PharmaSUG 2017 - Paper HA-04

Two Roads Diverged in a Narrow Dataset...When Coarsened Exact Matching is More Appropriate than Propensity Score Matching

Aran Canes, Cigna[1] Corporation

ABSTRACT

Coarsened Exact Matching (CEM) is a relatively new causal inference technique that allows the researcher to non-parametrically create a matched dataset to evaluate the effect of a treatment. Propensity Score Matching (PSM) is the older, more established technique in the literature. Both methods have been turned into SAS® macros which are used by many SAS data scientists.

In one particular instance, I was tasked to deal with a dataset generated from pharmacy data where it seemed that, because PSM would not work due to the ratio of cases to controls, the only viable alternative would be the use of regression. However, this alternative was unappealing as it imposes a linear model on the data. Remarkably, I found an N:N CEM was able to match 100% of the case population even though the ratio of cases to controls was close to 1:1.

In this paper, I show why PSM was not feasible in this particular instance and why CEM provided a non-parametric alternative to regression. The paper contributes to the growing literature on the subject of non-parametric observational studies, both inside and outside the SAS community, by providing a real-life example of a seemingly intractable problem from the perspective of PSM which could in fact be solved by the use of CEM. Readers of this paper will be introduced to the techniques, and instead of general advice about their applicability, provided a real world example where one method (once implemented in SAS) was clearly preferable to the other.

INTRODUCTION

Even though causal inference is a relatively new branch of statistics, with Donald Rubin and Paul Rosenbaum's landmark paper The Central Role of the Propensity Score in Observational Studies for Causal Effects[2] only published in 1983, there is already a well-developed literature not only on propensity score analysis but other causal inference techniques as well.[3] There is even a wide variety of textbooks available for use in universities.[4] Much of this literature examines only the theoretical properties of different causal inference techniques, but there is also a wealth of papers comparing these approaches from a practical perspective, to which I have contributed previously.

While this paper does not try to resolve the question of which causal inference technique should be used in all circumstances, I do point out a real world case in which a particular method, Coarsened Exact Matching (CEM), was applicable where Propensity Score Matching (PSM) was infeasible. This is somewhat of a departure from the overall trend in the literature, where the dimensionality problem[5] is often cited to show why Propensity Score Analysis is the only practical method of choice because exact matching is impossible.[6] The key factor which makes CEM the preferred method is that it allows for N:N matching, which is impossible using a propensity score approach. I believe that, once familiar with CEM, many researchers will find similar instances where they can draw a causal inference which is impossible to find using PSM.

In this paper, I will briefly summarize the problem of causal inference, PSM and some reasons why such techniques have become recommended instead of simply performing regressions. I will then explain the mechanics of CEM to those who do not know of this causal inference technique. The rest of the paper will then be devoted to a real-world example that illustrates when CEM is a practical method while PSM is not feasible. I'll then conclude with a summary of the results and recommendations for further reading.

THE BASIC PROBLEM OF CAUSAL INFERENCE

When a new drug is developed or a particular promotional campaign is run, the researcher often wants to know what the effect of this treatment was. The basic problem of causal inference is that, while one can evaluate the outcomes of those who received the treatment, one cannot simultaneously observe these same individuals not receiving the treatment and so cannot directly assess the treatment effect.

The most statistically rigorous response is to run a randomized experiment, ideally with blocking[7], in which the case and control populations have similar characteristics on all key confounding variables (such as gender, age, or other confounders particular to a certain treatment) because those who received the treatment were selected at random.

Thus, if the populations are similar, or the same, on all variables which could have an effect on the outcome the researcher can find the Average Treatment Effect (ATE) simply by evaluating:

$$ATE=E(Y(1))-E(Y(0))$$

where 1 stands for those individuals who received the treatment and 0 stands for the control.

PROPENSITY SCORE MATCHING

The key insight of Rubin and Rubinstein that sparked the modern causal inference revolution is that one can simulate a randomized control trial on retrospective data by knowledge of the mechanism by which participants in the experiment chose to be either case or control. Thus, if outcomes for users of drug A are being compared with users of drug B, one may want to know what comorbidities each patient has, their age, gender, amount of prior year cost, etc. in order to determine who was prescribed drug A versus who was prescribed drug B. Once all the variables are known and measured that contributed to the assignment mechanism, a logistic regression can be run to give the probability that an individual used either drug A or drug B. This predicted probability can then represent the propensity score and patients can be matched based on similar propensity scores to create a simulation of a randomized controlled experiment. One can then find the ATE as:

$$ATE=E(Y(1)|X)- E(Y(0)|X)$$

where X are the confounding variables one is matching on.

WHY REGRESSION IS NOT A GOOD CAUSAL INFERENCE TECHNIQUE

Before learning new and somewhat complicated statistical techniques, or, so to speak, joining the causal inference revolution, one may ask why wouldn't one simply evaluate the outcome by performing a regression with the treatment as a dummy variable and controlling for the confounders by making them independent variables? There are many reasons why this is not a good choice. The most important are:

Performing a regression implies unrealistic assumptions about knowledge of the data generation process.[8] How does one know that the outcome is related to the treatment and independent variables linearly (as all linear regression assumes)? Even if the variables are related linearly how is one to know what particular functional form this relationship takes? It is unlikely that when one retrospectively looks at medical claims data, for example, one can be confident that one knows the true process by which the data were generated.[9]

There may be individuals in either the case or control cohort who simply had no comparable individual in the contrasting population. Extrapolating results to these individuals is dubious at best.

Matching can help alleviate these problems by not making parametric assumptions, excluding participants for whom there is no appropriate match and not making the data fit a particular functional form.

PROPENSITY SCORE MATCHING--CONTINUED

Although preferable to regression analysis, Propensity Score Matching has itself sparked a considerable literature critical of the plausibility of its assumptions.[10] In particular, what is known as the ignorability assumption has been questioned. This assumption states, loosely, that, conditional on all the variables which determined assignment to case or control the outcome is independent of any other variable. In many instances, such as a health care company trying to figure out why patient X used drug A instead of drug B, this seems like an unrealistic assumption. The distance between what can be known based on medical claims data and the actual patient/doctor decision is just too great to imagine that one has accurately modeled this decision with the variables at hand.

Another criticism, the one I would like to examine in detail in this study, is the fact that, since each observation gets its own particular propensity, given the continuous and categorical variables used in the logistic regression, there is no theoretically derived tolerance to use in matching individuals to one another. Even more, if one wants to use all the information possible in an N:N match, it is computationally impossible to calculate the difference in propensity score between all permutations once the number of observations is sufficiently large (>1,000).

COARSENED EXACT MATCHING

Coarsened Exact Matching resolves this issue. In CEM, instead of matching based on a composite propensity to be in either treatment or control in order to create a quasi-randomized sample, one simply coarsens each variable to a reasonable degree of clustering for each variable and then performs an exact match on these coarsened variables.

For example, if one has years of education it is probably going to be difficult to match all individuals precisely if one is also matching on gender, age, income, etc. Instead, one can coarsen this variable to less than high school, high school graduate, some college, college graduate, and post-graduate studies and have enough similarity between individuals for the purposes of a particular study.

By not computing a composite score for each observation, and then having to select a certain tolerance, one can easily compute how many cases and controls fell into each permutation and then weight the controls based on how many cases they matched to. For example, suppose one had a dataset with only nine observations in which one case matched to two controls and four different cases matched to two other controls. One would assign of a weight of 4/2*4/5=1.6 to the 2 controls which matched to four cases and 1/2*4/5=0.4 to the 2 controls which matched to one case. Each case would receive a weight of one. In this way, all the useful information in the case and control datasets can be included in the analysis, a significant advantage over PSM. When one encounters larger, and thus more realistic, numbers of cases and controls the same method of weighting every control by the number of cases it matched to and giving each case a weight of one is used.

Moreover, when the ratio of cases to controls falls somewhere below 1:3, propensity score matching, even with replacement, can become problematic as one cannot retain enough individuals in the post-matched sample to evaluate the ATE properly. And, as previously stated, N:N matching is often not computationally possible.

CASE STUDY OF N:N COARSENED EXACT MATCHING

I was given the task at Cigna of evaluating a very large treatment where there were only seven confounders that needed to be controlled for. What made this analysis difficult was the ratio of cases to controls: 525,788 treated to 483,339 controls. Furthermore, the customers had statistically and substantively different averages across these confounders making the p-value of all significance tests far below an alpha level of 0.05 (see table 1). A first reaction may be to respond by stating that the analysis cannot be done since there are more treated than controlled. Therefore, even if the populations were relatively balanced across the confounders (which they are not) it would not be feasible to perform a 1:1 or N:1 match simply because there are not enough controls to make a valid comparison.

Fortunately, CEM allows for N:N comparisons.

Prior to matching, I determined appropriate coarsening on all confounders. Zip codes were divided into five regions of the United States where one could expect to find relatively similar costs. Age was bucketed into six categories which correspond to different stages of development where similar health care costs are experienced. Customer Cost Share was divided into three buckets of low, medium and high. An overall measure of the overall health of customer, called ERG, was divided into 29 different categories after trial and error determined that this led to insignificant differences between case and controls. I then used a Coarsened Exact Matching Macro developed by Professor Gary King and several colleagues.[11]

Pre-Matching Statistics on Confounders

	Treatment	Control	P-Value
N	525,788	483,339	N/A
Male	51.64%	52.08%	<0.0001
Midwest	13.42%	13.75%	<0.0001
South	51.25%	54.65%	
Northeast	19.69%	19.79%	
West	15.53%	11.81%	
Other	0.11%	.01%	
17 or Younger	2.78%	2.34%	<0.0001
18 to 24	2.21%	2.19%	
25 to 34	5.22%	4.80%	
35 to 44	15.33%	13.89%	
45 to 54	32.25%	31.61%	
55 to 65	42.21%	45.16%	
0%-19% Cost Share	44.77%	42.58%	<0.0001
19%-58% Cost Share	39.97%	41.07%	
>=58% Cost Share	16.35%	15.26%	
HRA/HSA	29.99%	35.71%	<0.0001
PHS Plus	70.63%	79.67%	<0.0001
ERG	2.38	2.47	<0.0001
Specialty Med Utilizer	3.58%	3.45%	0.0003

Post-Matching Statistics on Confounders

	Treatment	Control	P-Value
N	525,788	483,339	N/A
Male	51.64%	51.64%	1
Midwest	13.42%	13.42%	1
South	51.25%	51.25%	
Northeast	19.69%	19.69%	
West	15.53%	15.53%	
Other	0.11%	.11%	

17 or Younger	**2.78%**	**2.78%**	**1**
18 to 24	**2.21%**	**2.21%**	
25 to 34	**5.22%**	**5.22%**	
35 to 44	**15.33%**	**15.33%**	
45 to 54	**32.25%**	**32.25%**	
55 to 65	**42.21%**	**42.21%**	
0%-19% Cost Share	**44.77%**	**44.77%**	**1**
19%-58% Cost Share	**39.97%**	**39.97%**	
>=58% Cost Share	**16.35%**	**16.35%**	
HRA/HSA	**29.99%**	**29.99%**	**1**
PHS Plus	**70.63%**	**70.63%**	**1**
ERG	**2.38**	**2.38**	**0.6530**
Specialty Med Utilizer	**3.58%**	**3.58%**	**1**

Table 1. Pre and Post Matching Differences in Treatment and Controls

Remarkably, given the differences in the pre-matching population, 100% of both the cases and controls could be matched via the N:N CEM macro. And, while statistical significance tests are not necessarily the best guide to whether confounders have been appropriately controlled for12, all post-matching differences between the confounding variables were statistically insignificant.

In other words, without losing a single customer, I had transformed over 1,000,000 customers, simply by using appropriate weights on each control, into a matched dataset on which I could evaluate the average treatment effect.

Without exaggeration, I would describe the transformation of this dataset as near magical. Not only is there no statistically significant difference, there is actually no difference at all on all the confounders except ERG which, because it is continuous, had to allow for some differences within each of the 29 bins. This is accomplished by using

100% of the information in the pre-matched data was used in the match, a result which is almost impossible to achieve in PSM. While I cannot guarantee similar results for all uses of CEM in retrospective analysis, the overwhelming success of the method in this instance should be strong support for its use by others encountering similar datasets where one needs to evaluate the treatment effect with a ratio of case to controls that approximates 1:1.

CONCLUSION

There have been a lot of previous work comparing both the theoretical and experimental properties of different kinds of causal inference. What I hope I have achieved in this modest contribution, is to point out that N:N Coarsened Exact Matching often makes possible causal inferences which cannot be made with methods, such as PSM, which don’t easily allow for N:N matching.

REFERENCES

[1] “Cigna” as used herein refers to operating subsidiaries of Cigna Corporation, including Cigna Health and Life Insurance Company and Cigna Health Management, Inc. All Cigna products and services are provided exclusively by such operating subsidiaries

[2] Donald B. Rubin and Paul R. Rosenbaum, The Central Role of the Propensity Score in Observational Studies for Causal Effects, Biometrika Vol 70, 1983 41-55.

https://academic.oup.com/biomet/article/70/1/41/240879/The-central-role-of-the-propensity-score-in

[3] See, for example, Judea Pearl, Madelyn Glymour and Nicholas P. Jewell, 2016. Causal Inference in Statistics: A Primer. New York, NY. John Wiley and Sons

[4] Examples are:

Guido W. Imbens and Donald W. Rubin, Causal Inference for Statistics, Social and Biomedical Science. An Introduction, 2015. New York, NY. Cambridge University Press.

Shenyang Gao and Mark W. Fraser, 2015. Propensity Score Analysis, Statistical Methods and Applications. Thousand Oaks, CA. Sage Publications.

[5] The dimensionality problem is that, as the number of confounders increases, it becomes increasingly difficult to find exact matches between case and controls.

[6] On this see Gary King and Richard Nielsen, Working Paper, Why Propensity Scores Should Not Be Used for Matching. http://j.mp/1sexgVw

[7] Idem

[8] See Guido W. Imbens and Donald B. Rubin,opus cited.

[9] Gary King and Richard Nielsen, opus cited.

[10] See Guido W. Imbens and Donald B. Rubin, opus cited pp 257-280.

[11] The Macro can be found at: http://gking.harvard.edu/cem

[12] See Guido W. Imbens and Donald B. Rubin, opus cited pp.349-358.

[13] Idem.

ACKNOWLEDGMENTS

The author would like to thank Mr. Anthony Sumner and Mr. Jigar Shah for providing me the opportunity to conduct this analysis and for helpful suggestions throughout.

RECOMMENDED READING

- Guido W. Imbens and Donald W. Rubin, *Causal Inference for Statistics, Social and Biomedical Science. An Introduction*
- Shenyang Gao and Mark W. Fraser, *Propensity Score Analysis, Statistical Methods and Applications.*
- *Papers on Coarsened Exact Matching may be found at http://gking.harvard.edu/cem*

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Aran Canes
Enterprise: Cigna Health and Life Insurance Company
Address: 701 Corporate Center Drive
City, State ZIP: Raleigh, NC 27607
Work Phone: 919-854-7261
E-mail:aran.canes@cigna.com