DATA, SCIENCE, & NURSING VALUE

Title: Propensity Scores and Regression Adjustment

AUTHORS

Marcelo Coca Perraillon, PhD
Assistant Professor
Department of Health, Systems, Management & Policy
University of Colorado School of Public Health
Aurora, CO.

Kelsey M. Owsley, MPH
Doctoral Candidate
Department of Health, Systems, Management & Policy
University of Colorado School of Public Health
Aurora, CO.

John M. Welton, PhD, RN, FAAN
Professor
University of Colorado College of Nursing
Aurora, Colorado

Abstract

Propensity score methods are commonly used in the analysis of observational data as an alternative or in combination with regression adjustment. Both propensity scores and regression adjustment require the same assumptions to obtain causal estimates of treatment effects: all confounders must be observed and included in the models. However, propensity scores are helpful when there is lack of overlap in one or more covariates.

Propensity score methods are widely used in the analysis of observational data. Many articles and textbooks are available describing propensity scores and best practices for its implementation (Austin 2008; Hill 2008; Guo and Fraser 2014; Gelman and Hill 2006). In this article, we focus our attention on the assumptions needed for propensity scores to provide unbiased estimates of causal effects and the difference between regression adjustment and propensity scores. Although propensity scores are helpful for diagnosing potential problems that may bias treatment effect estimates, they are somewhat limited as a tool for providing estimates of causal effects since the assumptions needed are the same as assumptions for regression adjustment: all potential confounders must be accounted for in the model (Austin, Grootendorst, and Anderson 2007; Heckman, Ichimura, and Todd 1997). Furthermore, regression adjustment and propensity scores yield similar results in many circumstances. We introduce three hypothetical examples to describe the key concepts of lack of balance and lack of overlap and illustrate the use of propensity scores with a study examining the relationship between nurse practitioners and quality of care.

Balance, overlap, and regression adjustment

Consider a hypothetical trial studying the effect of nurse experience on patient outcomes. Suppose it would be possible to randomize a large group of patients to be treated by either an experienced nurse (treated group) or a recent graduate (control group). The outcome is a continuous index measuring health status after 30 days (for simplicity, we can assume that the health index is normally distributed). Because of randomization into treatment groups and large sample size, both treatment and control groups are likely to be comparable. The outcome of the experiment would be the same regardless of which group receives the treatment. Randomization guarantees that there are no unmeasured confounders (i.e. treated and control groups on average have similar observed and unobserved characteristics). It would be appropriate to estimate treatment effects by a simple t-test for independent samples or a regression model in which the health index is the outcome and the only predictor is an indicator variable equal to 1 if the patient received care from an experienced nurse. Testing if the coefficient on the indicator variable equals zero – that is, testing if the mean outcome is the same in both groups--is equivalent to a t-test of independent samples.

Now consider a slightly different version of this experiment. Supposed that due to ethical concerns a review board determined that patients of higher acuity should have a greater likelihood of receiving care from an experienced nurse. The (hypothetical) investigators determined that patients who score in the upper quartile on the Nursing Outcome Classification (NOC) will be 40% more likely assigned to be assigned into the treated group. This is an example of conditional randomization; treatment assignment is conditional on a covariate or set of covariates. In this version of the experiment, patients are randomized conditional on their NOC score. Consequently, patients in the treated group are not comparable to those in the control group. On average, acuity in the treated group is higher than that of the control group. However, both groups have patients of varying acuity levels. Higher acuity patients have a greater chance of being in the treated group but are not excluded from the control group. In this example, a t-test of mean

outcomes is no longer appropriate. A regression model controlling for NOC score must be used. In other words, the statistical model must (figuratively) “hold acuity constant” in order to estimate unbiased treatment effects. In this sense, regression adjustment is one of the oldest causal inference methods available.

This second version of the experiment provides an example of lack of *balance* but complete *overlap*. Imbalance occurs if the distributions of relevant pre-treatment variables differ for the treatment and control group (e.g. NOC score mean and/or standard are different). Lack of complete overlap occurs if there are regions in pre-treatment variables where there are treated observations but no controls, or controls but no treated observations (Gelman and Hill 2006). Figure 1 shows examples of lack of balance and lack of overlap for one variable. In first version of the experiment, unconditional randomization, there is balance and overlap in NOC scores because randomization guarantees that both groups have the same distribution of NOC scores. In the second version of the experiment, there is overlap but no balance (Figure 1, Panels B or C).

Finally, consider a third version of this experiment, one that is more likely to occur when observational data are used; that is, situations in which the investigator does not have control over treatment assignment. In this version of the experiment, patients who score high on the NOC are only assigned to experienced nurses. Patients who do not score high are randomly assigned to either group. Consequently, the control group has no patients of high acuity. However, both groups have patients of low-medium acuity who did not score high on the NOC. This is an example of both lack of balance and lack of complete overlap (Figure 1, Panels E and F). This third experiment could be analyzed with a regression model controlling for acuity, but it is not certain that the estimated treatment effects are unbiased because information from patients of high acuity in the treated group is used to *extrapolate* the outcome in the control group. However, the control group does not have patients of high acuity. Thus, the counterfactual does not exist for the high acuity subgroup. If the analysis were restricted to patients that are not of high acuity, then the experiment is similar to the first case and treatment effects would be unbiased.

When there is lack of overlap, as in the third experiment, propensity score methods offer a way to perform an analysis that limits the estimation to regions where a comparison can be made (i.e. where overlap occurs). However, if overlap is not a problem, then propensity score and regression adjustment provide similar answers (Drake 1993; Shah et al. 2005). In the examples above, lack of overlap can be easily diagnosed graphically as in Figure 1, but in practice lack of overlap can be due to multiple variables. An additional advantage of propensity scores is that the propensity score is a one-number summary of all covariates and can be used to diagnose lack of overlap created by multiple variables.

Estimation and use

The propensity score is the probability that an observation is assigned to a treatment given the value of covariates. In other words, it is the *propensity to receive treatment* conditional on the value of covariates. This probability can be estimated with a logistic model in which the outcome is an indicator variable equal to 1 if the observation received treatment. The predictors in the model are potential confounders (variables correlated with both the treatment and outcome). The fundamental theorem of propensity scores states that observations that have the same propensity score have the same distribution of all the variables that were used to estimate the propensity score (Rosenbaum and Rubin 1983). Graphs like Figure 1 of the propensity score distribution can be used to diagnose lack of overlap in *all variables*. Note that is in this sense that the propensity score provides a summary of overlap for all the variables used in the estimation of the propensity score.

Besides diagnosing lack of overlap, the propensity score can be used to restrict the estimation to regions where there is overlap (also known as "common support") in three ways: matching, stratification, and inverse probability of treatment weighting (IPTW). In matching, each treated observation is paired to one or more "similar" observations from the control group. The propensity score is used as the metric summarizing similarity. If a treated observation does not have a similar match in the control group, that observation is excluded from the analysis. In stratification, the propensity score is divided into categories –quintiles, for example—and then comparisons are made within each category. With IPTW, regression models are weighted by the inverse of the propensity score. This allows for comparable observations to be given given more weight. One advantage of IPTW over matching is that all observations are included in the analysis, thus increasing sample size. For more details on each propensity score approach, Guo and Fraser (2014) provide an excellent guide to propensity scores and their application.

One source of common confusion regarding propensity scores is that there is no guarantee that groups are comparable in covariates that were not included in the propensity score model. Therefore, regardless of how the propensity score is implemented –matching, stratification, or IPTW—the internal validity of the study must be judged based on whether *all potential confounders were included in the model*, and not by whether the method created comparable groups. Showing that groups have similar observed characteristics after matching, stratification, or IPTW does not prove that groups have similar *unobserved* characteristics. All the factors that affect selection into treatment and the outcome must be accounted for to obtain valid estimates of causal effects (Perraillon, Welton, and Jenkins 2019).

Example

Kurtzman and Barnow (2017) is a typical application of propensity scores. The authors were interested in the causal relationship between quality outcomes and Nurse Practitioner (NP) practice restrictions. Advocates for NP practice restriction laws argue that physicians provide higher quality of care than advanced practice providers in primary care settings. Yet, there is little evidence substantiating these claims in times when many geographical areas lack enough primary care physicians. At the same time, states that adopt more liberal practice authority laws for NPs may have a greater proportion of areas with limited access to primary care and have different characteristics than states with stricter practice authority laws. These same characteristics may influence individual health outcomes, and thus are confounders that could bias comparisons.

The study compared patient visits seen by a NP in states with restrictions on NP practice authority to states where NPs have practice independence. Data came from the 2006-2011 National Ambulatory Medical Care Survey (NAMCS). The propensity score model included numerous characteristics that are expected to influence practice laws and quality outcomes. The authors then generated matched groups using the propensity score, discarding observations without a good match. The study found no differences in quality between states that restricted NP practice authority compared to states that allow NP practice independence. The validity of this conclusion depends on whether all potential confounders were included in the propensity score model. The authors did not compare their results to regression adjustment.

Conclusion

Propensity score methods and regression adjustment are closely connected. Both methods require the same assumptions for causal inference: all potential confounders must be observed and included in the statistical model. However, propensity score methods are useful in situations when there is lack of overlap in one or more variables. If lack of overlap is not a concern, propensity scores and regression adjustment provide similar estimates.

References

Austin, Peter C. 2008. 'A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003', *Statistics in medicine*, 27: 2037-49.

Austin, Peter C, Paul Grootendorst, and Geoffrey M Anderson. 2007. 'A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study', *Statistics in medicine*, 26: 734-53.

Drake, Christiana. 1993. 'Effects of misspecification of the propensity score on estimators of treatment effect', *Biometrics*: 1231-36.

Gelman, Andrew, and Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models* (Cambridge university press).

Guo, Shenyang, and Mark W Fraser. 2014. *Propensity score analysis: Statistical methods and applications* (SAGE publications).

Heckman, James J, Hidehiko Ichimura, and Petra E Todd. 1997. 'Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme', *The review of economic studies*, 64: 605-54.

Hill, Jennifer. 2008. 'Discussion of research using propensity-score matching: Comments on 'A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003'by Peter Austin, Statistics in Medicine', *Statistics in medicine*, 27: 2055-61.

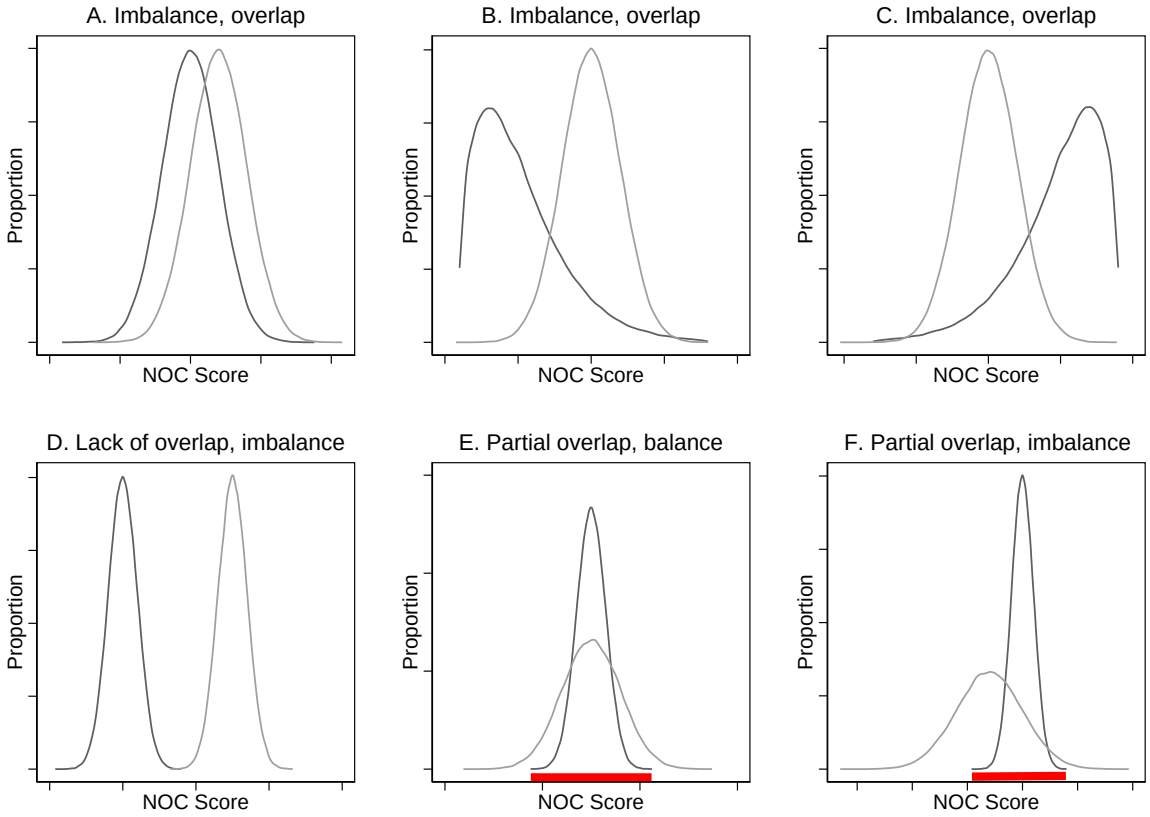
Kurtzman, Ellen T, and Burt S Barnow. 2017. 'A comparison of nurse practitioners, physician assistants, and primary care physicians' patterns of practice and quality of care in health centers', *Medical Care*, 55: 615-22.

Perraillon, Marcelo, John Welton, and Peggy Jenkins. 2019. 'Measuring Nursing Care Value, Big Data, and the Challenges of Estimating Causal Effects', *Nursing Economic$*, 37: 213-15.

Rosenbaum, Paul R, and Donald B Rubin. 1983. 'The central role of the propensity score in observational studies for causal effects', *Biometrika*, 70: 41-55.

Shah, Baiju R, Andreas Laupacis, Janet E Hux, and Peter C Austin. 2005. 'Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review', *Journal of clinical epidemiology*, 58: 550-59.

Figure 1. Balance and overlap



Simulated distribution of NOC scores. Panels A, B, and C show situations in which there is complete overlap but lack of balance (both average and standard deviation are different). Panels D shows distributions with both lack of balance and overlap. Panel E shows distributions with balance in means but lack of complete overlap. Panel F shows both lack of balance and overlap. The thick lines on x-axis in panels E and F show regions of overlap. Propensity scores could be used to analyze E and F, but not D. Regression adjustment and propensity scores would yield similar estimates in panels A to C, since there is complete overlap. NOC: Nursing Outcome Classification. Adapted from Gelman and Hill (2006).