Propensity Score Analysis of Data From Case-Control Studies

Irina Bondarenko * Trivellore Raghunathan†

Abstract

Propensity Score (PS) methods are useful to adjust for many covariates and to account for lack of randomization in many prospective or cohort observational studies. The goal of this paper is to develop and evaluate propensity score method for retrospective or case-control study designs which are also used in public health research. The primary purpose is to adjust for a large number of covariates in estimating the odds ratio, a common measure used in retrospective studies, and assessing its statistical significance. We use counterfactual probability of exposure under the control conditions to group subjects and then perform a stratified analysis. We conducted a simulation study to evaluate the repeated sampling properties of the associated point and interval estimates of the odds ratio.

Key Words: Cohort Studies, Causal Inference, Stratification, Confounding, Retrospective Studies, Propensity Score

1. Introduction

Propensity score approach is a powerful and popular method of analyzing data from observational studies. This framework allows adjustment for many covariates and to account for lack of randomization of exposure or treatment variables in cohort or cross-sectional studies. Specifically, if E is a binary exposure indicator variable, Y is an outcome, and X is a collection of covariates then propensity score (PS) is defined as conditional probability of exposure given the covariates.Throughout this paper $e(X) = Pr(E = 1|X)$ denotes the propensity score. Rubin and Rosenbaum have shown that propensity score $e(X)$ is an efficient summary of the covariates X and could be used to compare the outcome Y across the two treatment groups $E = 1$ and $E = 0$ conditional on the $e(X)$. (Rosenbaum & Rubin (1983)).

This framework provides an umbrella for various methods based on stratification, matching and weighting adjustments. Essentially, stratification and matching methods are aimed to achieve an effect of pseudo-randomization in observational study settings. The weighting adjustment compares pseudo populations under the two treatment scenarios.

Implicitly, propensity score methods assume that neither the study design nor the outcome variables affect the exposure probabilities. This assumption is satisfied in self-representing cohort or cross-sectional study designs but is violated in case-control study settings (and more broadly in the complex survey settings). The case-control study involves outcome based sampling where cases ($Y = 1$) are recorded first and then controls ($Y = 0$) are sampled from the study population to match certain distributional (frequency matching) or actual characteristics (matching on actual values of the variables) of cases. The exposure is then acertained on the the cases and controls. The ratio of cases to controls and the sampling mechanism is an important part of the design. Thus, case-control data consists of two different samples with a fixed ratio of cases to controls. In these settings probability of exposure is affected by the ratio of cases to controls, and the joint distribution of $f(E, X)$ is distorted. Hence, the case-control studies does not fit into the usual

*Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48108

†Survey Research Center, Institute for Social Research., 426 Thompson Street, Ann Arbor, Michigan 48106

causal inference framework for randomized or prospective studies. Here, therefore, we take the view of adjusting for a large number of covariates while inferring about the odds ratio, a common parameter used to assess the strength of association between exposure and outcome.

Propensity score methods for the case-control study settings were addressed in a number of papers. Robins (1999) suggested the propensity score estimated from the sampled non-cases within the marginal structural model framework. Mannson at el (2007) conducted a simulation study to compare different PS methods in the case-control study settings. Allen and Satten (2011) proposed to compare distributions of exposures that would be found among case and control participants if both groups had the same distribution of confounding, using conditional probability of being a case.

In this paper we describe an alternative propensity score based method which uses the counterfactual or estimated probability of being exposed had the case been a control and then match or group with propensity of exposure for controls. Additional stratification based on the propensity of being a case further eliminates the effect of confounding.

1.1 Proposed methodology

Let $e_o(X) = Pr(E = 1|X, Y = 0)$ be the propensity score model for the controls. A logistic or probit regression model is posited, the maximum likelihood approach is used to obtain the estimated vector of regression coefficient, $\widehat{\beta}_o$ to construct the estimated propensity scores $\widehat{e}_o(X)$. The design features such as stratification or weighting is incorporated in the estimation process.

We apply the regression coefficient estimate $\widehat{\beta}_o$ to the covariates of the cases to obtain the counterfactual propensity scores $\widehat{e}^*_o(X)$. The propensity scores $\widehat{e}_o(X)$ and $\widehat{e}^*_o(X)$ are appended and then the subjects are stratified based on the score. This approach only partially accounts for the imbalances because the distribution of X for the cases is not explicitly used in the construction of the propensity score. In the presence of confounding, when X affects Y and E, the distribution of X is different not only by exposure E, but by the outcome Y. We propose to additionally match or group subjects on the conditional probability of being a case given the covariates after grouping or matching on the counterfactual propensity of exposure.

Epstein (2007) discussed the use of the stratification score defined as $s(x)$ as $Pr(Y = 1|X)$ in retrospective studies and had shown that correctly specified stratification score is a retrospective balancing score for a case-control study, meaning that $Pr(X|Y, s(x)) = Pr(X|s(x))$. Thus, it can be used to balance the distribution of confounders X for cases and controls within each stratum based on the $e_0(x)$ values.

In summary, the proposed procedure involves the following steps:

1. Estimate counterfactual propensity of exposure as defined earlier.
2. Stratify subjects into strata based on the score (say p strata)
3. Within each stratum estimate the propensity of being a case, $s(x)$, given the covariates X.
4. Create further stratification (say, r strata) based on $s(x)$ within each of the p strata.
5. Use Cochran-Mantel-Haenszel test assess the statistical significance of common odds ratio across $p \times r$ strata to assess the effect of exposure on the outcome.

1.2 Simulation Study

We conducted a simulation study to evaluate the properties of the proposed method. The simulation study was a factorial design (2x2x4) with the two levels of the marginal probability of the exposure, two levels of the strength of association between covariates and the exposure and four levels of the strength of association between covariates and the outcome. For each of theses scenarios we simulated a cohort population, then drew a case-control sample and applied our method.

To simulate the cohort population we used a design similar to that of Mansson (2007). For each scenario, we simulated 1000 replicates of cohort populations with size 20000. In each population, the covariates were 5 independent standard normal deviates, x_1, x_2, x_3, x_4, x_5. Exposure E was determined using the logistic model,

$$\begin{aligned}\text{logit}(\Pr(E=1)) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \\ \alpha_3 x_3 + \alpha_4 x_4 + \alpha_5 x_5 \, .\end{aligned} \tag{1}$$

The outcome $Y = 1$ was generated using the logistic model,

$$\begin{aligned}\text{logit}(\Pr(Y=1)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \\ \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \log(OR)E \, .\end{aligned} \tag{2}$$

Parameter OR had two levels: $OR = 1$ corresponding to no effect of exposure E on the outcome Y, and $OR = 5$ - a strong effect of E on the outcome.

The intercept α_0 was chosen to ensure that the marginal probability of exposure in the cohort is equal to 0.15 or 0.5. The intercept β_0 was selected to fix the the probability of outcome as 0.01 (rare outcome).

The parameter vector α defined the strength of association between the covariates x_1, x_2, x_3, x_4, x_5 and the exposure, E. The two levels were $\alpha^{mod} = (0.3, 0.3, 0.3, 0.3, 0, 3)^T$ (moderate association), and $\alpha^{str} = (0.6, 0.6, 0.6, 0.6, 0, 6)^T$ (strong association).

The parameter vector β defined the strength of association between the covariates x_1, x_2, x_3, x_4, x_5 and the outcome, Y. The four levels of β were

$\beta^{str} = (0.6, 0.6, 0.6, 0.6, 0.6)^T$ (strong association),

$\beta^{mod+} = (0.3, 0.3, 0.3, 0.3, 0.3)^T$ (medium strength association and similar to association between exposure and covariates),

$\beta^{mod-} = (-0.3, -0.3, -0.3, -0.3, -0.3)^T$ (medium strength association opposite to the direction of association between covariates and E),

$\beta^{no} = (0, 0, 0, 0, 0)^T$ (no confounding).

Next, from each cohort population, we drew a case-control population under 2x2 factorial design. Specifically, we used two fixed ratios of cases to controls: 1 or 2, and two sampling schemes for controls: simple random and stratified random sampling. Under each setting, we first selected all cases with $Y = 1$ from the underlying cohort population, and sampled a fixed number of controls based on pre-fixed ratio of cases to controls and the sampling mechanism. For stratified sampling, values of the covariates x_1, x_2, x_3 were dichotomized and controls were sampled to match the distribution of the cases in each stratum. The mean sample size of the resulting case-control sample was 400 when the sampling ratio was 1 and 600 when the sampling ratio was 2. We also fixed the number of strata for the two propensity scores at $p = r = 5$ (that is, a total of 25 strata).

The proposed method is labeled $s(e_o(x))$. For comparison we included results from two additional methods: $(e(x))$ stratification into 5 strata based on the propensity score directly estimated on the entire case-control sample; $(e_0(x))$ stratification into 5 strata based on the counterfactual propensity scores.

Table 1 shows percent bias and under coverage rates for all three methods when case-control sample was selected using sample random sample and the sampling ratio of 1. Overall, the proposed method $s(e_0(x))$ shows the lowest percent bias and is closest to the nominal coverage rates. In contrast, the $e(x)$ method results in substantial underestimation (negative percent bias) of the true OR when the association between covariates and E is concordant with that of Y. When the association was discordant then stratification on $e(x)$ resulted in overestimation of the true OR. The stronger the association between covariates and the exposure or outcome, the more bias introduced by estimation based on $e(x)$. Undercoverage reached 15% for some scenarios.

Stratification on $e_0(x)$, the counter factual propensity score, reduced bias and improved coverage. However, for the scenarios with strong confounding,($\beta = \beta^{str}$), a substantial residual bias remained after stratification. On the other hand, the proposed method reduces bias in all but one scenarios below 3% and boosts coverage to above 94%.

Table 2 shows percent bias and under coverage rates for all three methods for stratified random sampling and fixed ratio of cases to controls equal to 1 . The results of application of the proposed method to the case-control sample selected using stratified random sample were similar to the results shown in Table 1.

1.3 Conclusion

Propensity score based methods have several advantages over the regression models. With a large number of covariates and potential for collinearity, regression models can give unreliable results. Stratification or matching on propensity scores allow to effectively reduce dimensionality of covariate space. This strategy has a special importance for the case-control studies given relatively low number of participants in such studies and a large number of covariates. Since the propensity score methods requires evaluation of the overlap between scores between cases and controls, they provide insights into collinearity among the covariates. Also, propensity score methods are essentially semi-parametric and allow additional flexibility when effect of the covariates on the outcome is non-linear.

In this paper we proposed the propensity score based method for case-control studies and examined its sampling properties for a number of simple scenarios. The proposed method is shown to have low bias and good coverage rate when correct models are used for stratification and propensity scores. Robustness of the method when one or both scores are misspecified needs to be evaluated. We evaluated use of the proposed method for the stratification and need and need to extended it for other complex sampling designs.

REFERENCES

Allen A.S., Satten G.A. Am J Epidemiol. (2011), "Control for confounding in case-control studies using the stratification score, a retrospective balancing score," *Am. J. Epidemiol.*, 173(7), pp.:752-760.

Epstein M.P., Allen A.S., Satten G.A. (2007). "A simple and improved correction for population stratification in case-control studies, *Am J Hum Genet.*, 80(5), pp.921930.

Mansson R., Joffe M.M., Sun W., Henessy S. (2007), "On the Estimation and Use of Propensity Scores in Case-Control and Case-Cohort Studies," *Am. J. Epidemiol.*,166, pp.332-339.

Robins J.M. (1999), "Comment on Choice as an alternative to control in observational studies, *Stat.Sci.*, 14, pp.281-93.

Rosenbaum, P. R., Rubin, D. B. (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects, " *Biometrika*, 70, pp. 41-55.

Table 1: Simulation Results for Simple Random Sampling Design.

				Bias(%)			Undercoverage (%)		
True OR	$P(E=1)$	α	β	$e(x)$	$e_0(x)$	$s(e_0(x))$	$e(x)$	$e_0(x)$	$s(e_0(x))$
1	0.15	0.3	0	-0.3	-0.2	-0.5	5.7	5.5	5.1
			0.3	-10.4	-4.3	-0.2	7.1	4.3	4.2
			-0.3	9.9	5.2	0.8	6.7	5.4	4.4
			0.6	-23.1	-8.3	-1.9	17.7	5.8	6.0
		0.6	0.0	0.2	0.0	0.2	5.2	5.3	5.0
			0.3	-12.5	-8.0	-0.8	7.7	4.0	4.3
			-0.3	13.9	10.9	1.5	5.9	5.0	4.3
			0.6	-23.3	-11.1	-0.1	14.0	7.2	4.3
	0.5	0.3	0.0	0.2	0.1	0.5	4.7	4.7	4.6
			0.3	-7.5	-3.6	-0.2	6.7	5.7	6.5
			-0.3	7.0	3.6	0.5	7.9	6.2	5.7
			0.6	-13.6	-5.5	-0.7	9.1	3.9	4.5
		0.6	0.0	-1.1	-1.1	-1.4	4.5	4.7	4.9
			0.3	-7.0	-4.5	0.7	5.7	5.3	4.7
			-0.3	7.5	5.4	0.3	6.2	5.1	3.8
			0.6	-15.4	-10.5	0.1	8.8	6.3	5.6
5	0.15	0.3	0.0	-2.4	-0.4	-0.4	5.1	4.6	4.5
			0.3	-12.6	-2.4	-0.8	8.2	4.8	5.9
			-0.3	6.6	5.2	-1.9	7.7	6.4	5.0
			0.6	-20.7	-3.2	-2.1	14.9	5.1	4.5
		0.6	0.0	-4.5	-1.8	-1.4	5.6	5.1	5.2
			0.3	-14.7	-5.2	-0.6	9.3	5.3	5.4
			-0.3	6.9	5.1	-2.8	6.7	5.5	3.9
			0.6	-25.7	-10.0	-2.6	12.3	4.5	3.5
	0.5	0.3	0.0	-0.2	0.5	0.5	4.1	3.3	3.6
			0.3	-9.8	-4.2	-1.8	5.1	4.7	4.9
			-0.3	6.3	6.7	0.5	6.6	5.8	4.9
			0.6	-15.4	-4.5	-2.6	10.1	5.7	4.8
		0.6	0.0	-3.9	-2.3	-0.9	5.7	5.9	4.8
			0.3	-12.5	-8.6	-2.0	6.6	5.3	4.7
			-0.3	6.1	5.6	-0.6	6.5	5.6	3.8
			0.6	-20.5	-14.8	-5.2	5.5	4.3	4.2

Table 2: Simulation Results for Stratified Random Sampling Design.

				Bias(%)			Undercoverage (%)		
True OR	$P(E=1)$	α	β	$e(x)$	$e_0(x)$	$s(e_0(x))$	$e(x)$	$e_0(x)$	$s(e_0(x))$
1	0.15	0.3	0	-0.7	-0.8	-0.7	4.8	4.4	4.3
			0.3	-8.7	-4.5	-1.1	8.2	5.5	5.2
			-0.3	9.2	4.8	1.8	5.3	4.1	4.0
			0.6	-13.9	-5.6	-0.1	11.1	4.6	3.9
		0.6	0	-0.4	-0.5	-0.4	5.1	5.2	4.1
			0.3	-6.0	-5.7	-0.4	11.0	9.4	4.8
			-0.3	6.8	6.5	0.1	9.3	8.9	4.0
			0.6	-9.4	-8.7	-1.1	18.5	15.0	5.1
	0.5	0.3	0	1.2	1.1	1.2	5.7	5.5	5.8
			0.3	-5.8	-3.0	-0.8	5.8	4.9	3.9
			-0.3	6.2	3.4	0.5	7.3	5.8	5.0
			0.6	-9.7	-4.3	0.2	8.5	4.4	5.2
		0.6	0	0.6	0.3	0.6	4.9	4.4	5.1
			0.3	-6.0	-4.4	-0.6	6.0	4.1	4.5
			-0.3	6.0	4.2	0.2	6.8	5.1	5.4
			0.6	-12.8	-9.4	0.0	9.2	5.2	5.0
5	0.15	0.3	0	-2.8	-1.8	-2.2	4.7	4.8	4.5
			0.3	-9.6	-1.3	-1.9	7.0	4.7	5.0
			-0.3	5.1	3.8	-2.4	7.3	7.1	5.8
			0.6	-15.0	-0.1	-2.1	11.5	3.0	5.0
		0.6	0	0.3	1.0	2.0	5.3	5.5	5.7
			0.3	-4.3	-0.5	1.5	8.1	5.0	4.9
			-0.3	6.9	7.3	4.1	12	13.5	8.7
			0.6	-6.3	-0.5	1.5	12.1	3.8	4.7
	0.5	0.3	0	-1.4	-0.9	-1.7	4.4	4.3	5.1
			0.3	-9.2	-3.6	-2.9	6.4	5.3	5.2
			-0.3	4.2	4.4	-0.5	7.2	6.5	5.1
			0.6	-11.4	-0.4	-0.3	6.0	4.4	5.4
		0.6	0	-3.0	-1.7	-1.6	6.3	6.5	5.3
			0.3	-9.3	-4.5	-1.5	4.6	4.2	4.3
			-0.3	3.9	4.1	-1.2	5.7	4.9	4.4
			0.6	-16.7	-7.4	-2.7	6.0	3.8	3.4