Question 2-1

[Answer] From low $R^2 = 0.4194$ and the number of samples($n = 814$), the linear regression model seems to somewhat explain the dependent variable $score_i$ in data science course. The average score is 13.98 (from constant). From t test to coefficients, $gpa(11.25/0.78 = 14.42 > 1.96)$, $hsgpa(2.57/1.26 = 2.04 > 1.96)$, and $mathstat(4.41/0.78 = 5.65 > 1.96)$ has strong positive correlation with $score$. On the contrary, the hour of studying or $work(-0.157/0.04 = -3.93 < -1.96)$ shows strong negative correlation with $score$. However, mother's bachelor degree $mothcoll(-0.728/0.796 = -0.91 > -1.96)$, $sat(0.742/0.122 = 0.61 < 1.96)$ and father's bachelor degree $fathcoll(0.18/0.766 = 0.23 < 1.96)$ show meaningless correlation with $score$. I think that the negative correlation from $work$ seems wrong as generally studying a course long time increases its score. For estimating causality we have to do randomized experiment on data independently from the treatment for each factor. Even after that there seems to be omitted variables. $work$ has endogeneity. These faults violates $GMA3R_{mi}$.

Question 2-2

[Answer] As shown above, $hsgpa$ shows strong causality with score in data science course while sat has small causality. Therefore, $hsgpa$ is helpful to predict score in data science. If $hsgpa$ depends on quality of high school's education, we can rely on two stage linear regression and can capture the exact influence of $hsgpa$ on score in data science course by adjusting the error. In this case, an instrumental variable called $hseduq$ can be used as a regressor. For two stage linear regression, we have to check the instrumental validity of $hseduq$ ($E(hseduq_i u_i) = 0$). (In this case, there is only one instrumental variable it is not necessary to check intrumental relevance.) If the original linear regression is like:

$$\begin{aligned}\widehat{score_i} = 13.98 &+ 11.25gpa_i + 2.57hsgpa_i + 0.742sat_i - 0.157work_i \\ &+ 4.41mathstat_i - 0.728mothcoll_i \\ &+ 0.18fathcoll_i + u_i\end{aligned} \tag{1}$$

Now $\widehat{hsgpa}$ is regressed by:

$$\widehat{hsgpa_i} = \beta_0 + \beta_1 hseduq_i + v_i \tag{2}$$

After estimating $\widehat{hsgpa}$ from the above linear regression, we can change the original linear regression like this:

$$\begin{aligned}\widehat{score_i} = 13.98 &+ 11.25gpa_i + \gamma_2\widehat{hsgpa_i} + 0.742sat_i - 0.157work_i \\ &+ 4.41mathstat_i - 0.728mothcoll_i \\ &+ 0.18fathcoll_i + u_i\end{aligned} \tag{3}$$

After regression, γ_2 will more precisely explain the influence of $hsgpa$ on score in data science course by capturing the portion of $hseduq$ from u_i.

Question 2-3

[Answer] The claim that low R^2 is caused by omitted variables is right. However, $egpa_i$ is not a good explanatory variable for score in data science course. I think that $egpa_i$ can be used as instrumental variable to $hsgpa_i$ as a person with higher gpa in elementary school may have more chance to have high gpa in high school also. They are the same vector in the vector space point of view. Rather, I think that the time for studying data science in a week ($dswork_i$) or homework points for data sciences course ($dshwp_i$) are more explanatory. Some of their effects are explained by $work_i$ but still most of their effect are contained in u_i.

Question 2-4

[Answer] As R^2 does not have significant change ($0.4194 \to 0.4188$), two explanatory variables $mothcoll_i$ and $fathcoll_i$ have no explanatory power on score in data science course if other explantory variables are controlled. If the coefficient of $mothcoll_i$ is γ_5 and $fathercoll_i$ is γ_6, whether we have the influence of having parents with bachelor degree on score in data science course can be explained by f test where H_0: $\gamma_5 = \gamma_6 = 0$ and H_1: $\gamma_5 \neq 0, \gamma_6 \neq 0$. Its test statistic is:

$$\frac{(RRSS - URSS)/q}{URSS/(n-p)} = \frac{(RRSS - 0.4194)/2}{0.4194/(814-8)} \tag{4}$$

We compare the above value with the critical value of $F_{1-\alpha}(2, 814-8)$ and check. If test statistic is less than the critical value so we accept the null hypothesis, which means that having parents with bachelor degrees is not helpful to predict the score in data science course. The reason why the sign of the coefficient in $mothcoll_i$ is negative and why the explanatory power of $mothcoll_i$ and $fathcoll$ is different is there are omitted variables like $wealth$ and $famlyincome$ having influence on $mothcoll_i$ and $fathcoll$. It is a result of mis-specification.

Question 2-5

[Answer] Suppose that parents' education upto graduate school is $pgedu_i$ and parents' differentiating college major is $pdmajor_i$. I think that in the vector space point of view, $mothcoll_i$, $fathcoll_i$, $pgedu_i$, and $pdmajor_i$ are the same. Therefore, adding those variables into linear equation does not improve R^2 of the linear regression.

Question 2-6

[Answer] Suppose that an indicator for the family's wealth level is $wealthtown_i$ and the family income is $fincome_i$. If we want to use $wealthtown_i$ and $fincome_i$ as the instrumental variables to $mothcoll_i$ and $fathcoll_i$, we have to check their instrumental validity. In other words, we have to show

$$Cov(wealthtown, u) \neq 0 \tag{5}$$

and

$$Cov(fincome, u) \neq 0 \tag{6}$$

Also we can check which instrumental variable is better by checking their instrumental relevance from comparing their followings:

$$\frac{Cov(wealthtown_i, score_i)}{Cov(wealthtown_i, mothcoll)} \tag{7}$$

or

$$\frac{Cov(fincome_i, score_i)}{Cov(fincome_i, mothcoll)} \tag{8}$$

in the case of $mothcoll_i$. In $fathcoll_i$, it is similar.

Question 2-7

[Answer] I cannot support coding class for data science career. First, the linear regression for $hsgpa_i$ just use a single variable regression. There are many other omitted variables (Spurious Regression). The regression is wrong and we cannot rely on its R^2 value.