STA502: Math & Stat for MBA
Problem Set 5

Question 1. Consider the following hypothetical example. A researcher at SIAI wants to evaluate the effectiveness of a math program in high school designed to increase number of students studying mathematics and statistics (M&S) at college and university. Some secondary schools offer the program and others do not

1. You have data from a sample of 25-year olds. The data include each individual's current labor market income and whether or not the secondary school the individual attended offered the (M&S) program at the time the individual attended. Write down the equation for the bivariate regression using these two variables, with labor market income as the outcome variable. Define the variables precisely. How would you interpret the regression coefficients (describe them as if it is given to a general audience)? Would you expect the slope coefficient to capture the causal effect of explanatory variable on the outcome? Why or why not?

2. A colleague makes the following claim: "You do not need to worry about establishing causality. The data were gathered using a random sample, so they are not biased." Critically assess this claim.

3. Another colleague worries that for the secondary schools that offer M&S programs, the quality of teaching varies significantly among schools. Without controlling for the quality of M&S programs, she believes the analysis is likely going to suffer from omitted variable bias. Evaluate her assertion.

4. The program was funded by tax revenues raised by local districts. Therefore, you gather additional data and control for average district income in your regression. First, explain as if to a general audience what it means to control for a variable in a regression. Second, describe what you expect would happen to the slope coefficient you described above when you control for income. Clearly state any assumptions you make.

Solution.

1. The bivariate regression equation is:

$$\text{LaborMarketIncome}_i = \alpha + \beta\, M\&S\, d_i + \epsilon_i$$

where $\text{LaborMarketIncome}_i$ is the labor market income of individual i and $M\&S$ and d_i is an indicator variable equal to 1, if individual i attended a secondary school that offer the M&S program, α is the constant in the regression equation and ϵ_i is the error term, which captures all of the other factors that determine individual i's labor market income.

The estimated coefficient $\hat{\alpha}$ tells us the average income for those individuals who did not attend a secondary school offering the program and $\hat{\beta}$ tells us the mean difference in the average income for those who did. We would not expect this to capture the causal effect of attending a school with the program. These is likely significant selection bias: individuals who attended schools with the program in the program almost likely to differ systematically from those who did not. For example, they or their parents may have cared more about M&S education and thus chosen to send them to such schools.

2. The claim is incorrect. Out colleague appears to be confusing random samples with random experiments or any other plausibly random assignment of a treatment. A randomized experiment eliminates selection bias, and thus allows us to establish causality, by assigning treatment status independently from potential outcomes. A random sample selects members of the population to tbe in the sample with a known probability (in a simple random sample, this probability is constant for all individuals in the population). While appearing in the sample is unrelated to treatment status (or related in a known way), a random sample puts no restrictions on the relationship between treatment status and potential outcomes. Selection bias is still a concern.

3. Due to the lack of pertinent regressors, the validity of M&S education in the secondary schools may not be captured in the regression. What can be offered alternatively is an instrumental variable to circumvent the endogeneity, although the choice of right instrument is questionable, when it comes to quality of teaching.

4. Controlling for a variable: we would like to compare individuals who are identical in every way but for the fact that some went to schools with the program and others did not. For those characteristics we can observe, like district income, we can achieve this by matching, that is, comparing individuals with the same observable characteristics, differing only in their treatment status. Controlling for a characteristic in a multivariate regression is an automated way to do this. One can also describe using average income eliminates all the variation in treatment status that can be explained by district income (e.g. richer districts are more likely to have the program) and then looking at the relationship between the remaining variation in treatment status and income. This quetion looks for an indication that the student understands what multivariate regression really does.

   In addition, we can expect the program to be more common in wealthier districts. We would also expect, all other things equal, that students from wealthier districts would also tend to earn more. Taken together, this would lead to positive OVB: the regression that did not include district income would overstate the slope coefficient.

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

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

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 (act_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 estimated effect seems 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?

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?

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 parents' higher education?

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

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 $fathcoll_i$. For the claim to be valid, what are the necessary conditions?

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?

Solution.

Before we answer the problem, check the below first :
(1) Check the coefficient and their sign : We expect that the coefficient sign of work should be negative, and the others are positive. But the coefficient sign of $mothcoll_i$ not. We should test the significance of the variable.

(2) Check the standard error : $s.e.(\hat{\beta}) = \sqrt{(X'X)^{-1}\sigma^2}$. If $(X'X)$ is large, $s.e.(\hat{\beta})$ smaller. In the model, $hsgpa_i, mothcoll_i, fathcoll_i$ have high standard error, we should make conclusion that the range of the $hsgpa_i$ is narrower than the other variables. (In general, standard error of dummy variables are higher than the continuous variables)

(3) Check the p-value : By definition, p-value = P(statistic T is more extreme than the critical value | when H_0 is true). the p-value of $mothcoll_i, fathcoll_i$ is higher than significance level $\alpha = 0.05$, hence we couldn't reject the null hypothesis that there's no significance of the variables.
Test statistic T for the significance of coefficient : $T = \frac{\hat{\beta}-0}{s.e.(\hat{\beta})} \sim t_{(806,\frac{\alpha}{2})}$.
Although $hsgpa_i$ is significant under the $\alpha = 0.05$, T statistic is 2.04, and its p-value is 0.021. Hence it isn't significant under the $\alpha = 0.02$.

(4) Check the sample size and coefficient of determination : n = 814, $R^2 = 0.4194$. Sample size is large sufficiently, and R^2 isn't high enough but nor low.

1. Coefficient sign of $mathstat_i$ is positive, which we expected. And the T = $\frac{4.41}{0.78} = 5.65 > t_{(806,0.025)}$, we think the $mathstat_i$ is significant variable to explain the $score_i$.
   Interpretation : If student has taken a math&stat course, the expected score on "average" is 4.41 higher than who hasn't taken.

2. $hsgpa_i$ and sat_i is significant under the $\alpha = 0.05$, and the coefficient sign is also positive. Although the estimate of $hsgpa_i$ is greater than sat_i, t-statistic of $hsgpa_i$ is much lower than sat_i, since the standard error of $hsgpa_i$ is much larger than sat_i. We might guess that the information of $hsgpa_i$ is smaller than sat_i.
   If $hsgpa_i$ varies depending on quality of high school's education, $hsgpa_i$ has endogeneity because of omitted variable bias. More variable that explains the quality of high school's education required.

3. Researcher A's claim would be true, since $R^2 = 0.4194$, which isn't high enough. But is really $egpa_i$ raises R^2 significantly? First, $esgpa_i$'s vector space isn't quite different from gpa_i and $hsgpa_i$. It measures the pure $esgpa_i$, subtracting the effect of gpa_i and $hsgpa_i$. Besides, we think the pure $esgpa_i$ wouldn't significant, since there's a long time gap between elementary school and undergraduate.

4. F = $\frac{(RSS_U - RSS_F)/r}{RSS_F/(814-8)} = \frac{(1-0.4188)-(1-0.4194)/2}{(1-0.4194)/806} = 0.4165 \sim F_{(2,806)}$. Since $F_{(2,806,0.05)} = 3.00$, we conclude that these variables are jointly insignificant.
   Focusing on analytic side, that variables reflect the interest of education. Since both variables are insignificant, we should find the other substitute variable.

5. If the differentiated college major and parent's education up to graduate school is significant subspace of $egpa_i, mothcoll_i, fathcoll_i$, Researcher B's argument would be true. But extracting the significant subspace using only dummy variables is difficult, since dummy variable has less information than the continuous variables generally. Finding the suitable subspace of $egpa_i, mothcoll_i, fathcoll_i$ is the key.

6. We should check the validity and relevance of instrumental variables, $address_i$ and $income_i$. In terms of validity, $address_i$ and $income_i$ themselves don't affect the $score_i$. But in terms of relevance, is correlation of $address_i$ and $fathcoll_i, mothcoll_i$ high? Type of residences, how specific the address(city, district, town... even in the same town, each person's wealth and interest of education is different) is important.

7. Correlation does not imply causation. For a causal effect to be established, the data must be randomly sampled and "Ceteris Paribus" also be satisfied. Simple regression is extremely dangerous, because it shouldn't guarantee the random sampling and Ceteris Paribus conditions. In fact, we expect there's omitted variable bias and simultaneity. Besides, the high R^2 is derived from $hsgpa_i$, not $score_i$. We already concluded that the $hsgpa_i$ is weakly significant, hence after-school center's claim is still questionable.