Brief propensity scoring tutorial based on learning reported by SOWO 922 students

Alan R. Ellis, PhD, MSW
October 5, 2014

The propensity score (PS) is the probability of being assigned to a particular treatment, conditional on measured covariates (Rosenbaum & Rubin, 1983). When treatment groups are balanced on the PS, they are also balanced on observed covariates. Therefore, propensity scores are often used in observational studies to balance treatment groups on observed covariates in order to make valid comparisons between treatments, or between treatment and no treatment. Below are several comparisons between PSs and conventional regression modeling:

- The PS is typically used for comparative effectiveness studies (causal modeling); traditional models are also used for other purposes such as prediction, explanation, and correlational studies.
- When conducted properly, PS analysis and regression analysis often result in similar estimates (Stürmer et al., 2006).
- The PS is a summary score, which makes it especially useful in models of rare outcomes (<8 events per covariate) where it would be impossible or inadvisable to include many covariates in the outcome model (Cepeda et al., 2003).
- Although the positivity assumption (see below) applies to both regression and propensity scoring, the PS makes it easier to identify positivity violations because it's easy to check for overlap between groups on a single variable.
- Similarly, the PS makes it easy to detect heterogeneity in the treatment effect; this can be accomplished by stratifying on the PS.
- Traditional regression analysis doesn't have a "balance check" to assess the adequacy of control for measured covariates, but PS analysis does.
- PS analysis is more complicated and therefore a bit more difficult than traditional regression analysis.

Table 1 details the assumptions underlying PS analysis. Rosenbaum and Rubin (1983) described the positivity and exchangeability assumptions as part of strongly ignorable treatment assignment (SITA). They also described the stable unit-treatment value assumption (SUTVA). Cole and Hernán (2008) labeled positivity and exchangeability as such, labeled SUTVA as consistency, and added the assumption of correct model specification (which is inherent in Rosenbaum and Rubin's descriptions of PS analysis).

Table 1. Propensity scoring assumptions

Assumption	Violation	How to address potential violations
Positivity – The PS distributions of the treatment groups overlap completely (i.e., regardless of the actual treatment received, everyone has some nonzero probability of being assigned to each treatment). There can be either structural zeros or random zeros, which affects weighting, bias, and estimates	If a subgroup of people with a particular covariate value or range of covariate values is present in one treatment group, but not the other, then the treatment groups aren't comparable (or at least this subgroup can't be included in the comparison), because there's no one in the other treatment group to whom these subgroup members can be compared. Cole and Hernán (2008) note that either "structural zeros" (true positivity violations) or "random zeros" (empirical positivity violations) can occur. Random zeros are due to sampling variability. Positivity violations can cause extreme weights and can bias estimates.	Positivity violations can be detected by plotting PS distributions in an overlay plot or by looking at sample sizes by PS stratum and treatment condition.
Exchangeability – The PS balances groups only on measured confounders; therefore, the treatment groups are balanced on all confounders only if there is no unmeasured confounding.	If there's unmeasured confounding, then the results of the PS analysis will be biased.	Using theory, expert opinion, and prior evidence to select covariates related to the outcome protects against violations of the exchangeability assumption. Sometimes automated methods (e.g., the high-dimensional PS; Schneeweiss et al., 2009; Patorno et al., 2014) are used to select measured confounders into the PS model, but this approach is risky because controlling for instruments (i.e., variables that are associated with treatment but unassociated with the outcome) can amplify bias (Brookhart et al., 2006; Wyss & Stürmer, 2014).
Consistency (stable unit-treatment value assumption, or SUTVA) – The effect of a particular treatment for a particular person is constant. It doesn't depend on what treatment anyone else received. Also, a person's counterfactual outcomes don't depend on which treatment (or which version of the treatment) that person received.	Crossover (treatment contamination) is an example of violation of the consistency assumption. Another example is variation in treatment, as in the fertilizer example described by Rubin (1986).	Good study design prevents violations of the consistency assumption, or SUTVA. For example, using multiple sites with sufficient distance between sites can prevent treatment contamination; assessing treatment fidelity can guard against variation in the potency of the treatment or allow adjustment for such variation.
Correct model specification – All confounders are included in the PS model with the correct functional forms.	An incorrect PS model will cause bias due to imbalance on covariate distributions.	Conducting balance checks and using different functional forms (such as interactions or quadratic terms involving the variables with the greatest imbalances) prevents violations by reducing imbalance.

PS analysis includes the following steps:

(1) Based on theory, expert opinion, and previous evidence, select covariates that are related to the outcome (and may be related to treatment assignment). Because including instruments (i.e., covariates that are related to treatment but not to outcomes) can amplify bias, avoid including instruments in the PS model.

(2) Use a method such as logistic regression, boosted CART, or the covariate-balancing PS (CBPS) to estimate the PS, i.e., the probability of being assigned to a particular treatment given the selected covariates. (Alternative methods include discriminant analysis, probit regression, and machine learning models such as random forests.)

(3) PSs can be used to estimate different treatment effects, such as the average treatment effect in the population (ATE), in the treated (ATT), or in the untreated (ATU) (Table 2). Based on the treatment effect you wish to estimate, apply a PS application method: stratification or inverse-probability-of-treatment weighting (IPTW) to estimate ATE, or matching or standardized mortality ratio weighting (SMR weights, "weighting by the odds," ATT weights) to estimate ATT.

Table 2. Using PS application methods to estimate different treatment effects

Method	**ATE (in population)?**	**ATT (in treated)?**	**ATU (in untreated)?**
Weighting	IPTW (inverse-probability-of-treatment weights, i.e., inverse of the probability of the treatment actually received) T: 1/ps C: 1/(1-ps)	SMR (standardized mortality ratio) weights, ATT weighting, "weighting by the odds" T: 1 C: ps/(1-ps)	Analogous to SMR weights T: (1-ps)/ps C: 1
Matching	No	Yes – select treated observations and then find matches	Yes – select comparison observations and then find matches
Stratification *	Yes	No	No
Regression adjustment (ANCOVA)	Yes (but this would result in a conditional estimate unless marginal effects were specifically requested)	No	No

T = treated; C = comparison; ps = propensity score

* Guo and Fraser (2014) mention the potential use of stratification to estimate ATT (which would also mean that estimating ATU is possible), but they do not demonstrate this application and I have seen no example of it.

(4) If using weights, then check the weight distribution. With IPTW, the mean stabilized weight in the full sample should be approximately 1. If the mean is far from 1 and/or there are extreme weights, then there may be a positivity violation or a model specification problem (e.g., too many covariate categories or too many correlated covariates).

(5) Check covariate balance. Rubin (2001) introduced the B and R statistics, which are based on the PS logit, log(ps/(1-ps)). The most commonly used measure is the standardized absolute mean difference (SAMD).
   - B statistic – the standardized mean difference on the PS logit – should be near zero.
   - R – the variance ratio on the PS logit – should be near 1 (according to Rubin, ≤4/5 or >5/4 indicates at least moderate imbalance; ≤1/2 or >2 indicates severe imbalance).
   - SAMD on individual covariates – should be near zero (various guidelines: <.10, <.20, or <.25)
   - Median, mean (ASAMD), maximum SAMD across covariates – should be near zero (various guidelines: <.10, <.20, or <.25)

(6) If there's residual imbalance, then adjust the PS model as described in Table 1: add interactions or quadratic terms involving the covariates with the greatest imbalances. Other options include using more flexible functional forms such as polynomials or splines, or using nonparametric models (e.g., boosted CART) to estimate the PS.

(7) Examine the PS distributions by treatment group (e.g., create overlaid density plots or display sample size by PS stratum and treatment group). Check for sufficient overlap versus positivity violations. As demonstrated by Cole and Hernán (2008), it may be useful to check overlap on specific key covariates as well. (Comparing PS distributions before and after applying PS methods can also help verify that the methods worked as expected. For example, the treatment group should have higher PSs in general than the comparison group; also, irregularities in the distributions may indicate problems with variable coding or with positivity violations.)

(8) If there's not sufficient overlap, then trim observations from the non-overlapping portions of the distribution and proceed to make inferences for a narrower but less well-defined population; alternatively, conclude that the data do not provide sufficient information about the population of interest.

(9) Obtain the crude estimate of the treatment effect (e.g., the between-group difference in means or proportions with no adjustment whatsoever).

(10) Obtain the PS-adjusted estimate of interest (e.g., ATE, ATT, ATU).

(11) Compare the crude and adjusted estimates to see how much difference the PS adjustment makes (or how necessary the PS adjustment is). Progressively adding covariates to the PS model in blocks can provide additional information about this and can give a sense of whether and when a point of diminishing returns has been reached (i.e., additional adjustments make little difference).

(12) Check for heterogeneity of the treatment effect (i.e., variation of the treatment effect over the PS distribution). For example, stratify by the PS and check how much the treatment effect varies across strata. Variation could be real or could be due to unmeasured confounding.

(13) If using weights, examine the sensitivity of the estimates to various levels of truncation (e.g., truncating the lowest and highest weights according to percentiles 0 and 100, 1 and 99, 5 and 95, 10 and 90, 25 and 75, 50 and 50; weights below the lower percentile cutoff are increased to match the lower cutoff, and weights above the higher percentile cutoff are decreased to match the higher cutoff). Weight truncation provides a way of exploring the bias-variance tradeoff; truncating the weights eliminates extreme weights and therefore narrows the confidence intervals, but also introduces some bias because the weights are no longer exactly IPTW (Cole & Hernán, 2008). Cole (personal communication, April 17, 2013) once suggested choosing an estimate based on minimizing approximate mean squared error (MSE).

Assuming that the IPTW estimate is the truth, MSE can be estimated as bias squared plus variance (i.e., the square of the difference between the current estimate and the IPTW estimate, plus the square of the standard error of the current estimate). This procedure amounts to picking an estimate that is close enough to the IPTW estimate but has a narrower confidence interval.

(14) Apply asymmetric trimming in order to examine the sensitivity of the findings to observations in the tails of the PS distribution (i.e., observations that were treated contrary to prediction) (Stürmer et al., 2010). Sensitivity to these observations may indicate unmeasured confounding (Kurth et al., 2006); Lunt et al., 2009; Stürmer et al., 2010). Useful percentile cutoffs might be 0 and 100, 1 and 99, 2.5 and 97.5, and 5 and 95. Trimming involves excluding observations in both treatment groups whose PS is below the lower cutoff or above the upper cutoff. In contrast with the cutoffs for weight truncation, the cutoffs for trimming are calculated asymmetrically: the lower cutoff is based on the PS distribution in the treated, whereas the upper cutoff is based on the PS distribution in the comparison group. Therefore, applying the 0/100 pair of cutoffs implies restriction to the common support region, the area of overlap in PS distributions.

References

Many of the ideas in this document arose from the course readings, whose authors deserve credit; see the syllabus for details. Some of the more obvious citations have been included in the text above; the corresponding sources are listed here.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Sturmer, T. (2006). Variable selection for propensity score models. *Am J Epidemiol, 163*(12), 1149-56.

Cepeda, M. S., Boston, R., Farrar, J. T., & Strom, B. L. (2003). Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am J Epidemiol, 158*(3), 280-7.

Kurth, T., Walker, A., Glynn, R., Chan, K., Gaziano, J., Berger, K., & Robins, J. (2006). Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *Am J Epidemiol, 163*(3), 262-270. doi:10.1093/aje/kwj047

Lunt, M., Solomon, D., Rothman, K., Glynn, R., Hyrich, K., Symmons, D., & Stürmer, T. (2009). Different methods of balancing covariates leading to different effect estimates in the presence of effect modification. *Am J Epidemiol, 169*(7), 909-917. doi:10.1093/aje/kwn391

Rassen, J. A., & Schneeweiss, S. (2012). Using high-dimensional propensity scores to automate confounding control in a distributed medical product safety surveillance system. *Pharmacoepidemiology and Drug Safety, 21*(S1), 41-49.

Rosenbaum, P., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41-55.

Rubin, D. B. (1986). Comment: Which ifs have causal answers. *Journal of the American Statistical Association, 81*(396), 961-962.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Serv Outcomes Res Methodol, 2*(3), 169-188. doi:10.1023/A:1020363010465

Schneeweiss, S., Rassen, J. A., Glynn, R. J., Avorn, J., Mogun, H., & Brookhart, M. A. (2009). High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology (Cambridge, Mass.), 20*(4), 512-522. doi:10.1097/EDE.0b013e3181a663cc [doi]

Sturmer, T., Joshi, M., Glynn, R. J., Avorn, J., Rothman, K. J., & Schneeweiss, S. (2006). A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *J Clin Epidemiol, 59*(5), 437-47. doi:S0895-4356(05)00224-6 [pii] 10.1016/j.jclinepi.2005.07.004

Sturmer, T., Rothman, K. J., Avorn, J., & Glynn, R. J. (2010). Treatment effects in the presence of unmeasured confounding: Dealing with observations in the tails of the propensity score distribution--a simulation study. *American Journal of Epidemiology, 172*(7), 843-854. doi:10.1093/aje/kwq198

Wyss, R., & Sturmer, T. (2014). Commentary: Balancing automated procedures for confounding control with background knowledge. *Epidemiology (Cambridge, Mass.), 25*(2), 279-281. doi:10.1097/EDE.0000000000000068