Question 2. (A part of Q1 in MSc DS Prep exam Fall 2021) Using information on students in a large data science course, the following equation was estimated

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

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 mathstat course ($mathstat_i$), and binary indicators for whether mother and father have bachelor's degrees ($mothcoll_i$, $fathcall_i$).

1. Interpret the coefficient on mathstat and decide whether its estimated effect seems reasonable.

   [**Answer**] Interpretation of $mathstat_i$'s coefficient is :
   The **average** treatment effect of $mathstat_i = 4.41$ holding other regressors fixed(Cetris Paribus).

   $$\text{t-stat}_{ms} = \frac{\hat{\beta}}{S.E} = \frac{4.41}{0.78} = 5.65$$

   Reject H_0 since t-stat$_{ms} > 1.96$ at the 5% level of significance, $mathstat_i$ is significant.
   The $mathstat_i$'s estimated effect seems reasonable. The $score_i$ is expected to be high when a student has taken a mathstat course, and the positive sign reflects this.

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?

   [**Answer**] First, check the variables are significant by t-test.

   $$\text{t-stat}_{hsgpa_i} = \frac{2.57}{1.26} = 2.03 \qquad \text{t-stat}_{sat_i} = \frac{0.742}{0.122} = 6.08$$

   Both reject H_0 at 5% significance level. But if significance level is 1%, then hsgpa can not reject H_0. This is because $hsgpa_i$ is not very helpful in predicting performance in data science or $hsgpa_i$'s S.E is big.

   $$\begin{cases} hsgpa_i\text{'s big S.E} \rightarrow \text{small swing of data} \\ sat_i\text{'s small S.E} \rightarrow \text{big swing of data} \end{cases}$$

   There will be more information that sat_i can explain the $score_i$ than $hsgpa_i$ since S.E$_{hsgpa_i}$ > S.E$_{sat_i}$ which means $hsgpa_i$'s swing < sat_i's swing. For example, if a school gives good grades to most students through absolute evaluation, $hsgpa_i$'s swing ↓ and the S.E ↑. In this case, $hsgpa_i$ would not be able to explain the score well, so the t-value was ambiguous.

   If $hsgpa_i$ varies significantly, depending on quality of high school's education, then $hsgpa_i$ has endogeneity since omitted variable bias. so A3 assumption is violated. Then we can't use our model because $\hat{\beta}$ will be inconsistent and biased. To solve this problem, We have to find data such as school grades or find instrumental variable and add them to the model.

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?

   [**Answer**]
   I think grades in elementary school will not affect grades in college. $egpa_i$ will have a time effect because there is a big time lag between elementary school and university.
   From a vector space perspective, $egpa_i$ will explain the information except for the parts explained by the $hsgpa_i$ and sat_i, but it is unlikely to help explain the $score_i$. And since the correlation between the three variables is high, using $egpa_i$ does not seem to significantly increase R^2. Considering that the significance of $hsgpa_i$ was also ambiguous, $egpa_i$ is probably not significant.

4. When $mothcoll_i$ and $fathcall_i$ 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 parents' higher education?

   [Answer] First, check the variables are significant by t-test.

   $$\text{t-stat}_{mothcoll_i} = \frac{0.728}{0.796} = 0.91 \qquad \text{t-stat}_{fathcoll_i} = \frac{0.18}{0.766} = 0.23$$

   Do not reject H_0, Both are not significant.
   Perform F-test to determine if there is multicollinearity.

   $$\begin{cases} H_0 : \beta_7 = \beta_8 = 0 \\ H_1 : \text{at least one of } \beta_j \neq 0 \end{cases}$$

   $$F = \frac{(RRSS - URSS)/2}{URSS/(814-8)} \quad \sim F(2, 806)$$

   If the null hypothesis is not rejected, there is multicollinearity.
   But, when $mothcoll_i$ and $fathcall_i$ are dropped from the equation, the R^2 becomes 0.4188. This means that even if $mothcoll_i$ and $fathcall_i$ are dropped from the model, the explanatory power of the model does not differ. And Occam's Razor proposes that the model with fewer parameters to be preferred to the one with more. Then, we can avoid unnecessarily reducing the degree of freedom and avoiding overfitting too.
   As a result of the t-test and the F-test, $mothcoll_i$ and $fathcall_i$ are not significant, so there is no 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.
   It is strange that the signs of the coefficients of the $mothcoll_i$ and $fathcall_i$ are different. Since the two variables reflect parents' enthusiasm for education, a positive sign is expected. However, the $mothcoll_i$'s negative sign is because the correlation between $mothcoll_i$ and $fathcall_i$ is high. However, the sign is not important because the $mothcoll_i$ and $fathcall_i$ are insignificant as a result of the t-test.

5. Researcher B argues that, in addition to $egpa_i$, $mothcoll_i$ and $fathcall_i$, parents' education upto graduate school or differentiating college majors must be better variables than simple dummy indicator of bachelor's degrees. Provide your rebuttal.

   [Answer]
   It would be better if we could be divided into variables that only capture subspace that significantly explain $score_i$. But in reality it will be very difficult. There is a limit to explaining only with dummy variables. Whether the variables we find are significant can be determined through the F-test.

6. The school collects survey statistics from students that include each student's address and family's income. Researcher C wants to use the 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 $fathcall_i$. For the claim to be valid, what are the necessary conditions?

   [Answer] For Researcher C's claim to be valid, the IV must meet the following two conditions :
   ① Validity : IV needs to be uncorrelated with the error
   ② Relevance : IV needs to be correlated with the endogenous regressor
   However, the address does not meet ① Validity, i.e the address will be correlated with the error.
   For example, the form of owning a house may be different. The house may belong to them or may be monthly rent. In addition, there may be a gap between the rich and the poor in the same city, so there is a problem of not being able to decide how far to split the address.

   Another thing to think about here is that as a result of the t-test/F-test in (4), the $mothcoll_i$ and $fathcall_i$ were insignificant.

   $$Var(\widehat{\beta_{IV}}) = \frac{\sigma^2}{n \cdot \text{Var}(x) \cdot Corr(x, IV)^2}$$

   The Relevance can not be 100% $\Rightarrow \text{Var}(\widehat{\beta_{IV}}) > \text{Var}(\widehat{\beta_{OLS}})$
   Thus, using IV does not make it significant. Instead, the endogeneity problem could be solved.
   But the best way is to find better data and put it in the model. And if two variables are simply dropped from the equation, omitted variable bias can occur.

7. An after-school education center for computer coding advertises that earlier exposure to coding is critical for students' 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?

   [**Answer**]
   I will not support my children's coding class for data science career because Correlation $\neq$ Causality.
   $R^2 = 0.90$ means high correlation, not causality.
   Correlation = Causality, when data is randomly sampled and Cetris Paribus.