Question 1. In coding manic countries without fundamental science in education, lack of mathematically trained data scientist is a big problem causing students to fail to get as quality education as they should. In SIAI's seminar paper of 2023, after running nation's first mathematical data science program for two years, an experiment is designed to improve math absence in data science education. Out of 120 students, half were chosen at random to take part in the experiment.

For the treated students, teachers were from mathematics department or trained by hardcore math in their graduate school education. However, the students in the control group were given classes by coding maniacs without proper mathematical statistics training at all. They all have elementary level mathematics, and the class materials are only copied from Github codes. Students in both treatment and control groups were tested before and after the experiment to measure how much they had learned.

Denote by $X_i = 1$ students in the treatment group and $X_i = 0$ those in the control group. The results reported below are the coefficient and standard error on X in regressions of the form:

$$y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + \epsilon_i$$

Where W_i are other controls. The results for different dependent variables are as follows:

	Dependent variable	Estimated coefficient on X (standard error)	Other controls
(1)	Fraction of Math trained teachers Prior to experiment	0.02 (0.10)	None
(2)	Teacher Quality	1.37 (2.01)	None
(3)	Students test score Prior to experiment	0.02 (0.10)	None
(4)	Fraction of Math trained teachers During experiment	0.20 (0.03)	None
(5)	Students test score After experiment	0.16 (0.10)	None
(6)	Students test score After experiment	0.15 (0.06)	Yes - student's score prior to experiment

1. Why do you think SIAI report the results presented in rows (1)-(3). What should you be looking for in these results?

2. Comment on the results presented in rows (4)-(6).

3. Compare the results reported in rows (5) and (6). Offer an explanation for the differences you observe.

4. Suppose you are interested in the relationship between student test scores, T_i, and the fraction of math-trained teachers, M_i, You propose the model:

$$T_i = \gamma_0 + \gamma_1 M_i + u_i$$

   Explain why a simple linear regression of T_i on M_i is likely to lead to an inconsistent estimate of γ_1. Explain how instrumental variables can help us obtain a consistent estimate of γ_1. What would you use as the instrument and why is this instrument valid?

5. Using the results presented in the Table above, what will be the IV estimate of γ_1?

Question 2. Consider the following specification

$$A_t = \beta_0 + \beta_1 Y_t + \epsilon_t$$

where A_t is aggregate value of a variable at time t, and Y_t is aggregate value of another variable at t. We would like to estimate the marginal propensity of A to Y, β_1, but we also know that $Y_t = A_t + B_t + C_t$, where B_t is aggregate bs at time t and C_t is aggregate cs at time t. Assume that ϵ_t is a mean zero, iid error distributed $N(0, \sigma_\epsilon^2)$. It is believed that A_t and Y_t are endogenous with respect to the disturbance ϵ, while variables B_t and C_t are weakly exogenous.

1. **True, False or Uncertain:** "Estimating a regression of A on Y using OLS will yield inconsistent and downward biased estimates of β_1." (Hint: $W_t = B_t + C_t$)

Question 3.

1. You regress Data Science exam score as post-SIAI effect on the same individual's engineering training before SIAI. You observe a number of the individual's undergraduate classmates. You have a valid instrument for individual's pre-PDSI engieering education. You first estimate the OLS model and then the IV model. You find the OLS model yields a negative effect of pre-SIAI engineering training on post-SIAI Data Science exam performance, but the IV estimate yields an even more negative estimate. Why?

2. You are regressing y on k exogenous regressors including a constant. The k-th regressor is measured with white-noise error. This biases the OLS estimates of the impact of the regressors down. True or false?