Report\_5

Alvaro Zapata Rojas

20173414

What is the research question of both articles?

The Hünermund work addresses the resent trend in ML research for methods that work with large number of regressors and help to determine which are the most relevant such as Lasso and l2 boosting. The authors call these methods Double Machine Learning (DML). They highlight the fact that these methods will throw biased results if correlation between regressors are not included. If those correlations are not included the authors call for a naive Lasso. Thus, the best approach is to focus the analysis based on orthogonal, or doubly robust, moment conditions that are insensitive to approximation errors stemming from regularization. This paper explores the consequences of violations of *unconfoundedness* for DML due to the presence of bad controls in the conditioning set and focus on the LASSO case.

This assumption can easily be violated, if X includes variables that are not fully exogenous. In this paper, we explore the consequences of violations of unconfoundedness for DML due to the presence of bad controls in the conditioning set.

In a first step, we make precise the notion of bad controls in regression analyses. We then show in simulation studies that DML is very sensitive to the inclusion of bad controls.

bad controls, although they do not necessarily exert a causal influence, are often highly correlated with the treatment or the outcome (since they are related to unobservable that affect D or Y ).

o that researchers cannot rely on simple rules of thumb for variable inclusion. Instead, each potential control requires its own careful identification argument based on economic theory, which is difficult to provide if the feature space is large and ultimately undermines the purpose of automated model selection.

One particularly common strategy to identify the ACE is to control for confounding influence factors via covariate adjustment. This strategy can be rationalized with the help of the backdoor criterion.

Deleting edges emitted by D ensures that all directed, causal paths between D and Y are removed from G. -> backdoor criterion. X is backdoor admissible if it blocks every path between D and Y.

Consequently, a backdoor admissible set X blocks all non-causal paths between D and Y , while leaving the causal paths intact.

What are the strengths and weaknesses of both papers' approaches to responding to that question?

The Hünermund paper first expose the consequences of using bad control variables from a theorical explanation. The theory exposed says that to get the Avarege Causal Efect (ACE) one must control the confounding influence factor with help of the backdoor criterion. On a broad explanation this criterion consist in blocking all non-causal paths between D (treatment) and Y (outcome), while leaving the causal paths intact.

Ten the authors put the theory into practice by simulating data where the correlation between the controls and the treatment is higher correlation than with the outcome. This way they show which method selects fewer bad controls in every causal scenario.

Finally, they apply the same approach to empirical data on wage gap. The marital status gets picked differently as a control variable in every case to which the authors suggest the necessity of economic theory to justify the picking of control variables.

How do the papers advance knowledge about the question, i.e., what is the contribution? (If you can't find any contributions, ask yourself why the editor and the referees decided to publish the articles).

This paper shows that the selection of a control variable is not that automated as it is presumed despite the popularity of DML methods. I think that this study shows how sensitive conclusions are in relation to the methods used.

What would be one or two specific and valuable next steps to move forward on this question?

A further step is to consider these results in data-driven causal learning based on a minimal set of assumptions where AI learns from causal discovery algorithms.

Since these methods are “imported” from other disciplines outside economics, one must consider which causal model to use in what case. As authors suggest: “DML is a highly effective tool within a particular domain, but needs to be considered in a broader context of possible causal structures that can occur in applied empirical settings” (Hünermund 2022).

What is the research question of both articles?

Cinelli 2022 is a crash course explanation of the implication of bad controls in regressions. This problem arises when one must decide whteher to include one variable or not. Good controls are variables that must be added to the regression equation to eliminate what came to be known as “omitted variable bias”. Bad controls produce unintended discrepancies between the regression coefficient and the effect that the coefficient is expected to represent. Thus, “bad controls” are variable that could be affected by the treatment and amplify the bias of the results. The authogs use directed acyclic graphs (DAGs) to portrait in an interactive way the the proble of bad controls and how to avoid them.

Cinelli expands on Húnermund by the concept of blocking the spurious paths between X and Y to estimate causal effect of X on Y. This will be the guiding principle for deciding whether or not Z should be included in the regression equation,

bad controls might just as well be dependent variables too. Good controls are variables that we can think of having been fixed at the time the regressor of interest was determined.

What are the strengths and weaknesses of both papers' approaches to responding to that question?

One strength is that the Cinelli paper assists the work of Hünermund(2022) by giving a more friendly explanation (for example when explaining main sources of association and its conditioning implications: Mediators, Common causes and common effects).

Also Cinelli expands on the various forms of control by presenting 18 models where the Z (variable which one must decide whether to add or not) is in different causal association an could be either a good control, bad control and neutral.

Finally it sets various examples of how previous studies have led to bias conclusion.

How do the papers advance knowledge about the question, i.e., what is the contribution? (If you can't find any contributions, ask yourself why the editor and the referees decided to publish the articles).

This is a well written article with practical examples on the field of causality. This contributes by making it easier to understand how to identify good controls and bad controls.

What would be one or two specific and valuable next steps to move forward on this question?

Sometimes is difficult to plot causal effects on a DAG plot and a sensitive analisis must be implemented.