Report 4

Alvaro Zapata

20173414

What is the research question of the article?

The article presents the usefulness of the double-lasso method to resolve the problem of too many covariates in dataset. It can be related that covariate controls by the researchers can denote “p-hacking” and bias results. Thus, the double-lasso offers a rigorous method to better selection of variables for the regressions. To demonstrated the reach of the double-lasso method, the authors apply it to four datasets from 4 different studies and demonstrate the usefulness on variable selection in both correlational analyses and experimental designs. This approach helps to answer the question on how useful this method is in the research field.

What are the strengths and weaknesses of the paper's approach to answering that question?

First the authors run 10,000 regressions on a dataset with different approaches.

regression including no covariates (“none”),

regression including all the covariates when possible (“all”)

regression including all covariates selected in a step-wise regression (“stepwise”),

regression choosing covariates to maximize the chances of the independent variable being significant (“p-hacking”)

Two-step multiple regression

Double-forward regression

as a baseline regression using the correct variables (“true”).

With this empirical approach they find that double-lasso selects with more precision the right number of covariates than the previous methods named.

Another of the main strengths of the article is the use of replication of previous studies that didn’t use the double-lasso method to select the covariates of the studies. By this, the authors compare the results of the original study and using the double-lasso method. In the first analysis it helps to identifying whether there is sufficient empirical justification for including the published covariates and interactions between covariates. The second analysis shows double lasso helps to determine the mediator as the correct one if it is sensible to a more complete model. The third analysis shows that double-lasso reduces the likelihood to include spurious covariates. Finally, the fourth analysis demonstrated that double-lasso helps testing whether randomization of data was successful.

On the other hand, the article doesn’t expand on the fact that lasso-estimated results cannot determine which variables make logical sense within the theoretically realm.

How does this paper advance knowledge about the question, that is, what is the contribution? (If you can't find any contributions, ask yourself why the editor and referees decided to publish the article)

This article contributes as a simply and well explained guide to the benefits of using double-lasso method in large datasets and empirical research. It answers the question on what the benefits of this method is by replicating previous empirical studies to show the reach of the method.

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

On my opinion it would also be important to show with same rigor what the limits of the double-lasso are. It is also worth noting that it is important to also thinks what the selected variables after the double-lasso mean for the research so it helps for a better interpretation of the results of the regression.

Thus, whether the relationship between life satisfaction and parenthood was significantly positive, non-significant or significantly negative depended on whether and which covariates were included. One important consideration is whether the potential covariates should be considered controls, on theoretical and logical grounds (e.g, as opposed to being analyzed as potential mediators). Assuming that the variables are valid potential controls, it is not clear which covariates should be included and what conclusion should be drawn.

The double-lasso addresses this concern by identifying whether there is sufficient empirical justification for including the published covariates and potentially identifying other covariate