Project 1: Exam score prediction

Description

This dataset includes three test scores of students at a public school and a variety of personal and socio-economic factors that may have interaction effects upon them.

1. Gender: Gender of the student (male/female)
2. EthnicGroup: Ethnic group of the student (group A to E)
3. ParentEduc: Parent(s) education background (from some_highschool to master's degree)
4. LunchType: School lunch type (standard or free/reduced)
5. TestPrep: Test preparation course followed (completed or none)
6. ParentMaritalStatus: Parent(s) marital status (married/single/widowed/divorced)
7. PracticeSport: How often the student parctice sport (never/sometimes/regularly))
8. IsFirstChild: If the child is first child in the family or not (yes/no)
9. NrSiblings: Number of siblings the student has (0 to 7)
10. TransportMeans: Means of transport to school (schoolbus/private)
11. WklyStudyHours: Weekly self-study hours(less that 5hrs; between 5 and 10hrs; more than 10hrs)
12. MathScore: math test score(0-100)
13. ReadingScore: reading test score(0-100)
14. WritingScore: writing test score(0-100)

Analytical questions (you may only answer some of them but properly addressing more will improve the rate of your report):

1. Using variables 1-11 as the covariates to predict Math, Reading and Writing scores.
2. Which factors (features) affect each test score signficantly? Are there interacting effects?
3. Are the optimal prediction models similar or different across the three test scores? Is it possible to leverage one score as the auxiliary information to learn the model for another score (still its model against variables 1-11) better?

Suggestions and tips:

In the report, you should describe your final model and interpret its parameters in an accurate and useful manner. It is expected that you would first examine the marginal distributions and pairwise relationships between variables (e.g., to check to see whether any nonlinearities are immediately obvious), that you would explore several candidate models, and explain why you selected your model. Also, you should check for violations of regression model assumptions, influential observations, multicollinearity, etc. In addition, there are a few missing entries in the data set of this project, you could simply drop them or impute them with some mean values. It would be great if you could be an active learner and try to figure out some more appropriate and systematical ways to handle them.

It would be helpful to be clear about your motivation for carrying out certain analyses as well as to be clear about interpretations of fitted model parameters. Your report should include a table summarizing parameter estimates associated with your final fitted model, characterizing predictor variables in a way that a reader can clearly understand.

Below you'll find some aspects to be addressed in your report. These are just a few suggestions, but feel free to add your own input/creativity to the analysis:

- Data exploration: descriptive statistics and visualization. You might want to, for instance:
    - Include a descriptive table with summary statistics for all variables;
    - Explore the distribution of the outcome and consider potential transformations (if necessary);
    - See if there are any unusual observations and consider them as potential outliers/influential points.
- In your regression model, be watchful for variables that are highly correlated and be selective in the variables you will include in your analysis.
- Consider selective interactions between variables.
- DO NOT IGNORE MODEL DIAGNOSTICS.