Question 2. Using information on students in a large data science course, the following equation was estimated

$$\begin{aligned}\widehat{score}_i = &\underset{(3.68)}{13.98} + \underset{(0.78)}{11.25} gpa_i + \underset{(1.26)}{2.57}\, hsgpa_i + \underset{(0.122)}{0.742}\, sat_i - \underset{(0.040)}{0.157}\, work_i \\ &+ \underset{(0.78)}{4.41}\, mathstat_i - \underset{(0.796)}{0.728}\, mothcoll_i + \underset{(0.766)}{0.18}\, fathcoll_i \\ &n = 814, \quad R^2 = 0.4194\end{aligned}$$

where the dependent variable, $score_i$ is the course total as a percentage of total points possible. The explanatory variables are the grade point average at the beginning of term (gpa_i), high school performance grade point average ($hsgpa_i$) and SAT score (sat_i), hours of work per week ($work_i$), a binary variable for whether a student has taken a math&stat course ($mathstat_i$), and binary indicators for whether mother and father have bachelor's degrees ($mothcoll_i$, $fathcoll_i$).

1. Interpret the coefficient on math&stat and decide whether its estiamted effect seems reasonable.

$$t = \frac{4.41}{0.78} = 5.7$$

   It is an explanatory variable that is highly correlated even if we think of it in common sense. Looking at the results, the coefficient value is relatively large and the t stat is high, so the estimated effect is considered reasonable.

2. Does high school performance (grade point average or SAT score) help predict performance in data science? You are also told that $hsgpa_i$ varies significantly, depending on quality of high school's education. How does this information affects your data science modeling strategy?

$$t \text{ of } hsgpa_i = \frac{2.57}{1.26} = 2.04, \qquad t \text{ of } sat_i = \frac{0.742}{0.122} = 6.08$$

   The t stat of the two explanatory variables is greater than 2, so it is large enough. Therefore, both variables can be said to be significant.
   A perfect score of gpa is 4.3 or 4.5. And the perfect score for the SAT is 1600(800 points in math alone). If sat_i is an unscaled SAT score, although the coefficient is a relatively small value, it has a dominant effect on the dependent variable.
   The issue of quality of education in high school is worth considering at $hsgpa_i$. But since the SAT is a standardized test, I don't think that issue is an problem at sat_i.
   $hsgpa_i$ may be underestimated or overestimated depending on the level of the individual's school. Therefore, in order to cotroll this effect, it is necessary to compare the academic achievements of each school. As a method, it can be used as a variable by averaging the SAT scores of students by school, determining the rank of the school, and multiplying it by a weight in $hsgpa_i$.

3. Researcher A claims that the lower R^2 is due to omitted variable. One of which is elementary school GPA ($egpa_i$). How do you value the claim?
   High school gpa is already used as an explanatory variable. Scores from older courses will have a higher correlation with high school gpa. It seems good not to use it even considering multicollinearity.
   Let's think in terms of common sense. If both the past and the present have high scores, and both are low scores, multicollinearity will be induced. If the score was high in the past but is now low, it can be considered that the achievement in higher education is low because the score is low now. Therefore, if the coefficient of $egpa_i$ comes out as a positive, it can be interpreted that the experience of studying well in the past has a positive effect even if a individual do poorly in high school. It makes sense, but I don't think it would have a high coefficient value.

4. When $mothcoll_i$ and $fathcoll_i$ are dropped from the equation, the R^2 becomes 0.4188. Is there any evidence that having a parent or both parents with a college degree helps predicting performance in data science, having controlled for the other explanatory variables? How do you interpret the signs of coefficients for parent's higher education?

$$t \text{ of } mothcoll_i = \frac{-0.728}{0.796} = 0.915, \qquad t \text{ of } fathcoll_i = \frac{0.18}{0.766} = 0.23$$

   Both variables have very low t stats. Therefore, it can be regarded as an insignificant variable.
   Social statistics show that among high-achieving students, there are many cases where parents are highly educated or have good economic background. Parental education may be a necessary but not sufficient condition for a student's academic achievement.

5. Researcher B argues that, in addition to $egpa_i$, $mothcoll_i$ and $fathcoll_i$, parent's education upto graduate school or differentiating college majors must be better variables than simple dummy indicator of bachelor's degrees. Provide your rebuttal. as answered above, Parental education may be a necessary but not sufficient condition for a student's academic achievement.
   The following is excerpted from the contents of ⌜ The Effect of Parents' Academic Background on Children's Academic Achievement: The mediation of poverty, delinquency, the parent-child relationship and self-esteem⌟.
   "First, it was found that, although parental education directly affects children's academic achievement, it also influences the academic achievement level by mediating variables such as **parent-child relationship and self-esteem**. Second, it was found that parents' educational background mediates **poverty and delinquency** and affects the level of academic achievement."
   It seems better to add a mediator variable like the quote above than to subdivide the parents' educational background.

6. The school collects survey statistics from students that include each student's address and family's income. Researcher C wants to use address as an indicator for the family's wealth level, and in combination of family income, she believes both variables can be good instruments for $mothcoll_i$ and $fathcoll_i$. For the claim to be valid, what are the necessary conditions?
   The condition of the instrumental variable should have a high correlation with the target explanatory variable and low correlation with the error term.
   I think it's a good try. However, although there is some correlation between parents' educational background and parents' economic power, it is thought that it will not be at an absolute level.

7. An after-school education center for computer coding advertises that earlier exposure to coding is critical for student's performance in data science. In the advertisement, it says coding class participation yields $R^2 = 0.90$ in a single variable regression to $hsgpa_i$. Given that, as a parent with college degree, will you support your children's coding class for data science career?
   Since it is a single variable regression, even if R^2 is high, the model is not reliable. I don't even know the coefficient values.
   I think it will allow the child to attend a coding class for interest and experience. It's not a for data science career.

References

[1] Jung Seb Lee, Yong Gyo Lee, The Effect of Parents' Academic Background on Children's Academic Achievement: The mediation of poverty, delinquency, the parent-child relationship and self-esteem, Korean Journal of Family Welfare Vol.16 No.4 (2011) 65-88