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

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

Exercise 1: Data Description and Visualization

1. How many observations are in the training data?
2. What is the average, minimum, and maximum grade in the training data?
3. Plot the histogram of the final math grades in the training data.
4. Plot the correlation between all variables. Which variables have the highest correlation?

Exercise 2: OLS

1. Estimate the linear regression model in which you regress the final grade on all the covariates by OLS using the training sample. Compute the MSE in the test sample. Why do we calculate the MSE in the test data and not in the training data?
2. Does the coefficient `schoolsup` have the expected sign? Which sign would you expect in a causal model? Does this imply that our prediction model is misspecified?
3. Try to find an OLS model which uses fewer covariates and has a lower MSE in the test sample.

Exercise 3: Lasso and Ridge

1. Estimate the linear regression model as before with the lasso penalty. Choose the optimal λ by a 5-fold cross validation. Compute the MSE in the test sample.
2. Estimate the lasso model with different seeds for the cross validation. Do the models always select the same regressors? What are the potential reasons?
3. Estimate the linear regression model as before with the ridge penalty. Choose the optimal λ by a 5-fold cross validation. Compute the MSE in the test sample.
4. Compare the OLS and Ridge coefficients for the pair of variables with the highest correlation. What is the effect of the ridge penalty on the coefficients?
5. Compare the MSE of all the three methods (OLS, Lasso and Ridge). Which one yields the best predictive model?
6. Plot the predicted grades (by OLS, lasso and ridge) against the actual grades and compare the methods.

Useful links:

- A description of the `glmnet` package is here: `https://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html`.