Simulation study on variable selection in propensity score models

Faith Lee

Nov 28 2016 (Updated Nov 17 2017)

1 Aim and problem definition

- Generate 1000 datasets, each with 2500 subjects
- Three standard normal covariates, ie. $x_{ij} \sim N(0,1)$
- True probability of being assigned to treatment (ie. $A_i = 1$) for the i-th individual given by **Note this is the probability of getting your treatment, T = 1 (T is binary ; T = 0 or 1.** :

$$P(A_i = 1|X_i = x_i) = \frac{exp(0.5x_{i1} + 0.75x_{i3})}{1 + exp(0.5x_{i1} + 0.75x_{i3})} \tag{1}$$

- The outcome Y_i, which is a count outcome, follows a Poisson distribution ($Y_i \sim Poi(\mu_i)$ and is related to the treatment and covariates in the following manner:

$$\mu_i = E(Y_i|X_i, T_i) = exp(0.5 + \frac{4}{1 + exp(-3x_{i1})} + x_{i2} + \lambda_0 A_i) \tag{2}$$

  where $\lambda_0 = 0.5$

- Fit the following logistic regression propensity score models:
  - $PS_1 = P[A_i = 1|X_1]$
  - $PS_1 = P[A_i = 1|X_1, X_3]$
  - $PS_1 = P[A_i = 1|X_1, X_2]$
  - $PS_1 = P[A_i = 1|X_1, X_2, X_3]$
- Then based on the estimated propensity scores from the models above, fit the outcome Poisson model with $E(Y|T) = exp(\alpha_0 + \gamma T)$ using the following
  - Matching

- Inverse probability weighting (IPW)
- Stratification : Here we define **5** stratums
- Subsequently compute the causal log ratio ($\log \hat{\lambda}$), the bias, the mean square error (MSE) and standard deviation (SD)

2 Introduction

In this simulation study, we are interested to investigate the importance of variable selection in propensity score model to estimate causal effects. Common methods to remove the effects of confounders when estimating the effect of treatment on outcome include propensity score matching, stratification on the propensity score and inverse probability weighting (IPW) on the propensity score. Through a Monte Carlo simulation study, we aim to compare the bias, mean squared error (MSE) and standard deviation (SD) of the treatment effect estimate $\hat{\gamma}$ under the different propensity score model specification.

3 Methods

We generate Y, which is a Poisson distributed count outcome, A, which is the binary treatment assignment indicator and 3 covariates X_1, X_2, X_3 which are independent and have standard normal distribution. From the data generation steps in the question, we see X_1 is the real confounder as it is related to the A and Y, X_2 is related Y only and X_3 is related to A only. We have learnt that X_1 should be included in the propensity score model. However, including X_2 can reduce the variance of the causal estimator. Inclusion of X_3 in the model will increase the variance as it is instrumental and not related to the outcome. The relationship can be represented through a causal diagram shown in Fig. 1.

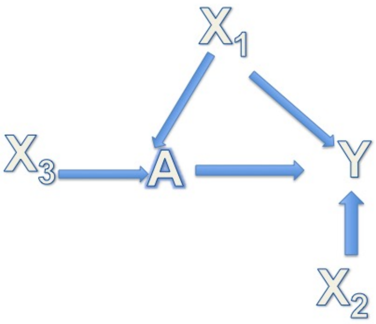

Figure 1: Causal diagram relating X_1, X_2, X_3, A and Y.

We are to investigate the four propensity score models through the matching, IPW and stratification :

- $PS_1 : P[A = 1|X_1]$(the confounders only model)
- $PS_2 : P[A = 1|X_1; X_3]$(the true propensity score model)
- $PS_3 : P[A = 1|X_1; X_2]$(the outcome associated model)
- $PS_4 : P[A = 1|X_1; X_2; X_3]$(the full model)

and subsequently compare their causal log rate ratio $\hat{\gamma}$ with the true parameter value, $\gamma_0 = 0.5$ through the bias, MSEs and SDs. The simulation is performed for B = 1000 replications with n = 2500 observations in each set of data.

4 Results and Discussion

4.1 Propensity Score Matching

The `R` package Matching was used to perform propensity score matching. The default one-to-one matching is used here.

Model	Performance Metric		
	Bias	SD	MSE
PS_1 (X_1)	0.00558	0.0847	0.00719
PS_2 (X_1, X_3)	0.00477	0.0957	0.00918
PS_3 (X_1, X_2)	0.00299	0.0711	0.00506
PS_4 (X_1, X_2, X_3)	0.00393	0.0881	0.00778

Table 1: Results from propensity score matching

All of our models included X_1 and it is necessary to do so as it is the only confounder. From the results, the model (PS_3) with X_1, X_2 has the lowest bias, MSE and SD of $\hat{\gamma}$. It shows that a model with X_1 only is not sufficient if there are other covariates associated with the outcome. Adding X_2 to the model already having X_1 improved the precision of the estimator. However, including X_3 to the model with X_1 and X_2 already in it (PS_4) increases the variance (by about 5.7%) and bias of the estimated effect. Adding variables that are related to the treatment but not to the outcome is not recommended as it can increase variation to the estimated treatment effect. From Table 1, the optimal model to estimate the causal treatment effect γ is PS_3.

4.2 Inverse Probability Weighting

The results showed that once again model PS_3 has the lowest bias, SD and MSE compared to all other models. From the results, adding X_3 to the model already having X_1 and X_2 increases the bias, SD and MSE of the estimated effect.

Model	Performance Metric		
	Bias	SD	MSE
PS_1 (X_1)	0.00597	0.0803	0.00648
PS_2 (X_1, X_3)	0.00593	0.0893	0.00802
PS_3 (X_1, X_2)	0.00466	0.0694	0.00483
PS_4 (X_1, X_2, X_3)	0.00513	0.0794	0.00632

Table 2: Results from using IPW on the propensity scores

4.3 Stratification

For stratification, for each dataset, the n = 2500 observations are divided into k = 5 strata on the basis of the quantiles of the estimated propensity scores from each of the four propensity score models. Similarly, the results showed that the model containing X_1, X_2 has the lowest bias, SD and MSE of the treatment effect estimate, γ compared to the other models. Adding X_3 to the model with X_1 and X_2 increased the bias and variance of the treatment effect, in fact it increased the variance by about 71.3% and the MSE by about 2.53 times. Here, we also see that adding X_3 is not recommended. However, including X_1 only in the model was also not sufficient as X_2 affects the outcome. Hence, Table 2 also shows that model PS_3 is the optimal model to estimate the causal treatment effect.

Model	Performance Metric		
	Bias	SD	MSE
PS_1 (X_1)	0.0211	0.0585	0.00387
PS_2 (X_1, X_3)	0.0421	0.0921	0.0102
PS_3 (X_1, X_2)	0.0202	0.0438	0.00232
PS_4 (X_1, X_2, X_3)	0.0409	0.0808	0.00820

Table 3: Results from stratification on the propensity scores

4.4 Discussion

The simulation study gave us consistent results by showing that PS_3 was the optimal model in all of the three methods used. Tables 1, 2 and 3 also showed that the addition of the X_3 which is the exposure-associated covariate to the model with X_1 and X_2 neither decreases the variance nor the bias of the estimator. Including all the covariates does not necessarily improve the performance of the model. We should therefore include covariates that are associated with the outcome, but it may not be recommended to include covariates that are associated with the exposure/treatment status only.

Comparing the methods, it is noticed that the stratification method resulted in larger bias of the estimated effect compared with IPW and propensity score matching. It also has the largest percentage change in bias, variance and MSE when we add X_3 to the PS_3 model.