

Please produce a report that answers the problems and add any code or documentation as an appendix. The report should be in pdf-format and approximately 10 pages. Make sure not to collaborate.

Problem 1: (Weight 20 pct, about 2 pages)

Consider the statements:

- “Controlling for variable z in a multiple regression, when the outcome variable of y and the regressor of interest is x , can approximately be thought about as regression of the independent variation in y on the independent variation in x .”
- “Controlling for a binary variable z in a multiple regression, when the outcome variable of y and the regressor of interest is x , amounts to more or less the same thing as deducting the group specific means - with groups defined by z - before running a simple regression of y on x .”
- “In a model where y is the outcome and x is the variable of interest, but where the effect of x on y depends on z , we can not interpret the estimates from a model where we leave out the interaction term between x and z .”

Please write a text that clarifies these statements: Are they true or false, if so, why, and could the statements be more precise?

Problem 2: (Weight 60 pct, about 6 pages - or a bit more if you have many graphs)

- a. (10 pct) Please read the paper “Does studying economics make you rich?” Explain the empirical strategy of the paper. What are the key assumptions for identification of the causal effects of the choice of major on future earnings?
- b. (10 pct) Please discuss the statement: “The study is only relevant for the subpopulation that chooses at least two bachelor level courses in economics and the empirical strategy does not even identify an average treatment effect for this group.”
- c. (40 pct) Attached is an artificial data set that contains data for future earnings, GPA for the two economics courses (“ x ” in the data set) and a dummy variable for the choice of major (“ $treat$ ” in the data set). Like in the paper, there is a discontinuity in the probability of treatment for $x=2.1$. Please perform a regression discontinuity analysis of these data. We want an analysis that is robust to “whimsical assumptions”, that is, perform the analysis in such a way that it is convincing that your results do not depend on arbitrary assumptions. We also want investigations into the key identifying assumption of the analysis. Please remember that graphs are key ingredients in regression discontinuity analyses.

Problem 3: (Weight 20 pct, approximately 2 pages)

There exists a large literature on the estimation of average treatment effects under the assumption of “selection on observables”. This key assumption is that the expected differences in the outcomes as treated and untreated for two individuals with the same observable characteristics will be equal to the average treatment effect. In general, we allow for the average treatment effect to be a function of these observables.

To actually estimate an average treatment effect with some precision, it may be necessary with some additional assumptions. Explain why it is sufficient to assume that we either specify the outcome equation correctly (in terms of how it depends on observables) or that we specify the treatment assignment equation correctly to successfully estimate the average treatment effect - and try to be more precise than the current phrase here: “successfully estimate the average treatment effect”.

Make a small simulation study that illustrates how we can set up the estimation of an average treatment effect in this “doubly robust” manner and show, using different data generating processes, that this robustness holds in practice, in contrast to for example a simple procedure based on “regression adjustment”.