E C O N O M I C S B U L L E T I N

Propensity Score Matching with Limited Overlap

Onur Baser
Thomson-Medstat

Abstract

In this article, we have demostrated the application of two newly proposed estimators which accounts for lack of overlap under propensity score matching on a case study involing the analysis of health expenditure data for the United States.

Citation: Baser, Onur, (2007) "Propensity Score Matching with Limited Overlap." *Economics Bulletin,* Vol. 9, No. 8 pp. 1-8
Submitted: March 17, 2007. **Accepted:** April 9, 2007.
URL: http://economicsbulletin.vanderbilt.edu/2007/volume9/EB-07I10005A.pdf

Introduction

Propensity score matching is a technique for removing possible selection bias on observables, now widely used in health services research. Propensity score matching specifies a function of measuring the proximity of one case to another, based on many observed characteristics; cases are then grouped to minimize the distance between matched cases. The literature presents several matching techniques, which Baser (2006) compared using a real-world example.

To work well, the propensity score method requires sufficient "support" of the groups and a strong overlap of the distribution of the variables in treatment and control groups. Insufficient overlap , may result in imprecise estimates that are sensitive to the choice of specification. Below we present some empirical evidence on the importance of this source of noncomparability bias.

Heckman, Ichimura and Todd (1998) and Dehejia and Wahba (1999) point out the empirical relevance of the overlap issue. Several techniques have been proposed to deal with the problem. Cochran and Rubin (1973) suggest caliber matching, wherein potential matches are dropped if the within-match differences in propensity score exceed one-fourth of the standard deviation of the estimated propensity score. Rubin (1977) suggests simply discarding all units with covariate values with either no treated or no control units. Dehejia and Wahba (1999) focus on the average treatment effect for the treated group and suggest discarding all controls with estimated propensity score below the smallest value of the propensity score among the treated group. Heckman et al. (1988) drop units from the analysis if the estimated density of the covariate distribution conditional on treatment status is below some threshold.

All these methods have drawbacks, since they rely on arbitrary choices regarding thresholds for discarding observations. None offer a formal justification, and evidence that they improve the efficiency of the estimands and reduce the bias is limited.

Crump et al. (2006) recently proposed a method that provides a systematic approach to account for subpopulations with limited overlap in the covariates. Thus far this method has not been applied to health services data, where propensity score matching is common practice.. The objective of this study is to apply the proposed methodology for adjusting for the lack of overlap to the estimation of healthcare expenditures. In the next section we briefly describe this method and how we applied it to it to Medstat MarketScan® data from Thomson.

Study Design and Methods

Two specific methods proposed by Crump et al. (2006) can be summarized as follows:

1. The first method focuses on average treatment effects within a selected subpopulation, defined in terms of covariate values by balancing possible two opposing effects: (a) the increase in variance of the estimated average treatment effect due to smaller (subpopulation) sample size (b) the decrease in variance of the estimated average treatment effect due to discarding observations whose efficient comparable representative is missing. Crump et al. formulate the optimum value of *a* that balances the two opposing effects and show that for a subpopulation whose estimated propensity scores *(e(x))* is in between *[a,1-a],* it is possible to estimate average treatment effect more precisely than the average effect for the entire population. This estimator is referred to as optimal subpopulation average treatment effect (OSATE).
2. The second method, called optimally weighted average treatment effect (OWATE), considers weighted average treatment effects with the weights depending only on the covariates. The optimal weight function is a function of the propensity score alone, proportional to the product of the propensity score and one minus the propensity score. Under homoskedasticity, the weight is simply *e(x) * (1-e(x))*. Formulas are presented at Crump et al. (2006).

Since the method inherently lowers the weight on high-variance observations and increase the weight on the observations with propensity score close to one-half, sub samples based on these estimators tend to be more balanced in the distribution of covariates.

Increase in precision of the estimates is another advantage of the proposed method. By discarding the observations for which average treatment effect cannot be estimated efficiently, the methods increases the internal validity at the expense of external validity. For most cases in health research, the former is more important.

More relevantly, in pharmacoeconomics, the primary interest might be to estimate the treatment effect of some group of patients in a broader population. Usually it is more difficult to find comparable match for sicker patients, so most often a more precise estimator is sacrificed by including these patients and their "not well matched" controls. The proposed two estimators, although based on subpopulations, allow us to make more precise inferences, rather than reporting potentially biased estimate for population average effect.

Finally, since true randomization is not possible in observational studies, any evidence that supports the reliability of our population average treatment estimator using propensity score matching is valuable information. In this respect, these estimators are useful because, if the variance reduction suggested by OSATE or OWATE estimators is not significant relative to the variance of average treatment effect, we can conclude that our population average effect is reliable.

Data Source

This retrospective claims analysis used data from the Medstat MarketScan Commercial Claims and Encounters (Commercial) Database for 1998–2004. These data include health insurance claims across the continuum of care (e.g. inpatient, outpatient, outpatient pharmacy, carve-out behavioral healthcare), plus enrollment data from large employers and health plans across the United States who provide private healthcare coverage for more than 33 million employees, their spouses, and dependents. This administrative claims database includes a variety of fee-for-service, preferred provider organizations, and capitated health plans.

The study population consisted of subjects with an International Classification of Disease (9th revision, or ICD-9) primary diagnosis of prostate cancer (185.0 236.5) in the inpatient, outpatient, or emergency department setting. For outpatient claims, we required that ICD-9 codes appear on two or more claims at least 30 days apart, in order to exclude patients with rule-out diagnoses only. We required that subjects were continuously enrolled for 12 months before the index date and 36 months after the index date, and had prescription drug benefits for the entire study period. The analytic sample consisted of 8,576 prostate cancer patients and 30,550 cancer-free patients. We calculated a baseline Charlson comorbidity index (CCI) score for each group and also used age, health plans (indemnity, POS, PPO, capitated POS), and geographic region for the estimation of propensity score.

The total cost of healthcare was measured as total medical costs for all inpatient, outpatient, pharmaceuticals, radiology, and emergency room (ER) visits in the three-year follow-up period. Costs incurred in 1999–2003 were adjusted to 2004 dollars based on the Consumer Price Index–Medical Component.

Results

This study was undertaken to answer several questions: (a) Is it possible to estimate burden of illness for prostate patients with covariate adjustment for our population?(b) If not, can we identify an optimal subpopulation that allows us to estimate the burden of illness? (c) How much does the precision of our estimates change if we shift population estimate to subpopulation estimate?

Table 1 presents summary statistics. It is evident that both cancer and non-cancer groups differ dramatically from the treatment group in terms of pre-period CCI, age, region, and most of the health plans; all of the mean are significantly different from zero, well beyond a 1% level of significance except the indicator “POS” and “North Central.” For the prostate cancer group, the pre-period CCI is 2.8075. For the cancer-free group, it is only 0.9197. Given the standard deviation of 1.7332, this sample suggests that simple covariance adjustment is unlikely to yield credible inferences.

Table 1. Covariate Balance for the Cancer Dataset

Variables	Overall Difference		Prostate Cancer	Cancer Free	Normalized Differences in Cancer and Non-Cancer Group Averages				
	Mean	STD	Mean	Mean	All	p-value	a<e(x)<1-a	Optimal Weight	Weighted Prop. Score
Pre-period CCI	1.3335	1.7332	2.8075	0.9197	1.0892	0.0000	0.7452	0.6585	0.9850
Age	73.1001	6.0942	74.8075	72.6208	0.3588	0.0000	0.1242	0.1458	0.2458
North central	0.3145	0.4643	0.3288	0.3105	0.0393	0.0013	0.0158	0.0246	0.3211
South	0.3600	0.4800	0.4005	0.3486	0.1081	0.0000	0.0587	0.0658	0.0959
West	0.1358	0.3426	0.0752	0.1529	-0.2267	0.0000	-0.1136	-0.1547	-0.1854
Indemnity	0.6568	0.4748	0.6884	0.6479	0.0852	0.0000	0.0125	-0.0012	0.0654
POS	0.0093	0.0959	0.0089	0.0094	-0.0056	0.6488	-0.0035	-0.0065	-0.0049
PPO	0.2790	0.4485	0.1785	0.3073	-0.2871	0.0000	-0.0570	-0.0576	-0.2032
POS capitated	0.0538	0.2256	0.1241	0.0341	0.3988	0.0000	0.1478	0.2458	0.3425

Normalized differences in cancer and non-cancer group averages are shown in the table. Even with sample size and ratio of cancer to non-cancer group fixed, the cumulative probability distribution function of statistics such as a t test (or corresponding p-values) can be uninformative if the variances of two samples are vastly different.

For this dataset, we estimated the propensity score using a logistic model with all nine covariates entering linearly. We then used the estimated propensity score to calculate the optimal cut-off point, *a*. The optimal cutoff point is a = 0.0211.

According to this criterion, 34,463 observations should be discarded. Out of the original 8,576 cancer and 30,550 non-cancer patients, only 1,752 patients from the cancer group and 2,912 patients from the non-cancer group were left. In Table 2, we present the number of observations in the various categories. Here OSATE methodology suggests dropping 32,680 out of 34,463 observations, leaving 1,783 observations, or just 5% of the original sample. This suggests that the covariate values for some non-cancer patients are so far from those of the cancer patients that attempting to estimate burden of illness for these covariate values would be unrewarding.

Table 2. Subsample Sizes for Analytic File, Propensity Score Threshold 0.0211

	E(x)<a	**a<e(x)<1-a**	**1-a<e(x)**	**All**
Cancer-Free Group	2,102	658	152	2,912
Prostate Cancer Group	245	1,125	382	1,752
All	2,347	1,783	534	4,664

Table 3 presents asymptotic standard errors for the difference in total healthcare expenditures between cancer patients and non-cancer patients. The first one is the standard error for the population average treatment effect (ATE). The second is the asymptotic standard error for the average treatment effect (OSATE) in the subpopulation with a<e(x)<1-a, for the optimal value of a=0.02111. The third is the standard error for the optimally weighted average treatment effect (OWATE). The last one is the asymptotic standard error for the average treatment effect for the treated (ATT).

Table 3. Asymptotic Standard Errors

	ATE	**OSATE**	**OWATE**	**ATT**
Asymptotic Standard Error	434.9984	2.0140	2.6521	5.2142
Ratio to All	1.0000	0.0046	0.0061	0.0120

Moving from the population average treatment effect to any of the three other estimands resulted in huge gains. Calculations suggest that OSATE lowers the variance by a factor of 0.0046, reflecting the sizeable difference between most of the controls and the treated patients, as well as the difficulty of estimating the population burden of illness. Thus large areas in the covariate space show essentially no treated units (prostate patients).

Conclusion

In practice, an important concern in implementing propensity score matching is the necessity of sufficient overlap between covariate distributions in the treatment and control groups, since limited overlap can result in estimators for average treatment effects with poor finite sample properties. In particular, such estimators can have substantial bias, large variances, and considerable sensitivity to the exact specification of the propensity score. In this case, optimal subpopulation can lead to precise estimators, which can be presented with the population average treatment effect.

In this article, we have demonstrated the application of two newly proposed estimators on a case study involving the analysis of health expenditure data for the United States. Lack of overlap is especially important in health expenditure data; given the significance level of difference in disease staging between treated and control groups. Our results show that, due to significant differences in covariates between the prostate cancer patient and non-cancer groups, we cannot estimate the burden of illness for our population precisely even after covariate adjustments between the two groups; the gap was simply too large to support reliable conclusions. However, we did identify an optimal subpopulation that would produce the efficient and precise estimates.

The methods suggested by Crump et al. (2006) and applied here are not relevant in all situations. First, these methods change the estimands. The estimators focus on the average effects for a subpopulation or a weighted subpopulation, so generalization to the larger population would not be correct. Instead of reporting solely the potentially imprecise estimate for the population average treatment effect, it has been proposed that we report estimates both for the population of interest and for the subpopulation, where one can make more precise inferences. Second, there may be important unobservable covariates for which these adjustment using observable covariates cannot account. Several techniques are available to control for unobservable factors, such as the Instrumental Variable method (Wooldridge, 2002) or Rosenbaum's bounding (Rosenbaum, 2002) approach, but these estimations are confounded by their own limitation.

This article does not provide detailed or rigorous consideration of the method that underlines the application. Curious readers are encouraged to consult the method article by Crump et al. (2006).

References

1. Baser O. Too much ado about propensity score models? Comparing methods of propensity score matching. *Value in Health*, 9(6): 377-385, 2006.
2. Heckman J., Ichimura H., Todd P. Matching as an econometric evaluation estimator. *Review of Economic Studies* 1998; 65: 261–294.
3. Dehejia R, Wahba S. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of American Statistical Association* 1999; 94(448): 1053–1062.
4. Rubin D. Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics* 1977; 2(1): 1–26.
5. Cochran WG, Rubin DB. Controlling bias in observational studies: A review. *Sankya, Series A* 1973; 35: 417–446.
6. Crump RK, Hotz JV, Imbens GW, Mitnik O. Moving the GoalPosts: Addressing Limited Overlap in the Estimation of Average Treatment Effects by Changing the Estimand, National Bureau of Economic Research, Technical Working Paper 330, September 2006.
7. Rosenbaum PR. Observational Studies. Series in Statistics. New York: Springer, 2002.
8. Wooldridge JM. Econometric analysis of cross section and panel data. Cambridge, Mass.: MIT Press; 2002.