Penalized Regression

The task of this exercise is to predict the performance of students in a math course. Based on these predictions, students in need of additional support are assigned to private lessons. The files `student-mat-train.Rdata` and `student-mat-test.Rdata` contain data about student achievements from Portuguese schools. They contain information about the math grade, socio-economic characteristics of the students, and school related features. Table 1 contains the description of the variables (see `https://archive.ics.uci.edu/ml/datasets/Student+Performance` for a more detailed data description).

Table 1: Description of the Variables

Variable	**Description**
G3	final math grade (numeric: from 1 to 20)
sex	student's sex (binary: 0 = male and 1 = female)
age	student's age (numeric: from 15 to 22)
address	student's home address type (binary: 0 = urban and 1 = rural)
Pstatus	parent's cohabitation status (binary: 0 = living together and 1 = apart)
Medu	mother's education (numeric: from 0 to 4)[a]
Fedu	father's education (numeric: from 0 to 4)[a]
famsize	family size (binary: = 0 if 3 or less family members and = 1 if more than 3 family members)
famrel	quality of family relationships (numeric: from 0 - very bad to 4 - excellent)
traveltime	home to school travel time (numeric: = 0 if $<$ 15 min, = 1 if 15 to 30 min, = 2 if 30 min to 1 hour and = 3 if $>$ 1 hour)
studytime	weekly study time (numeric: = 0 if $<$ 2 hours, = 1 if 2 to 5 hours, = 2 if 5 to 10 hours and = 3 if $>$ 10 hours)
failures	number of past class failures (numeric: = n if $0 \leq n < 3$, else = 4)
schoolsup	extra educational school support (binary: = 0 if no and = 1 if yes)
famsup	family educational support (binary: = 0 if no and = 1 if yes)
activities	extra-curricular activities (binary: = 0 if no and = 1 if yes)
paid	extra paid classes (binary: = 0 if no and = 1 if yes)
internet	Internet access at home (binary: = 0 if no and = 1 if yes)
nursery	attended nursery school (binary: = 0 if no and = 1 if yes)
higher	wants to take higher education (binary: = 0 if no and = 1 if yes)
romantic	with a romantic relationship (binary: = 0 if no and = 1 if yes)
freetime	free time after school (numeric: from 0 - very low to 4 - very high)
goout	going out with friends (numeric: from 0 - very low to 4 - very high)
Walc	weekend alcohol consumption (numeric: from 0 - very low to 4 - very high)
Dalc	workday alcohol consumption (numeric: from 0 - very low to 4 - very high)
health	current health status (numeric: from 0 - very bad to 4 - very good)
absences	number of school absences (numeric: from 0 to 93)

Note: [a] 0 = none, 1 = primary education (4th grade), 2 = 5th to 9th grade, 3 = secondary education or 4 = higher education

Group Home Assignment (max. 4 points)

The mandatory group home assignment has to be submitted before 12:00 o'clock prior PC-session 3. It is obligatory to solve the assignment in R Markdown with `echo = TRUE` option for every code chunk and generate a PDF file with the solution. Generating an HTML file from the R Markdown and converting it into the PDF is also possible. Make sure that the final PDF file has no readability issues. The PDF with the answers to the six questions below as well as the file with the R Markdown code has to be submitted via Canvas.

Download the data sets `student-mat-train.Rdata` and `student-mat-test.Rdata` from Canvas. Load the data into R. Install and load the packages `glmnet` and `corrplot`.

1. How many observations are in the training and test data? (0.5 points)
2. What is the average, minimum, and maximum grade in the training data? (0.5 points)
3. Plot the histogram of the final math grades in the training data. (0.5 points)
4. Explain shortly the difference between causal and predictive modelling. (0.5 points)
5. Choose five variables that you consider most relevant to predict the final grade. Estimate two models by OLS, the first with your chosen set of variables, the second including all first order interactions. Discuss the in-sample fit of the two models. (1 point)
6. Choose another five variables, **add** them to your variables set and generate two additional models analog to exercise 5). Split your data into a training and estimation sample. Plot both, the in-sample and out-of-sample fit. Which of the four model performs the best? (1 point)