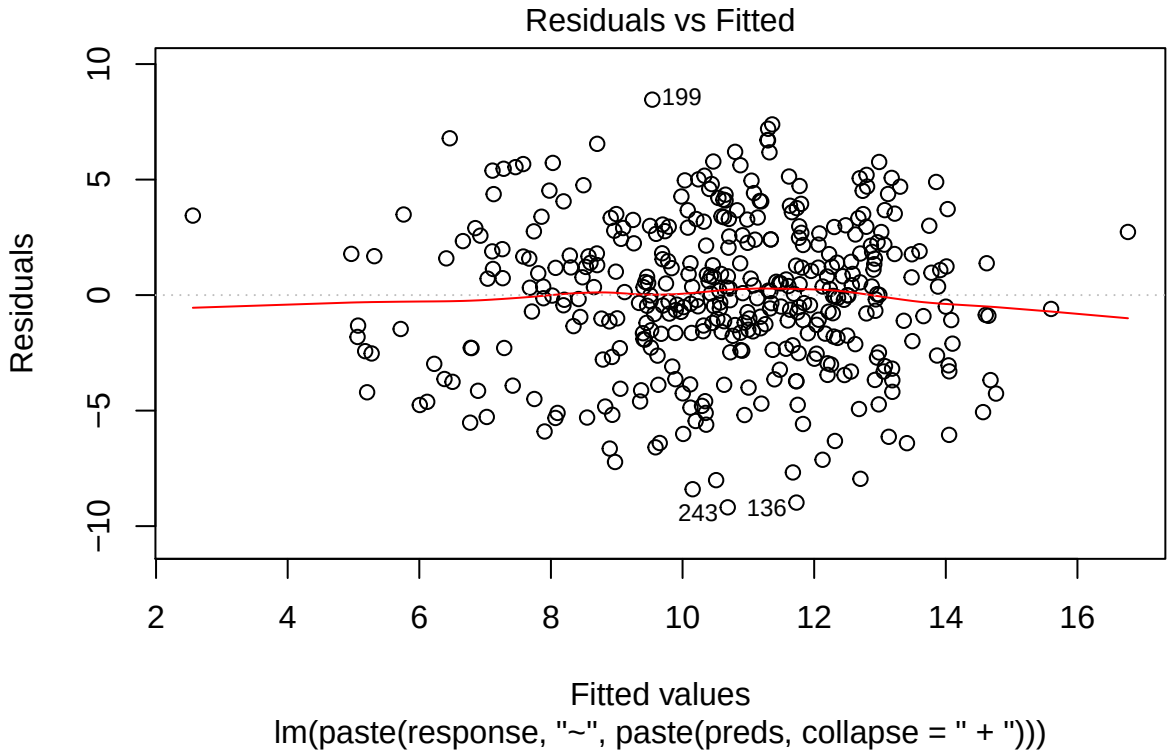


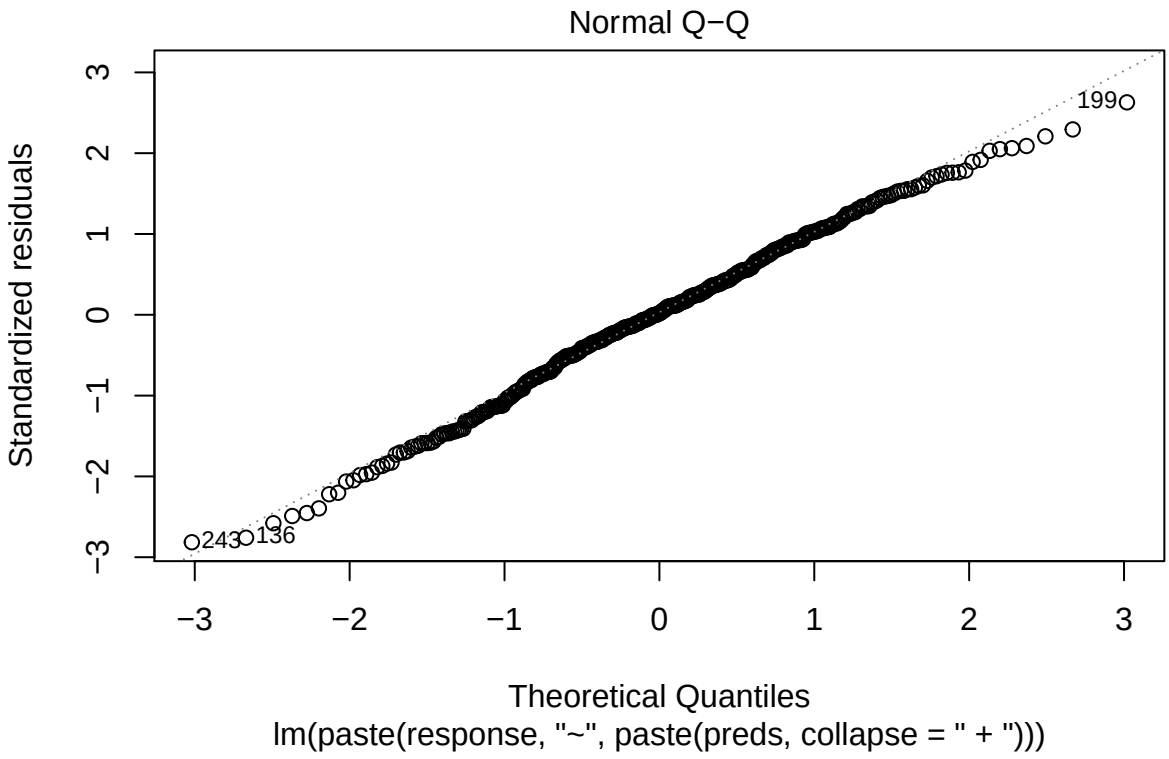

```
summary(both_model_math$model)
```

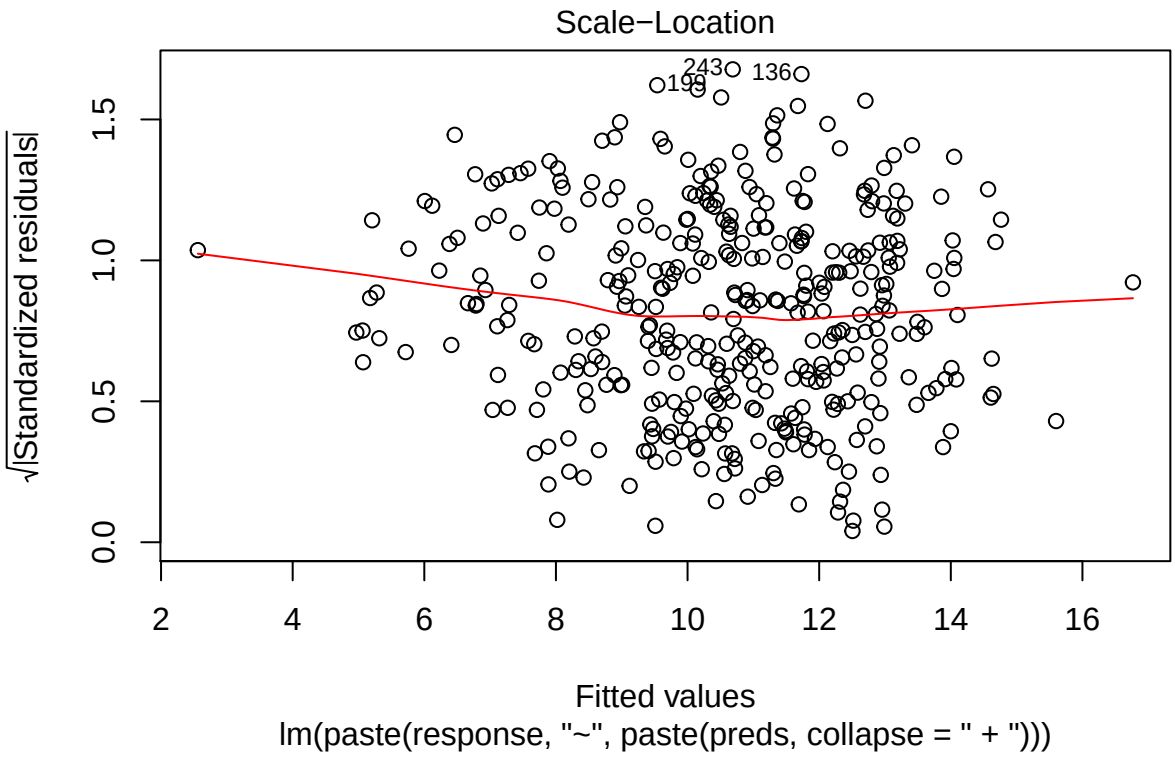
```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.1863 -2.0744  0.0639  2.2811  8.4597
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      10.2224     0.8236  12.412  < 2e-16 ***
## failures1+2+3+4  -3.1639     0.4517  -7.005 1.16e-11 ***
## Mjobhealth        0.8336     0.8501   0.981 0.327405
## Mjobother        -0.3238     0.5501  -0.589 0.556497
## Mjobservices      0.8891     0.6151   1.445 0.149159
## Mjobteacher      -1.1463     0.8102  -1.415 0.157910
## log_absences      0.6264     0.1656   3.783 0.000181 ***
## goout3           -0.5830     0.4375  -1.333 0.183468
## goout4+5         -1.6651     0.4404  -3.781 0.000182 ***
## sexM              0.9930     0.3792   2.619 0.009185 **
## freetime3        -0.8349     0.4791  -1.743 0.082228 .
## freetime4+5       0.1903     0.4849   0.392 0.694998
## studytime2        0.5615     0.4279   1.312 0.190244
```

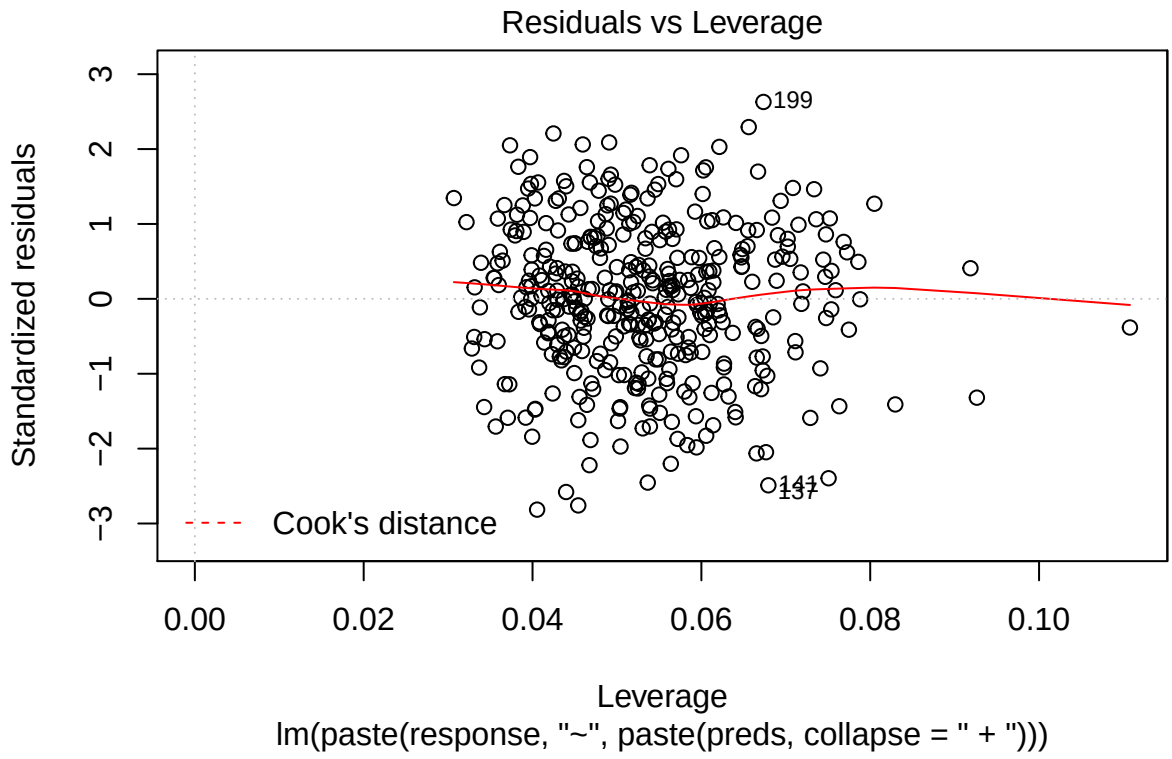
```
## studytime3+4      1.6183     0.5226   3.096 0.002107 **
## famsupyes        -0.8507     0.3639  -2.338 0.019931 *
## schoolsupyes     -1.2522     0.5180  -2.417 0.016120 *
## romanticyes      -0.8226     0.3690  -2.229 0.026410 *
## addressU          0.7724     0.4211   1.834 0.067419 .
## Medu2            -0.1670     0.5682  -0.294 0.769003
## Medu3             0.2081     0.5987   0.348 0.728404
## Medu4             1.3197     0.6882   1.918 0.055915 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.332 on 373 degrees of freedom
## Multiple R-squared:  0.3026, Adjusted R-squared:  0.2653
## F-statistic: 8.094 on 20 and 373 DF,  p-value: < 2.2e-16
```

```
plot(both_model_math$model)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(paste(response, "~", paste(preds, collapse = " + ")))
199
243
136
−3
−2
−1
0
1
2
3


Scale−Location
243
136
199
√|Standardized residuals|
0.0
0.5
1.0
1.5
2
4
6
8
10
12
14
16
Fitted values
lm(paste(response, "~", paste(preds, collapse = " + ")))




```
vif(both_model_math$model)
```

```
##                   GVIF Df GVIF^(1/(2*Df))
## failures      1.193372  1        1.092416
## Mjob          2.483592  4        1.120431
## log_absences  1.081973  1        1.040179
## goout         1.260723  2        1.059632
## sex           1.271720  1        1.127706
## freetime      1.308489  2        1.069529
## studytime     1.269537  2        1.061479
## famsup        1.113592  1        1.055269
## schoolsup     1.073266  1        1.035986
## romantic      1.072733  1        1.035728
## address       1.091627  1        1.044809
## Medu          2.420034  3        1.158698
```

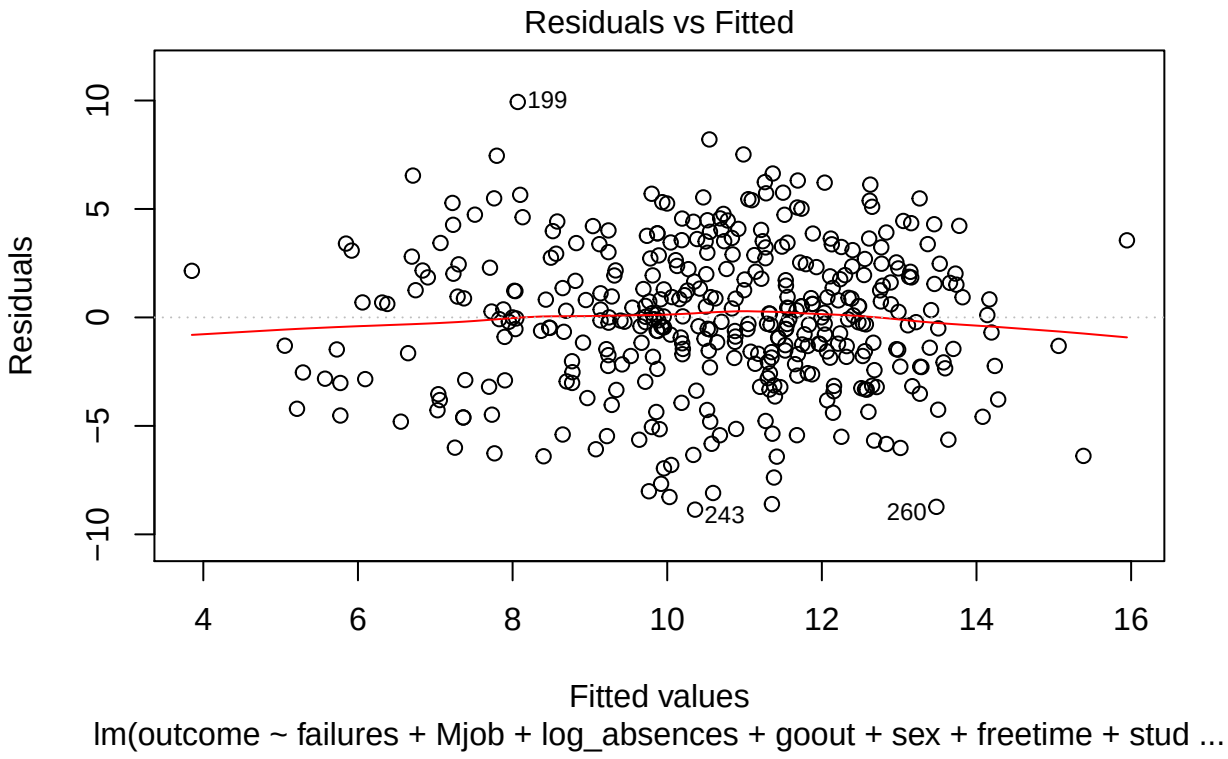
Making final Math model using result of stepwise selection

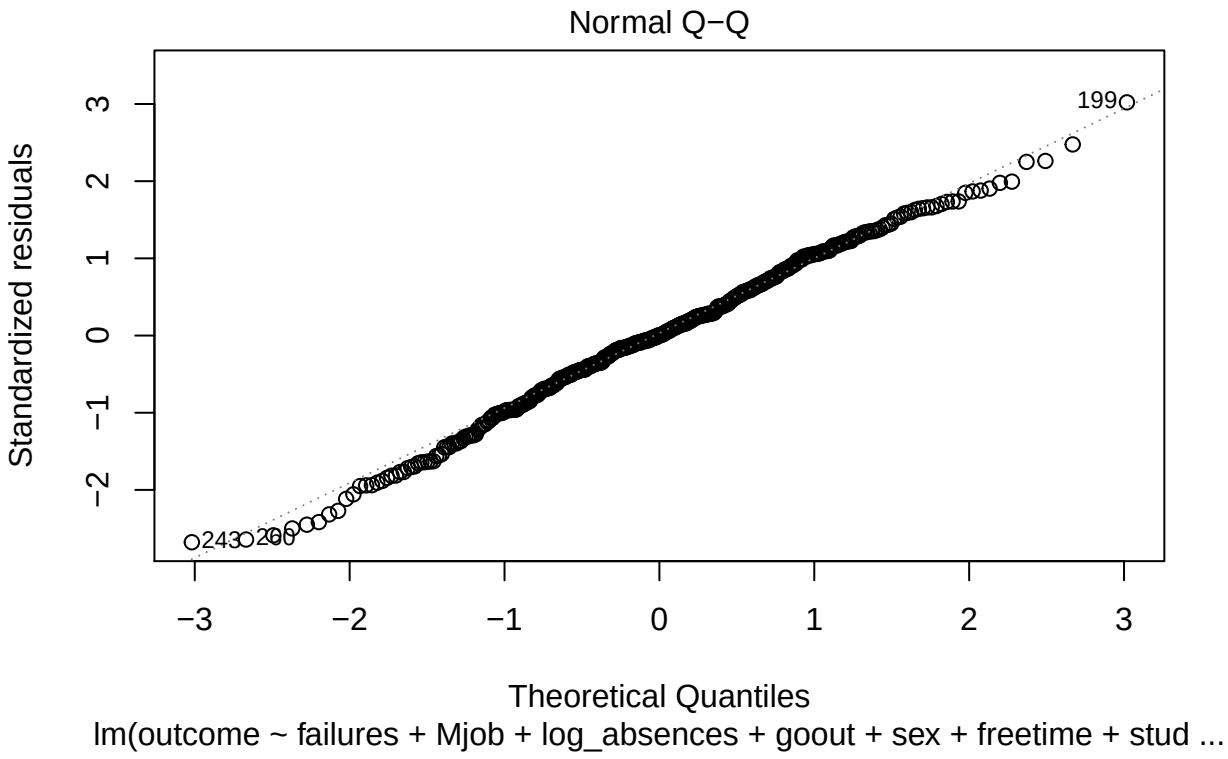
```
# Stepwise selection included 12 predictor variables
# This may lead to overfitting. Lets make a linear model with top 10 features used by stepwise regressi

final_model_math <- lm(outcome ~ failures + Mjob + log_absences + goout + sex + freetime + studytime + 
summary(final_model_math)
```

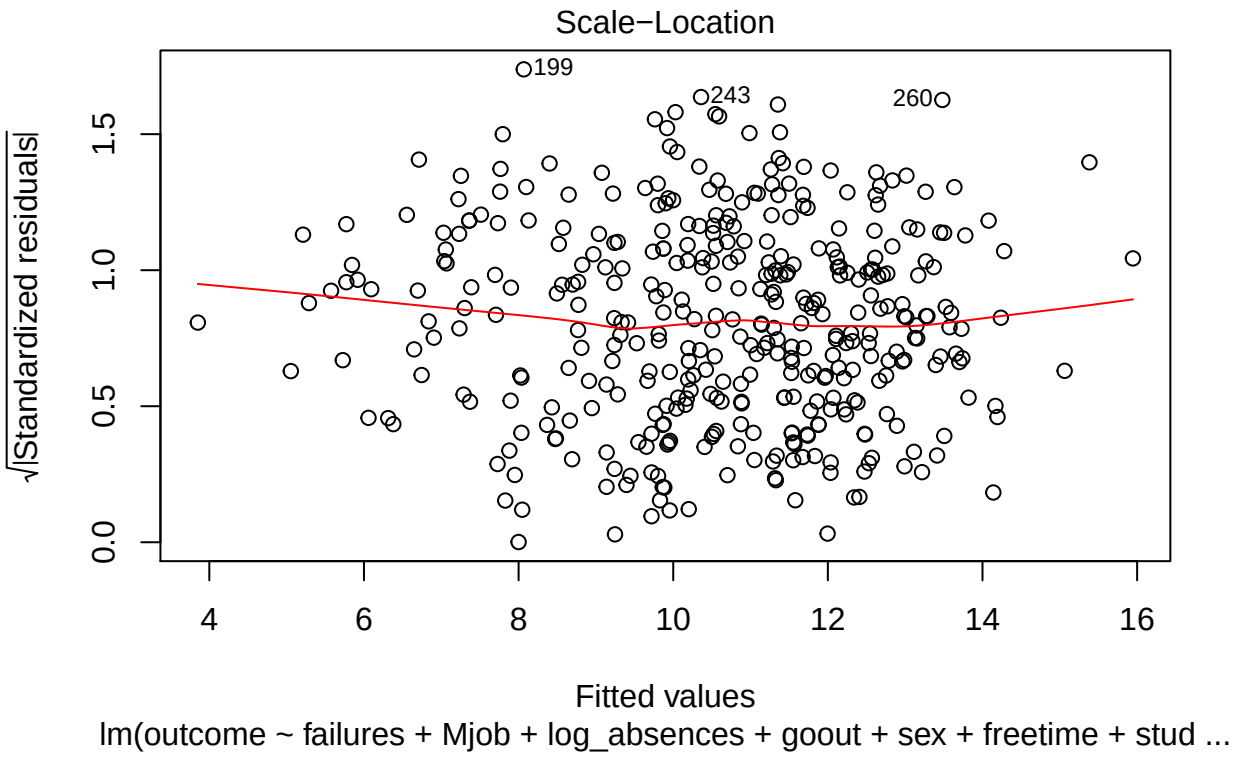
```
##
## Call:
## lm(formula = outcome ~ failures + Mjob + log_absences + goout +
```

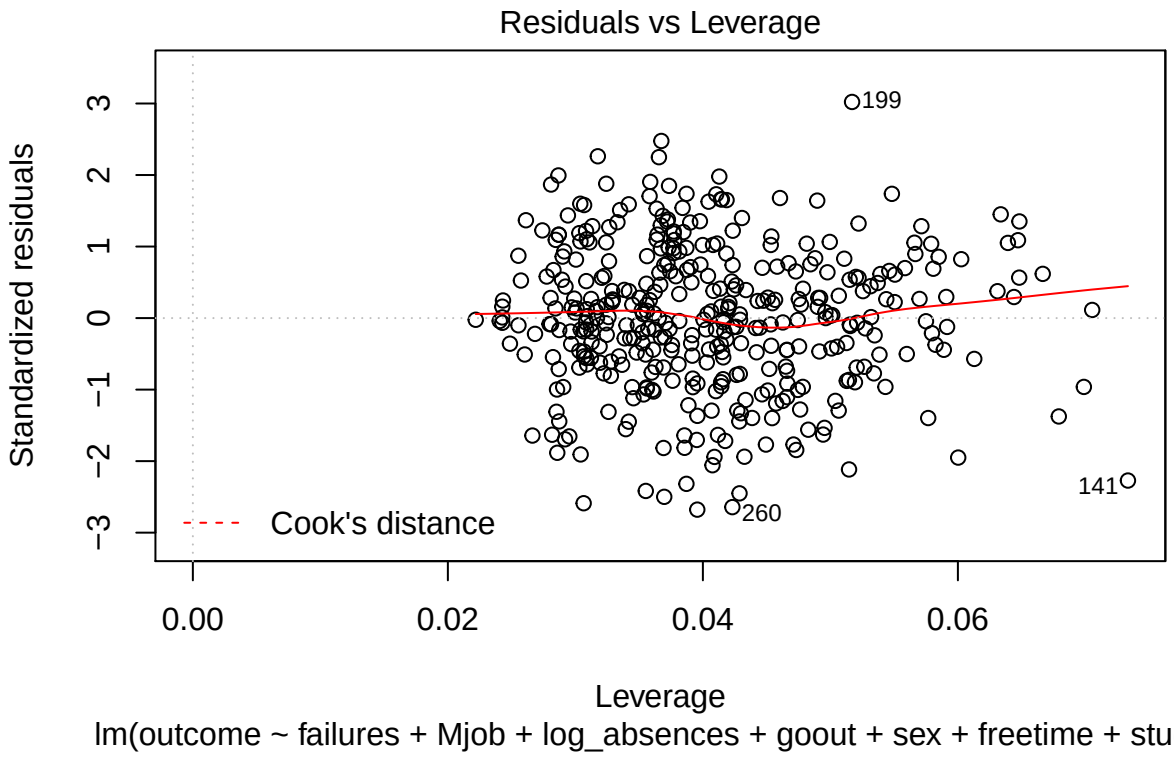
```
##     sex + freetime + studytime + famsup + schoolsup, data = math_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8601 -2.0685  0.0031  2.2535  9.9330
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     10.42960    0.75339  13.844  < 2e-16 ***
## failures1+2+3+4 -3.52385    0.43913  -8.025  1.3e-14 ***
## Mjobhealth       2.00659    0.75243   2.667 0.007987 **
## Mjobother       -0.02974    0.53167  -0.056 0.955424
## Mjobservices     1.45167    0.56233   2.582 0.010213 *
## Mjobteacher      0.25173    0.65702   0.383 0.701831
## log_absences     0.60471    0.16615   3.639 0.000311 ***
## goout3          -0.36450    0.43800  -0.832 0.405825
## goout4+5        -1.45465    0.44221  -3.289 0.001098 **
## sexM             1.03493    0.38240   2.706 0.007111 **
## freetime3       -0.88208    0.48302  -1.826 0.068615 .
## freetime4+5      0.09844    0.48960   0.201 0.840753
## studytime2       0.48490    0.42962   1.129 0.259754
## studytime3+4     1.50172    0.52443   2.863 0.004423 **
## famsupyes       -0.78224    0.36611  -2.137 0.033273 *
## schoolsupyes    -1.05483    0.52086  -2.025 0.043550 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.376 on 378 degrees of freedom
## Multiple R-squared:  0.2745, Adjusted R-squared:  0.2457
## F-statistic: 9.535 on 15 and 378 DF,  p-value: < 2.2e-16
```

```
plot(final_model_math)
```


Residuals vs Fitted
Residuals
Fitted values
lm(outcome ~ failures + Mjob + log_absences + goout + sex + freetime + stud ...
199
243
260
10
5
0
−5
−10
4
6
8
10
12
14
16


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(outcome ~ failures + Mjob + log_absences + goout + sex + freetime + stud ...
199
243
200
−3
−2
−1
0
1
2
3


Scale−Location
199
243
260
√|Standardized residuals|
0.0
0.5
1.0
1.5
4
6
8
10
12
14
16
Fitted values
lm(outcome ~ failures + Mjob + log_absences + goout + sex + freetime + stud ...




```
ncvTest(final_model_math)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.6987122, Df = 1, p = 0.40322
```

```
vif(final_model_math)
```

```
##                  GVIF Df GVIF^(1/(2*Df))
## failures     1.098742  1        1.048209
## Mjob         1.218027  4        1.024960
## log_absences 1.061000  1        1.030048
## goout        1.224451  2        1.051926
## sex          1.259963  1        1.122481
## freetime     1.288586  2        1.065438
## studytime    1.216980  2        1.050318
## famsup       1.098062  1        1.047884
## schoolsup    1.056907  1        1.028060
```

Language - GP

```
# stepwise regression
print("Both : Forward and Backward selection")
```

```
## [1] "Both : Forward and Backward selection"
```

```
model_gp_lang <- lm(outcome ~ ., data = gp_lang)
both_model_gp_lang <- ols_step_both_p(model_gp_lang)
```

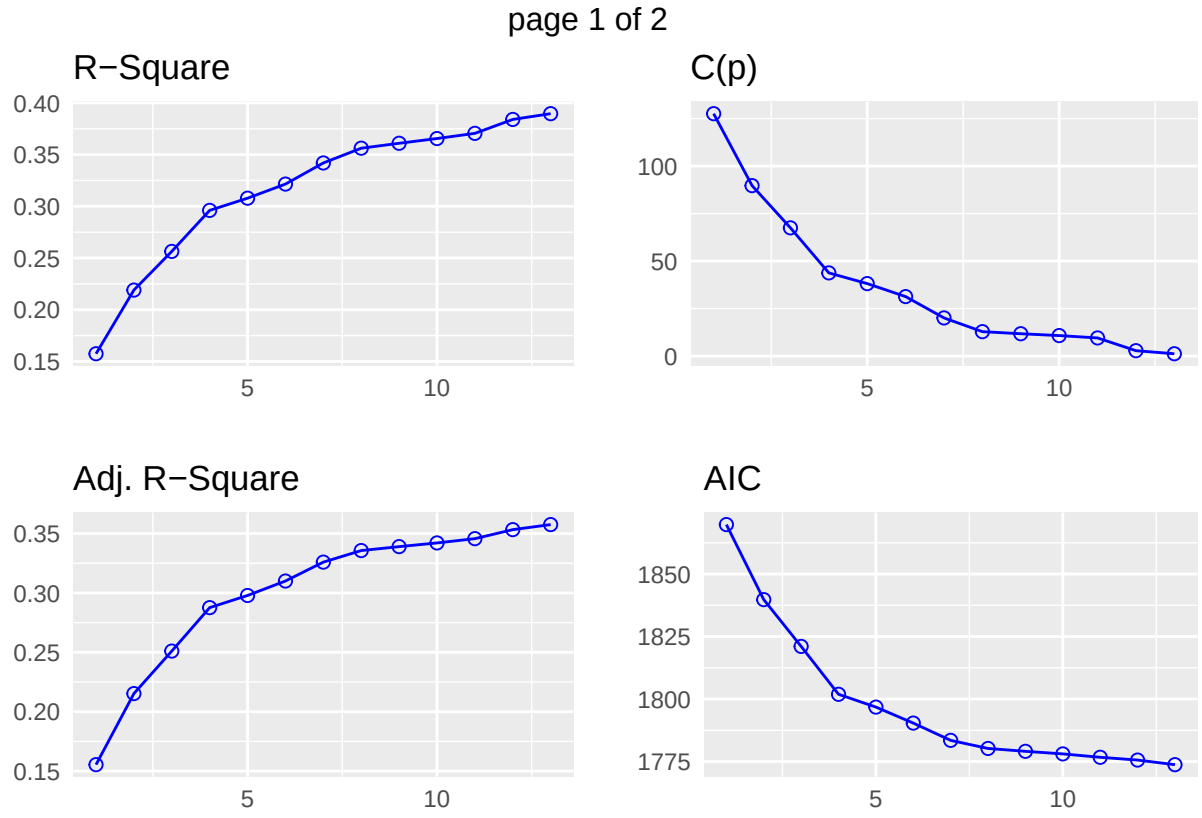
```
## Stepwise Selection Method
## ---------------------------
##
## Candidate Terms:
##
## 1. sex
## 2. age
## 3. address
## 4. famsize
## 5. Pstatus
## 6. Medu
## 7. Fedu
## 8. Mjob
## 9. Fjob
## 10. reason
## 11. guardian
## 12. traveltime
## 13. studytime
## 14. failures
## 15. schoolsup
## 16. famsup
## 17. paid
## 18. activities
## 19. nursery
## 20. higher
## 21. internet
## 22. romantic
## 23. famrel
## 24. freetime
## 25. goout
## 26. Dalc
## 27. Walc
## 28. health
## 29. log_absences
##
## We are selecting variables based on p value...
##
## Variables Entered/Removed:
##
## - failures added
## - higher added
## - schoolsup added
## - Walc added
## - sex added
## - log_absences added
## - Medu added
## - reason added
## - age added
## - health added
## - activities added
## - Fjob added
## - romantic added
##
## No more variables to be added/removed.
```

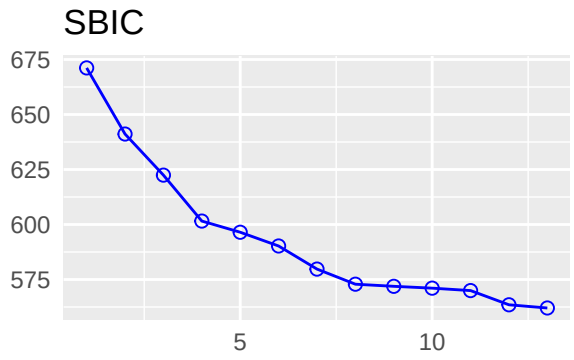
```
##
##
## Final Model Output
## ------------------
##
##                         Model Summary
## ------------------------------------------------------------------
## R                       0.624       RMSE                 1.925
## R-Squared               0.390       Coef. Var           15.606
## Adj. R-Squared          0.358       MSE                  3.706
## Pred R-Squared          0.324       MAE                  1.480
## ------------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## --------------------------------------------------------------------------
##                 Sum of
##                Squares         DF    Mean Square      F         Sig.
## --------------------------------------------------------------------------
## Regression     946.049         21          45.050    12.157    0.0000
## Residual      1482.255        400           3.706
## Total         2428.304        421
## --------------------------------------------------------------------------
##
##                                    Parameter Estimates
## ----------------------------------------------------------------------------------------------------
##            model      Beta    Std. Error    Std. Beta      t        Sig      lower     upper
## ----------------------------------------------------------------------------------------------------
##      (Intercept)     9.199         1.634                5.630    0.000     5.987    12.411
##  failures1+2+3+4    -1.790         0.338       -0.233   -5.295    0.000    -2.455    -1.125
##        higheryes     2.081         0.400        0.226    5.205    0.000     1.295     2.867
##     schoolsupyes    -1.615         0.292       -0.228   -5.530    0.000    -2.188    -1.041
##          Walc2+3     0.052         0.216        0.011    0.242    0.809    -0.373     0.478
##          Walc4+5    -0.715         0.278       -0.120   -2.568    0.011    -1.262    -0.167
##             sexM    -0.742         0.208       -0.154   -3.569    0.000    -1.151    -0.333
##     log_absences    -0.314         0.100       -0.130   -3.154    0.002    -0.509    -0.118
##            Medu2     0.125         0.306        0.024    0.408    0.684    -0.476     0.725
##            Medu3     0.401         0.320        0.071    1.254    0.211    -0.228     1.030
##            Medu4     0.724         0.318        0.141    2.274    0.023     0.098     1.350
##       reasonhome     0.392         0.238        0.073    1.649    0.100    -0.075     0.860
##      reasonother     0.291         0.417        0.029    0.698    0.486    -0.528     1.110
## reasonreputation     0.643         0.243        0.119    2.651    0.008     0.166     1.120
##              age     0.163         0.086        0.083    1.897    0.059    -0.006     0.332
##        health4+5    -0.401         0.196       -0.083   -2.050    0.041    -0.785    -0.016
##    activitiesyes     0.428         0.195        0.089    2.188    0.029     0.043     0.812
##       Fjobhealth    -0.319         0.683       -0.027   -0.467    0.641    -1.660     1.023
##        Fjobother    -0.835         0.493       -0.171   -1.694    0.091    -1.804     0.134
##     Fjobservices    -1.099         0.517       -0.200   -2.127    0.034    -2.115    -0.083
##      Fjobteacher    -0.136         0.606       -0.015   -0.224    0.823    -1.327     1.055
##      romanticyes    -0.389         0.204       -0.077   -1.911    0.057    -0.789     0.011
## ----------------------------------------------------------------------------------------------------
```

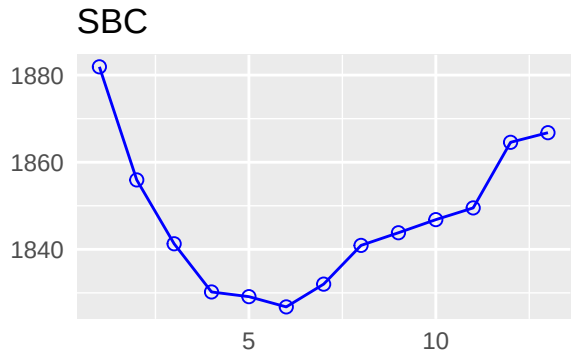
```
both_model_gp_lang
```

```
##
##                                 Stepwise Selection Summary
## ----------------------------------------------------------------------------------------------
##                          Added/                   Adj.
## Step      Variable       Removed     R-Square    R-Square      C(p)         AIC         RMSE
## ----------------------------------------------------------------------------------------------
##    1      failures       addition       0.157       0.155    127.5620    1869.7711    2.2071
##    2        higher       addition       0.219       0.215     89.7020    1839.7541    2.1275
##    3     schoolsup       addition       0.256       0.251     67.4850    1821.0457    2.0784
##    4          Walc       addition       0.296       0.288     43.7820    1801.8949    2.0270
##    5           sex       addition       0.308       0.298     38.1510    1796.7696    2.0124
##    6    log_absences       addition       0.322       0.310     31.3240    1790.3747    1.9949
##    7          Medu       addition       0.342       0.326     20.0900    1783.4672    1.9718
##    8        reason       addition       0.356       0.336     12.8620    1780.2276    1.9575
##    9           age       addition       0.361       0.339     11.7690    1779.0848    1.9526
##   10        health       addition       0.366       0.342     10.8060    1778.0514    1.9480
##   11    activities       addition       0.371       0.346      9.5560    1776.6996    1.9427
##   12          Fjob       addition       0.384       0.353      2.8480    1775.5855    1.9314
##   13      romantic       addition       0.390       0.358      1.2380    1773.7488    1.9250
## ----------------------------------------------------------------------------------------------
```

```
plot(both_model_gp_lang)
```








```
summary(both_model_gp_lang$model)
```

```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.8431 -1.3474 -0.1188  1.1334  5.6333
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       9.19898    1.63387   5.630 3.39e-08 ***
## failures1+2+3+4  -1.79020    0.33812  -5.295 1.97e-07 ***
## higheryes         2.08081    0.39979   5.205 3.11e-07 ***
## schoolsupyes     -1.61456    0.29194  -5.530 5.78e-08 ***
## Walc2+3           0.05248    0.21647   0.242 0.808556
## Walc4+5          -0.71480    0.27840  -2.568 0.010606 *
## sexM             -0.74239    0.20802  -3.569 0.000402 ***
## log_absences     -0.31382    0.09951  -3.154 0.001735 **
## Medu2             0.12458    0.30566   0.408 0.683804
## Medu3             0.40121    0.32004   1.254 0.210703
## Medu4             0.72393    0.31835   2.274 0.023494 *
## reasonhome        0.39221    0.23788   1.649 0.099985 .
## reasonother       0.29077    0.41657   0.698 0.485567
```

```
## reasonreputation  0.64312    0.24262   2.651 0.008351 **
## age               0.16286    0.08584   1.897 0.058514 .
## health4+5        -0.40082    0.19556  -2.050 0.041056 *
## activitiesyes     0.42767    0.19544   2.188 0.029233 *
## Fjobhealth       -0.31874    0.68250  -0.467 0.640742
## Fjobother        -0.83496    0.49295  -1.694 0.091080 .
## Fjobservices     -1.09921    0.51670  -2.127 0.034001 *
## Fjobteacher      -0.13570    0.60588  -0.224 0.822888
## romanticyes      -0.38901    0.20353  -1.911 0.056673 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.925 on 400 degrees of freedom
## Multiple R-squared:  0.3896, Adjusted R-squared:  0.3575
## F-statistic: 12.16 on 21 and 400 DF,  p-value: < 2.2e-16
```

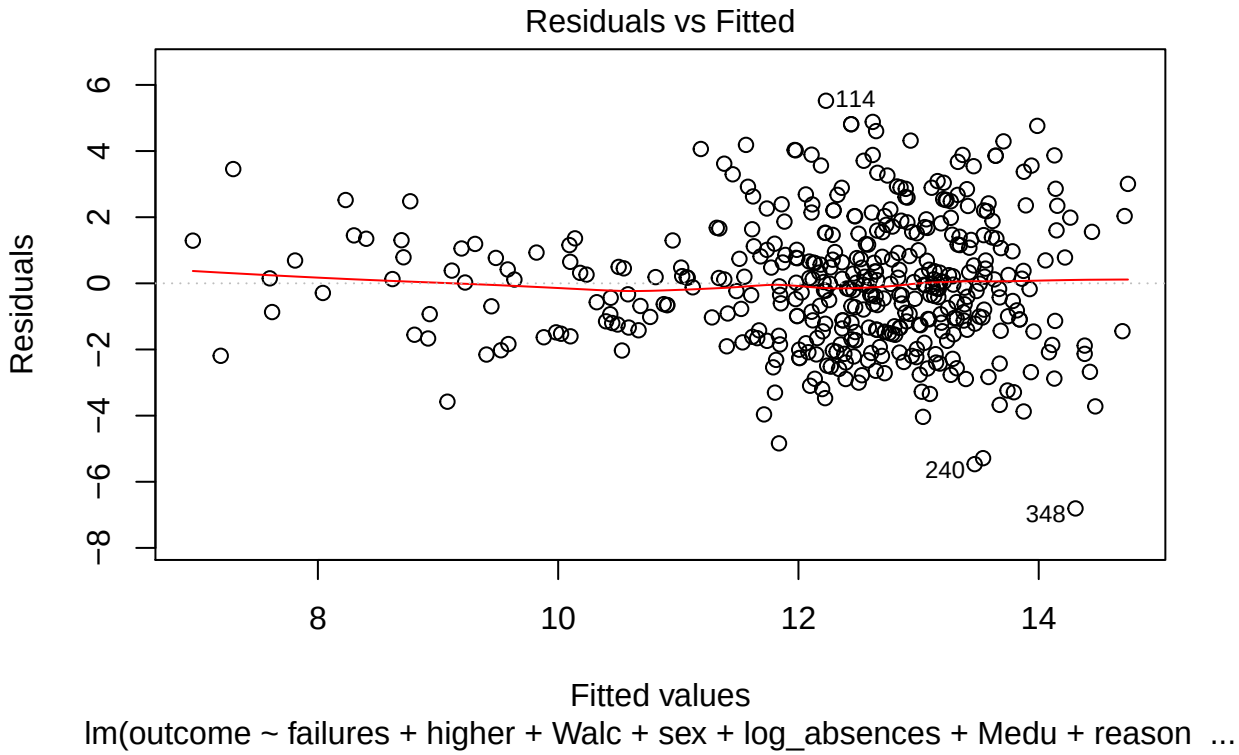
Final GP Language model

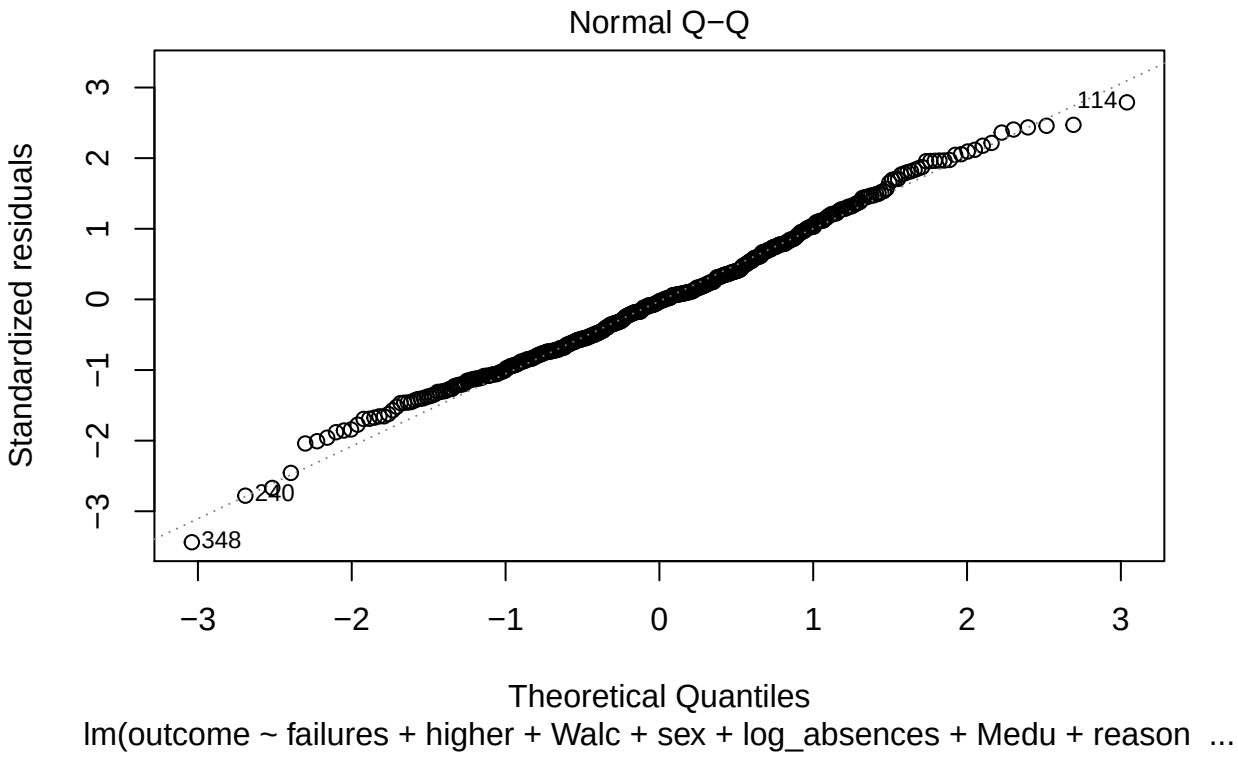
```
# Stepwise selection included 12 predictor variables
# This may lead to overfitting. Lets make a linear model with top 10 features used by stepwise regressi

final_model_lang_gp <- lm(outcome ~ failures + higher + Walc + sex + log_absences + Medu + reason + age
summary(final_model_lang_gp)
```

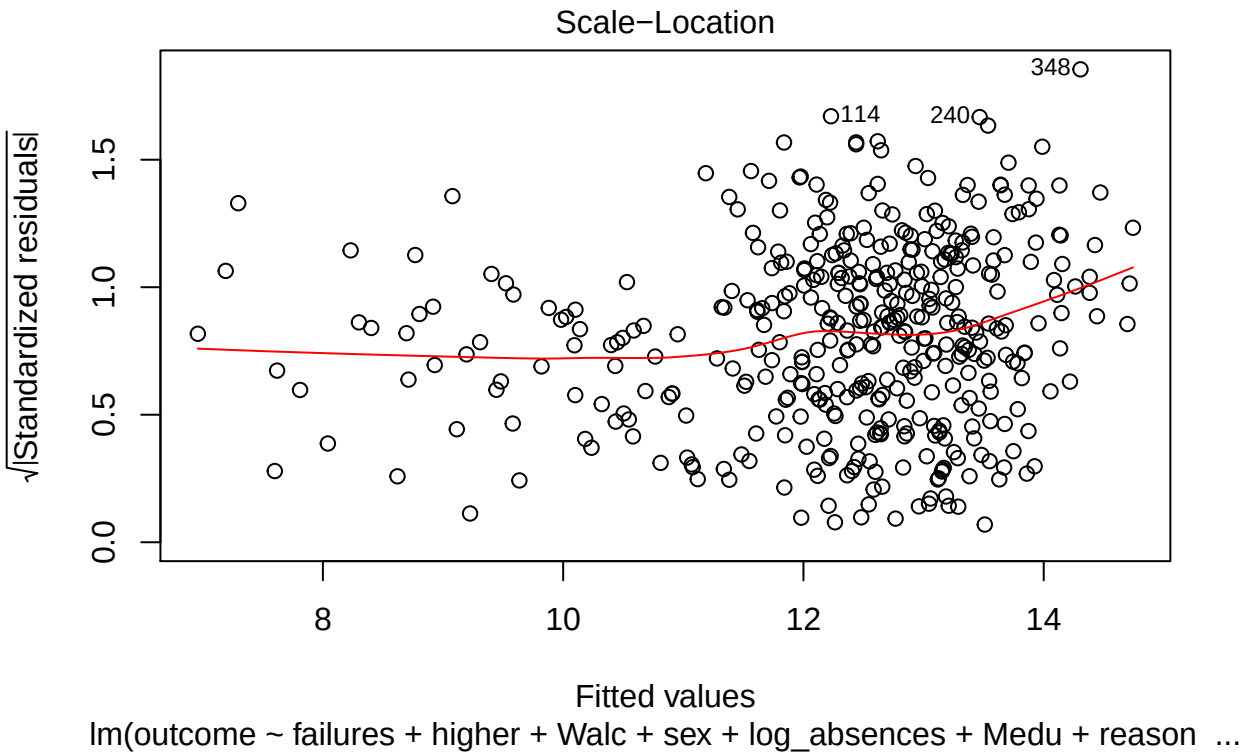
```
##
## Call:
## lm(formula = outcome ~ failures + higher + Walc + sex + log_absences +
##     Medu + reason + age + goout * romantic, data = gp_lang)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.8062 -1.4160 -0.0516  1.2954  5.5208
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            6.33058    1.59037   3.981 8.15e-05 ***
## failures1+2+3+4       -2.01133    0.35222  -5.710 2.19e-08 ***
## higheryes              2.10814    0.41728   5.052 6.64e-07 ***
## Walc2+3                0.14295    0.22742   0.629  0.52998
## Walc4+5               -0.68130    0.30955  -2.201  0.02831 *
## sexM                  -0.57423    0.21451  -2.677  0.00773 **
## log_absences          -0.26887    0.10404  -2.584  0.01011 *
## Medu2                  0.06746    0.31814   0.212  0.83218
## Medu3                  0.32233    0.33260   0.969  0.33306
## Medu4                  0.93410    0.32185   2.902  0.00391 **
## reasonhome             0.28138    0.24816   1.134  0.25751
## reasonother            0.32371    0.43368   0.746  0.45585
## reasonreputation       0.60838    0.25049   2.429  0.01559 *
## age                    0.27169    0.08880   3.059  0.00237 **
## goout3                -0.15629    0.31368  -0.498  0.61859
## goout4+5              -0.29369    0.31767  -0.925  0.35577
## romanticyes           -0.17350    0.37386  -0.464  0.64284
## goout3:romanticyes    -0.03033    0.53545  -0.057  0.95486
## goout4+5:romanticyes -0.29574    0.50107  -0.590  0.55538
```

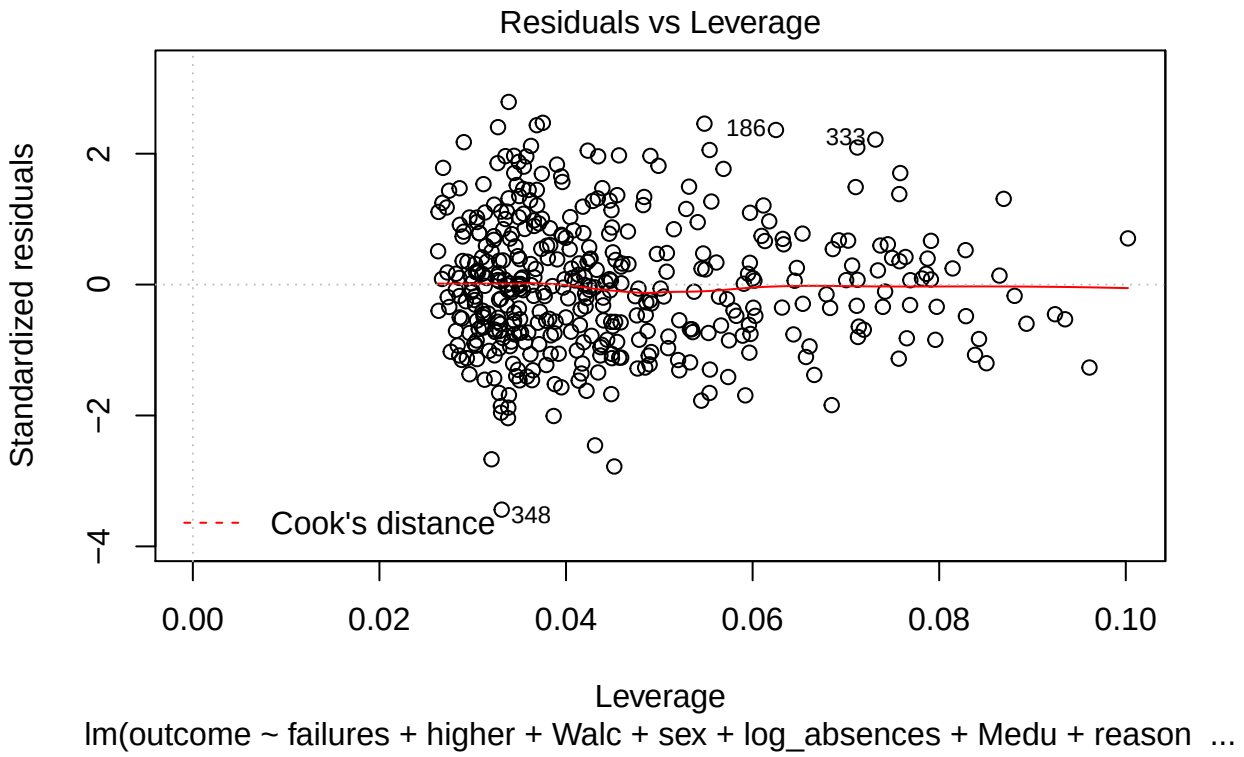
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.013 on 403 degrees of freedom
## Multiple R-squared:  0.3275, Adjusted R-squared:  0.2974
## F-statistic:  10.9 on 18 and 403 DF,  p-value: < 2.2e-16
```

```
plot(final_model_lang_gp)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(outcome ~ failures + higher + Walc + sex + log_absences + Medu + reason ...
114
240
348
−3
−2
−1
0
1
2
3


Scale−Location
348
114
240
√|Standardized residuals|
0.0
0.5
1.0
1.5
8
10
12
14
Fitted values
lm(outcome ~ failures + higher + Walc + sex + log_absences + Medu + reason ...




```
ncvTest(final_model_lang_gp)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 10.3578, Df = 1, p = 0.0012893
```

```
vif(final_model_lang_gp)
```

```
##                     GVIF Df GVIF^(1/(2*Df))
## failures        1.254746  1        1.120154
## higher          1.234177  1        1.110935
## Walc            1.439789  2        1.095405
## sex             1.179780  1        1.086177
## log_absences    1.105625  1        1.051487
## Medu            1.169208  3        1.026397
## reason          1.133327  3        1.021079
## age             1.217190  1        1.103264
## goout           2.892651  2        1.304139
## romantic        3.271947  1        1.808852
## goout:romantic 5.759781  2        1.549179
```

NCV of model is 0.001, but since we have large number of samples, we are good.

Language - MS

```
# stepwise regression
print("Both :n Forward and Backward selection")
```

```
## [1] "Both :n Forward and Backward selection"
```

```
model_ms_lang <- lm(outcome ~ ., data = ms_lang)
both_model_ms_lang <- ols_step_both_p(model_ms_lang)
```

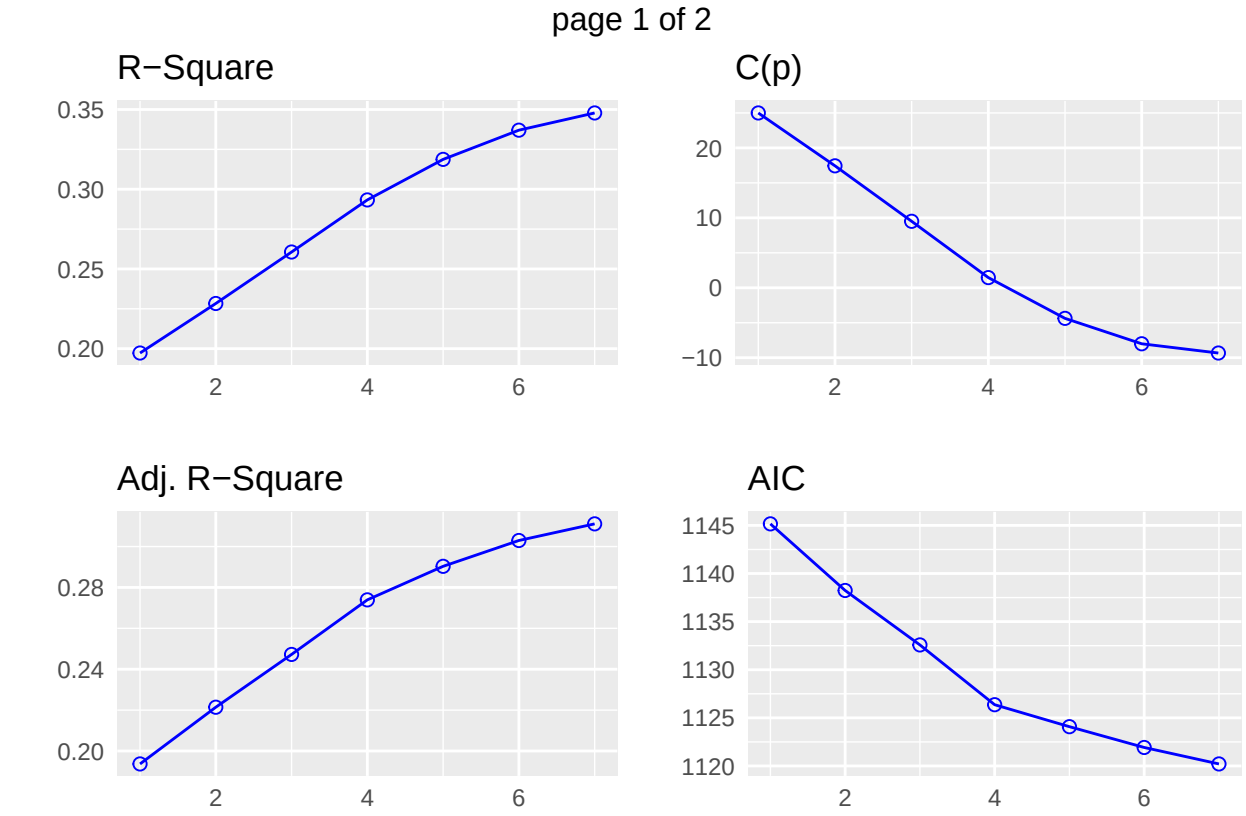
```
## Stepwise Selection Method
## --------------------------
##
## Candidate Terms:
##
## 1. sex
## 2. age
## 3. address
## 4. famsize
## 5. Pstatus
## 6. Medu
## 7. Fedu
## 8. Mjob
## 9. Fjob
## 10. reason
## 11. guardian
## 12. traveltime
## 13. studytime
## 14. failures
## 15. schoolsup
## 16. famsup
## 17. paid
## 18. activities
## 19. nursery
## 20. higher
## 21. internet
## 22. romantic
## 23. famrel
## 24. freetime
## 25. goout
## 26. Dalc
## 27. Walc
## 28. health
## 29. log_absences
##
## We are selecting variables based on p value...
##
## Variables Entered/Removed:
##
## - failures added
## - higher added
## - famrel added
## - studytime added
## - Fedu added
## - guardian added
## - sex added
##
## No more variables to be added/removed.
##
##
```

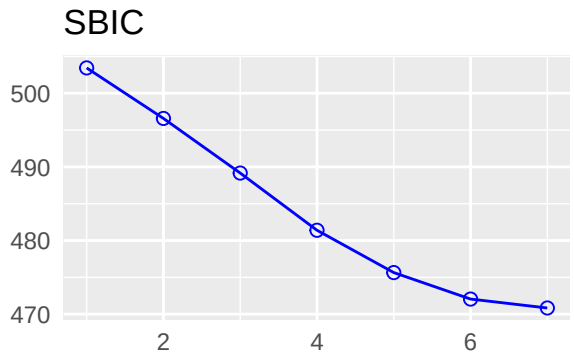
```
## Final Model Output
## ------------------
##
##                         Model Summary
## ----------------------------------------------------------------
## R                       0.590       RMSE                 2.793
## R-Squared               0.348       Coef. Var           26.538
## Adj. R-Squared          0.311       MSE                  7.800
## Pred R-Squared          0.268       MAE                  2.051
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## -------------------------------------------------------------------
##                 Sum of
##                Squares         DF    Mean Square      F        Sig.
## -------------------------------------------------------------------
## Regression     886.155         12         73.846    9.467    0.0000
## Residual      1661.462        213          7.800
## Total         2547.616        225
## -------------------------------------------------------------------
##
##                                    Parameter Estimates
## ----------------------------------------------------------------------------------------------
##           model      Beta    Std. Error    Std. Beta      t        Sig      lower     upper
## ----------------------------------------------------------------------------------------------
##     (Intercept)     9.146         0.756                12.099    0.000     7.656    10.636
## failures1+2+3+4    -3.041         0.485       -0.384   -6.271    0.000    -3.997    -2.085
##       higheryes     1.585         0.546        0.175    2.906    0.004     0.510     2.660
##         famrel4     1.397         0.481        0.208    2.904    0.004     0.449     2.345
##         famrel5     0.281         0.526        0.038    0.534    0.594    -0.756     1.318
##      studytime2     0.126         0.431        0.019    0.291    0.771    -0.724     0.975
##    studytime3+4     1.499         0.591        0.160    2.536    0.012     0.334     2.663
##           Fedu2     0.394         0.452        0.056    0.872    0.384    -0.497     1.286
##           Fedu3     1.256         0.588        0.130    2.135    0.034     0.097     2.416
##           Fedu4     1.656         0.633        0.163    2.617    0.010     0.408     2.904
##  guardianmother    -0.832         0.424       -0.117   -1.962    0.051    -1.667     0.004
##   guardianother     0.800         0.954        0.055    0.838    0.403    -1.081     2.681
##            sexM    -0.765         0.407       -0.109   -1.878    0.062    -1.568     0.038
## ----------------------------------------------------------------------------------------------
both_model_ms_lang
```

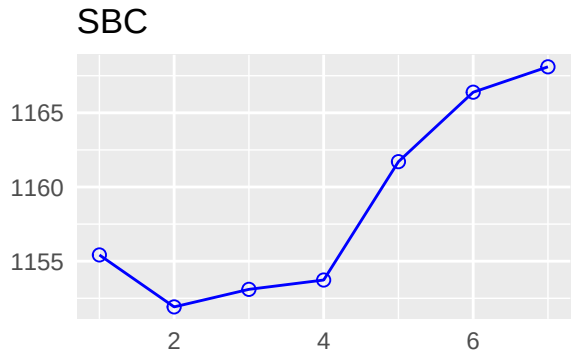
```
##
##                                Stepwise Selection Summary
## ----------------------------------------------------------------------------------------
##                       Added/                   Adj.
## Step    Variable     Removed     R-Square    R-Square     C(p)         AIC         RMSE
## ----------------------------------------------------------------------------------------
##    1    failures     addition       0.197       0.194    25.0010    1145.1590    3.0215
##    2     higher      addition       0.228       0.221    17.4400    1138.2369    2.9691
##    3     famrel      addition       0.261       0.247     9.5060    1132.5786    2.9195
##    4    studytime    addition       0.293       0.274     1.4550    1126.3662    2.8673
```

```
##     5       Fedu        addition        0.319        0.290     -4.3730     1124.0815     2.8347
##     6    guardian       addition        0.337        0.303     -8.0080     1121.9224     2.8093
##     7        sex        addition        0.348        0.311     -9.3300     1120.2117     2.7929
## ---------------------------------------------------------------------------------------------
```

```
plot(both_model_ms_lang)
```








```
summary(both_model_ms_lang$model)
```

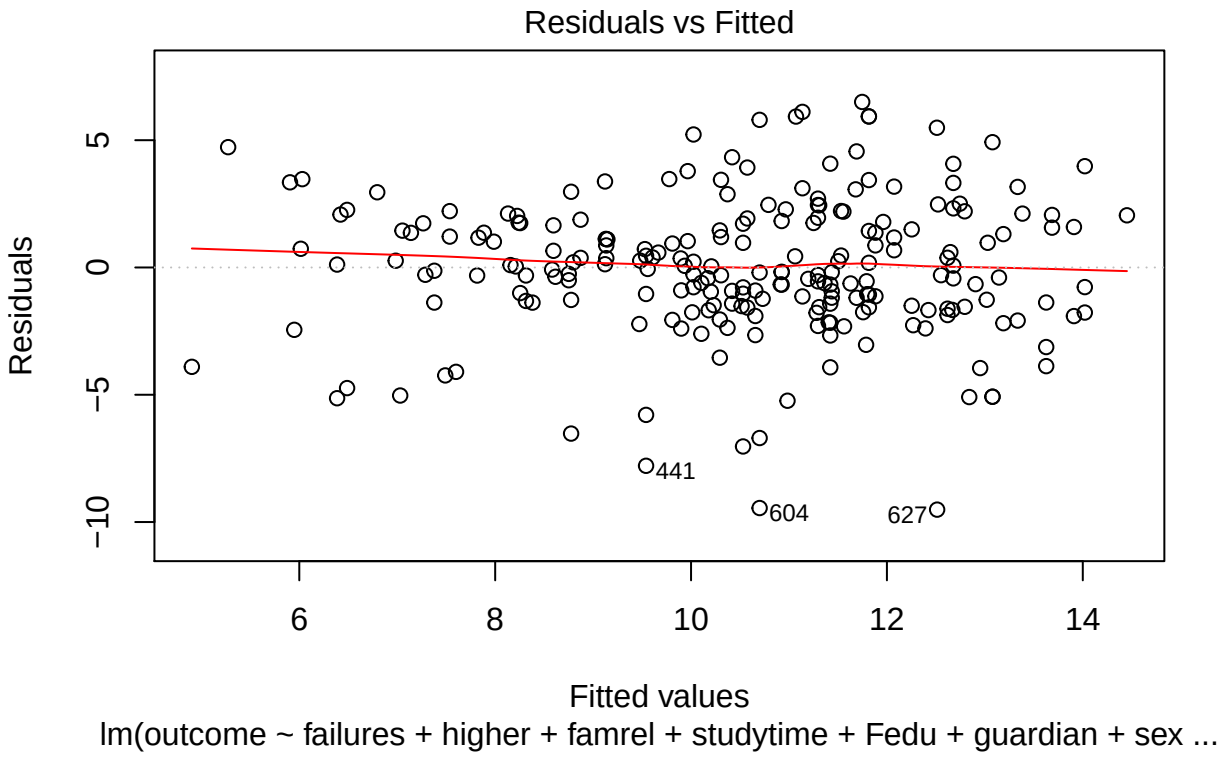
```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5127 -1.4976 -0.0695  1.7480  6.5023
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       9.1459     0.7559  12.099  < 2e-16 ***
## failures1+2+3+4  -3.0407     0.4849  -6.271 1.97e-09 ***
## higheryes         1.5852     0.5455   2.906  0.00405 **
## famrel4           1.3967     0.4809   2.904  0.00407 **
## famrel5           0.2809     0.5261   0.534  0.59403
## studytime2        0.1255     0.4308   0.291  0.77105
## studytime3+4      1.4986     0.5909   2.536  0.01193 *
## Fedu2             0.3943     0.4523   0.872  0.38432
## Fedu3             1.2563     0.5883   2.135  0.03387 *
## Fedu4             1.6562     0.6330   2.617  0.00952 **
## guardianmother   -0.8316     0.4239  -1.962  0.05110 .
## guardianother     0.7996     0.9543   0.838  0.40304
## sexM             -0.7651     0.4074  -1.878  0.06177 .
```

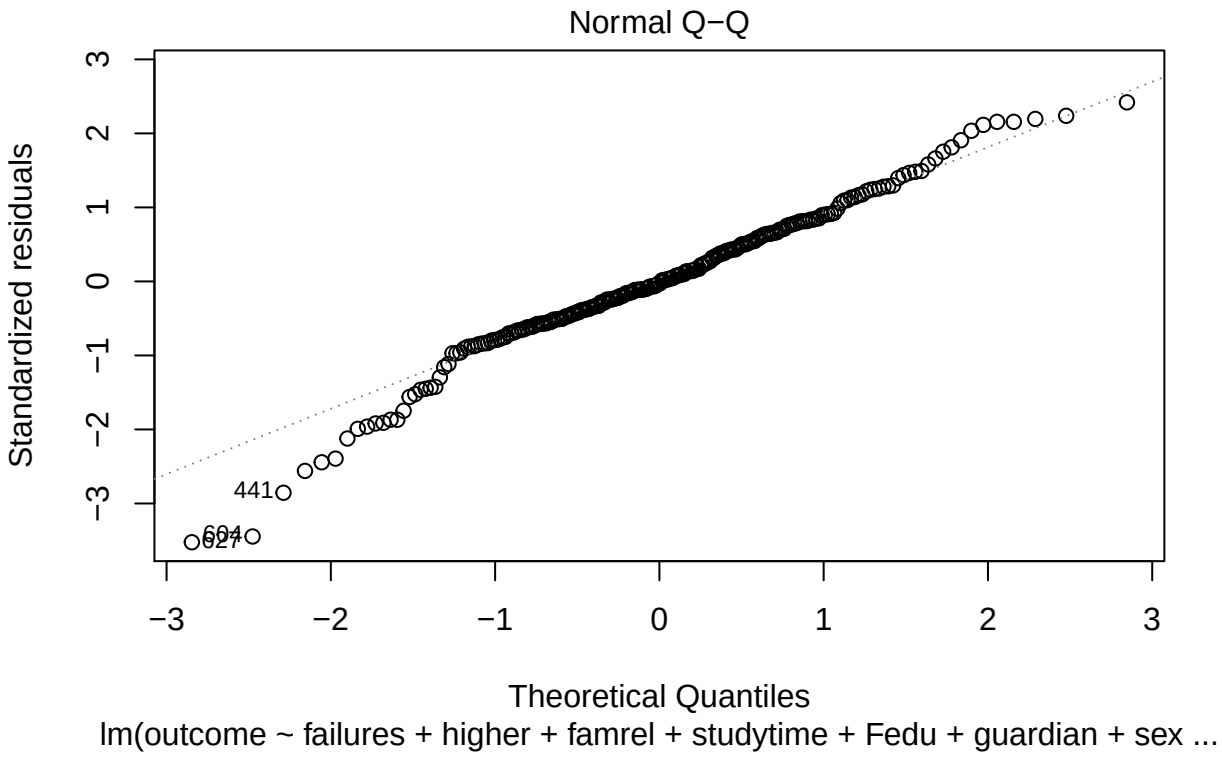
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.793 on 213 degrees of freedom
## Multiple R-squared:  0.3478, Adjusted R-squared:  0.3111
## F-statistic: 9.467 on 12 and 213 DF,  p-value: 1.237e-14
```

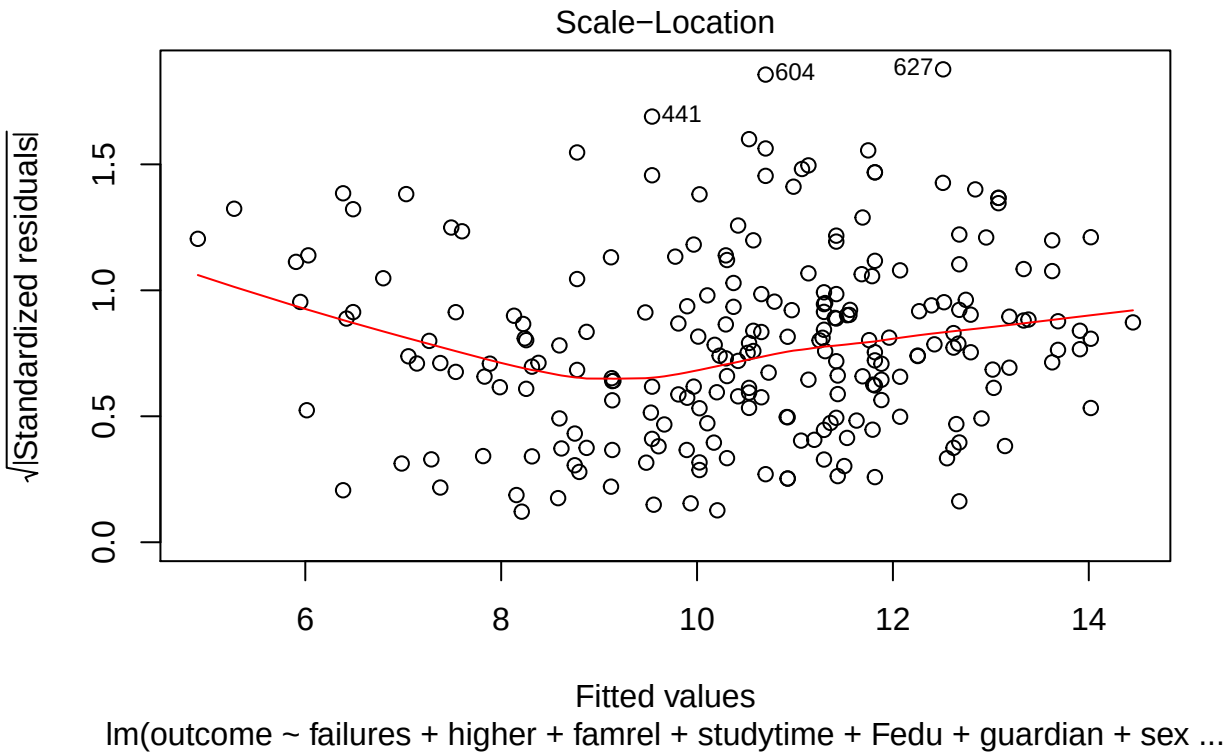
```
# Stepwise selection included 12 predictor variables
# This may lead to overfitting. Lets make a linear model with top 10 features used by stepwise regressi

final_model_lang_ms <- lm(outcome ~ failures + higher + famrel + studytime + Fedu + guardian + sex, data
summary(final_model_lang_ms)
```

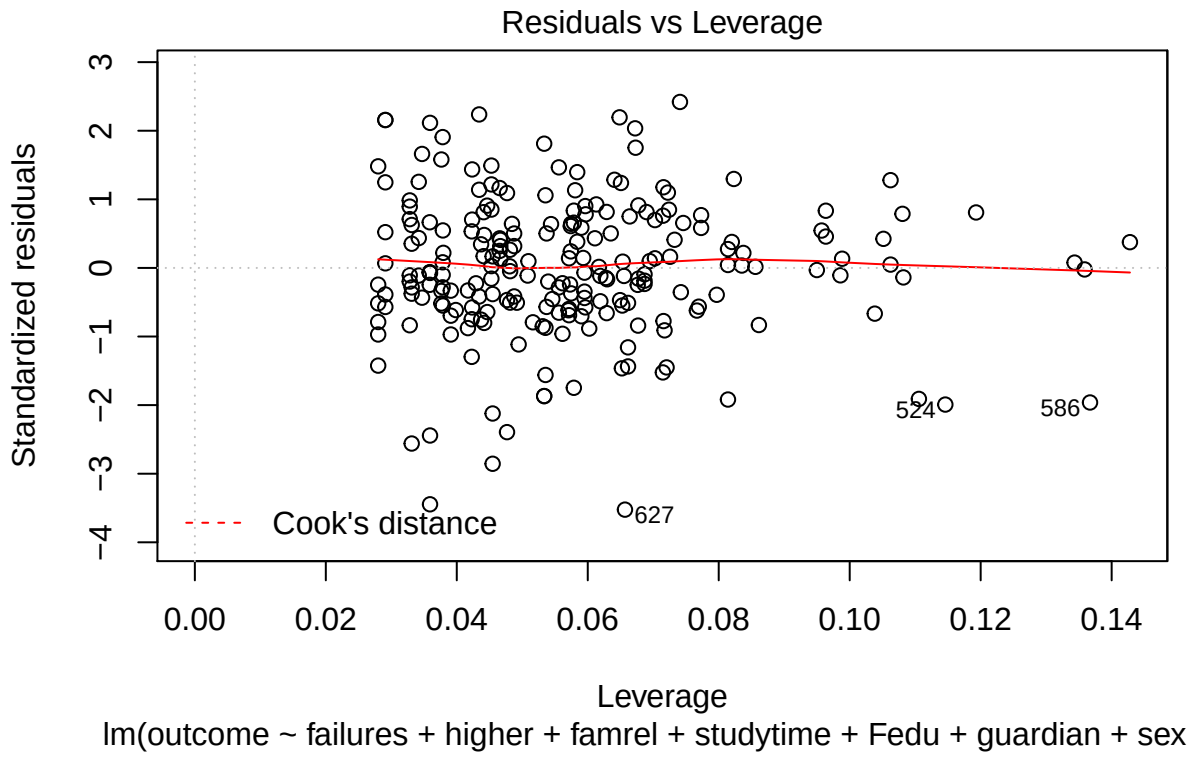
```
##
## Call:
## lm(formula = outcome ~ failures + higher + famrel + studytime +
##     Fedu + guardian + sex, data = ms_lang)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5127 -1.4976 -0.0695  1.7480  6.5023
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       9.1459     0.7559  12.099  < 2e-16 ***
## failures1+2+3+4  -3.0407     0.4849  -6.271 1.97e-09 ***
## higheryes         1.5852     0.5455   2.906  0.00405 **
## famrel4           1.3967     0.4809   2.904  0.00407 **
## famrel5           0.2809     0.5261   0.534  0.59403
## studytime2        0.1255     0.4308   0.291  0.77105
## studytime3+4      1.4986     0.5909   2.536  0.01193 *
## Fedu2             0.3943     0.4523   0.872  0.38432
## Fedu3             1.2563     0.5883   2.135  0.03387 *
## Fedu4             1.6562     0.6330   2.617  0.00952 **
## guardianmother   -0.8316     0.4239  -1.962  0.05110 .
## guardianother     0.7996     0.9543   0.838  0.40304
## sexM             -0.7651     0.4074  -1.878  0.06177 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.793 on 213 degrees of freedom
## Multiple R-squared:  0.3478, Adjusted R-squared:  0.3111
## F-statistic: 9.467 on 12 and 213 DF,  p-value: 1.237e-14
```

```
plot(final_model_lang_ms)
```


Residuals vs Fitted
Residuals
5
0
−5
−10
441
604
627
6
8
10
12
14
Fitted values
lm(outcome ~ failures + higher + famrel + studytime + Fedu + guardian + sex ...


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(outcome ~ failures + higher + famrel + studytime + Fedu + guardian + sex ...
441
604
627
−3
−2
−1
0
1
2
3


Scale−Location
604
627
441
√|Standardized residuals|
0.0
0.5
1.0
1.5
6
8
10
12
14
Fitted values
lm(outcome ~ failures + higher + famrel + studytime + Fedu + guardian + sex ...




```
ncvTest(final_model_lang_ms)
```

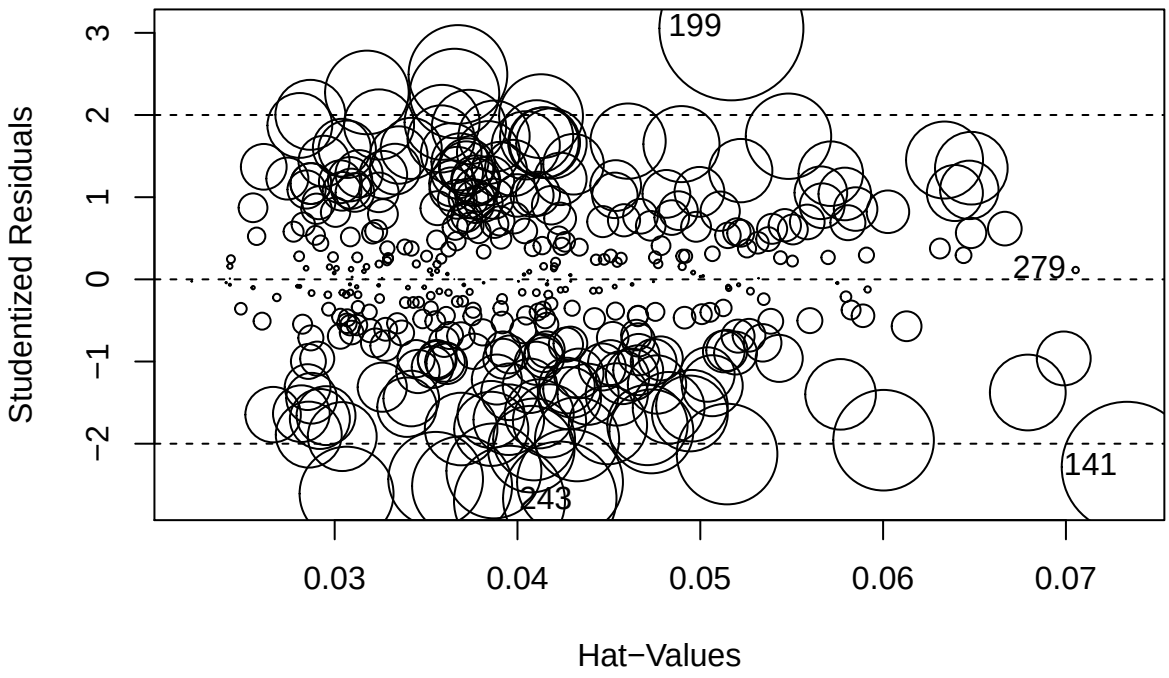
```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.574976, Df = 1, p = 0.44829
```

```
vif(final_model_lang_ms)
```

```
##                GVIF Df GVIF^(1/(2*Df))
## failures  1.223072  1        1.105926
## higher    1.180428  1        1.086475
## famrel    1.070635  2        1.017209
## studytime 1.224959  2        1.052036
## Fedu      1.186878  3        1.028966
## guardian  1.281355  2        1.063941
## sex       1.099842  1        1.048734
```

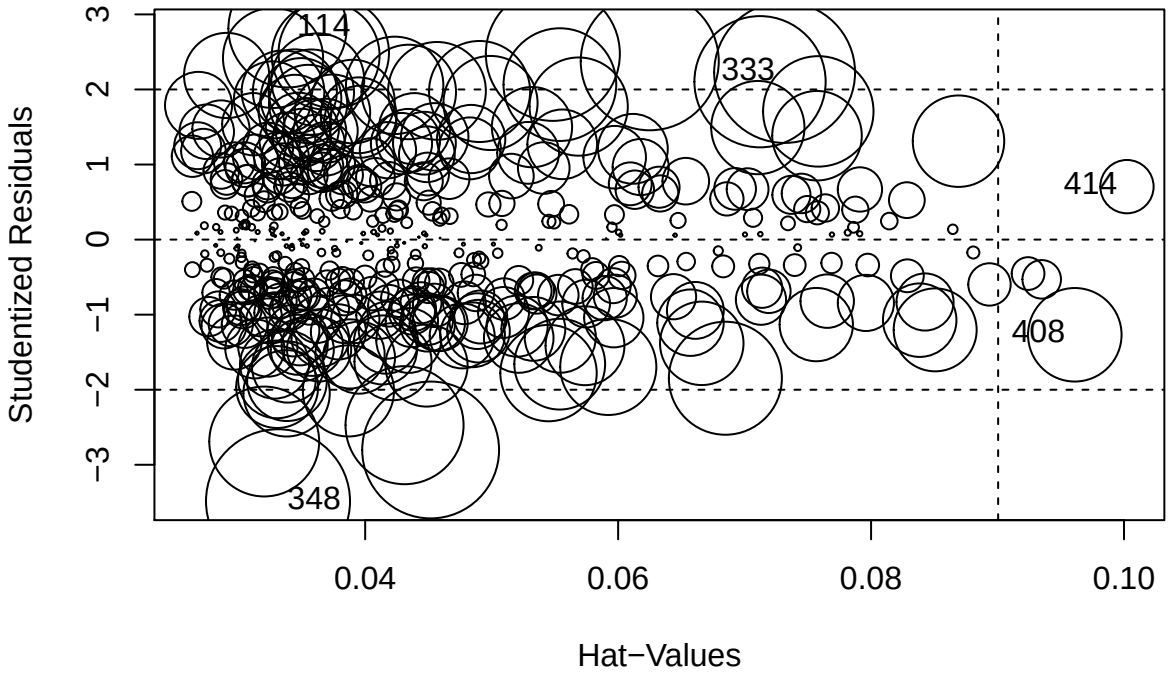
Diagnostic

```
influencePlot(final_model_math)
```



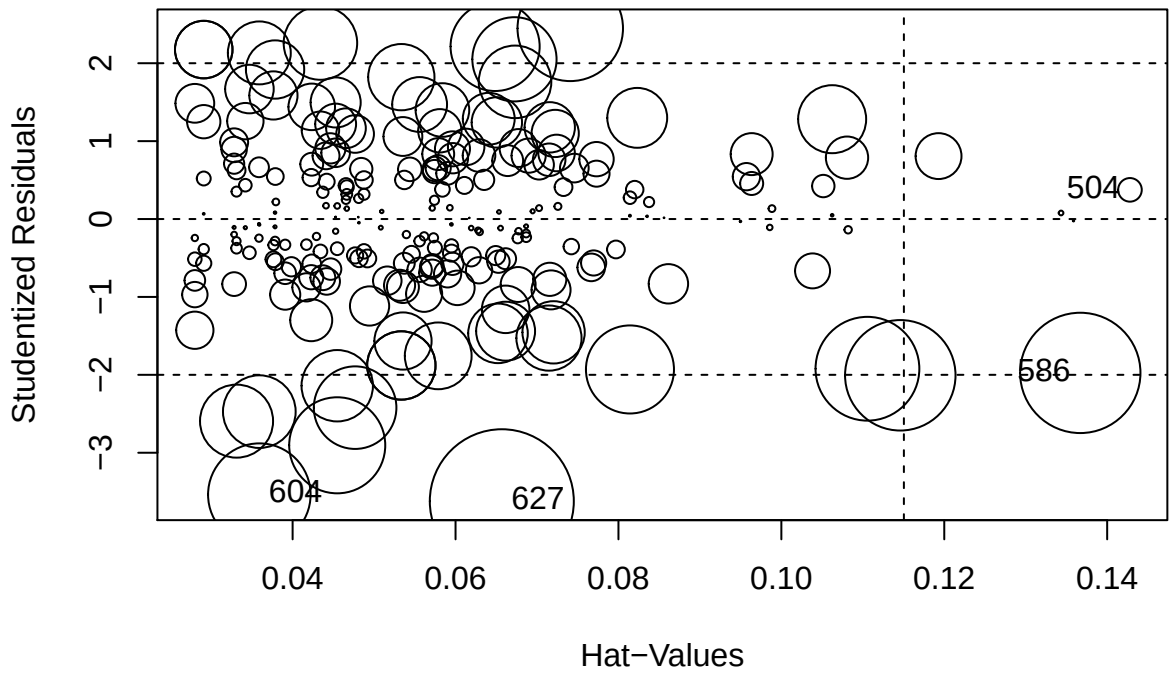

```
##         StudRes         Hat        CookD
## 141 -2.2838809 0.07333810 2.551632e-02
## 199  3.0546184 0.05170117 3.110869e-02
## 243 -2.7002383 0.03956234 1.846410e-02
## 279  0.1133592 0.07053286 6.110642e-05
```

```
influencePlot(final_model_lang_gp)
```




```
##        StudRes        Hat       CookD
## 114  2.8140275 0.03385812 0.014359267
## 333  2.2260645 0.07316359 0.020387880
## 348 -3.4856816 0.03310730 0.021306630
## 408 -1.2681714 0.09610746 0.008986438
## 414  0.7054168 0.10023220 0.002921168
```

```
influencePlot(final_model_lang_ms)
```




```
##         StudRes        Hat       CookD
## 504   0.3746621 0.14279658 0.001806036
## 586  -1.9750497 0.13670068 0.046875650
## 604  -3.5380149 0.03591184 0.034027195
## 627  -3.6226221 0.06568118 0.067144054
```