2012 CCPRC Meeting
Methodology Presession Workshop
October 23, 2012, 2:00-5:00 p.m.

Propensity Score Methods for Estimating Causality in the Absence of Random Assignment: Applications for Child Care Policy Research

Description

The goal of this session was to provide a detailed overview of propensity score methods--a promising approach for analyzing administrative data and for conducting policy-relevant research.

This session started with a presentation which included:

- Background on propensity score methods
- Guidelines for building a propensity model
- Steps for estimating propensity scores
- Use of propensity scores for matching, weighting, and subclassification
- Specific application of propensity score methods using real data.

The presentation was followed by a facilitated discussion during which participants addressed: ways in which propensity score methods can be applied to child care policy-related research; situations for which propensity scores methods are appropriate and effective; and the advantages and disadvantages of using propensity score methods for child care policy research.

Facilitator

Kathleen Dwyer, Office of Planning, Research and Evaluation

Presenter

Donna Coffman, Methodology Center, Pennsylvania State University

Discussant

Elizabeth (Liz) Davis, University of Minnesota

1. Documents in Session Folder

- “Propensity Score Methods for Causal Inference;” Donna Coffman (contact presenter to obtain presentation, dcoffman@psu.edu)

2. Summary of Presentations

- Opening: Kathleen Dwyer
  - Kathleen opened the session by indicating that this is the inaugural CCPRC methodology workshop. It is designed as a professional development opportunity to explore strategies for conducting rigorous research that addresses child care policy questions while making efficient use of existing resources such as administrative data.
  - Donna Coffman will provide a detailed overview of propensity score methods and then go through an example of a specific application of propensity score methods using real data. Liz Davis will present a summary of key points and then facilitate an

interactive discussion of propensity score methods and their application to child care policy research.

- **Summary of Presentation #1: Donna Coffman**
  - Propensity scores are based on what is called a counterfactual framework or the potential outcomes framework for causal inference; it has also been called Rubin's Causal Model. It has been thoroughly reviewed in articles and books.
  - In the simplest case, there is a treatment or exposure for each individual and this is denoted with an indicator variable *t* that equals one for a treated individual and zero when the person is not treated.
    - For example, children are randomly assigned to either receive Head Start (HS) or parental care at home. The observed outcome is *y*. Does the treatment cause a difference in the mean value of *y*?
    - We know that correlation is not causation; it is not sufficient to show a significant difference between mean outcomes for the HS group (treatment) versus the mean outcome for the parental care group (control). We need to rule out alternative explanations and systematic differences between the groups in addition to the treatment that they received.
    - To characterize a treatment effect in an observational study, Rubin introduced a notation for potential outcomes.
      - The treatment received by each subject is a dummy variable that's zero or one; y_{i0} is the potential outcome for subject *i* if the subject is in parental-based care and y_{i1} is the potential outcome for subject *i* if they go to HS. The difference between the two is the causal effect of receiving HS versus parental care for that individual.
    - Causality is defined in terms of outcomes for a particular individual, but causal inference can never be observed for any child because they either received HS or parental care (and we can’t redo history). That's why this called the counterfactual framework. However, by making certain assumptions, it is possible to estimate the causal effect (*average causal effect*) for the population.
    - Average causal effect for the population is distinct from *the average causal effect for the treated* (the causal effect of HS among those who received HS).
    - In a randomized experiment, the treatment is independent of potential outcomes. We take the average of the observed *y*s for those who received HS and get an unbiased estimate of the potential outcome of HS. And then we take the average of the observed values in the control group, and that's an unbiased estimate of the average potential outcomes under the control group. We estimate the difference in the means between the two groups and do a t-test.
    - In a typical observational study, it's unlikely that the treatment indicator will be independent of the potential outcomes. Subjects select their own treatments, which results in the treated and untreated groups being systematically different at baseline. The characteristics on which they differ are potential confounders and the differences between the means of the treatment and control groups are biased.
    - *Confounders* are pre-treatment variables that may jointly influence both the treatment and outcome. And the regression approach, which is the underlying idea of ANCOVA and regression modeling, is to define the causal effect as the

average difference in the mean response between treated and untreated persons holding constant these covariates, i.e., what is the mean difference between the treatment and the control group given that the child is female, that she qualifies for reduced free lunch, etc.

- Traditional methods, e.g., regression, model the relationship between the covariates and the outcome. Propensity scores control for confounding by modeling the relationships between the covariates and the treatment assignment.
- Rosenbaum and Rubin (1983) define the propensity score as the probability of receiving the treatment given covariates. These are similar to selection probabilities in sample surveys, but unlike survey selection probabilities, the propensities are unknown and must be estimated. The propensity score is the probability of selecting into the treatment. The propensity score is a single number summary of 40, 50, maybe 60 potential confounders.
- Propensity scores are not a panacea to making causal inferences in observational data, but propensity scores do allow many covariates to be summarized with a single summary number. Further, propensity scores don't require assumptions about the functional form, e.g., linearity or quadratic terms or how the covariates are related to the outcome.
- Treated and untreated persons with identical propensity scores have identical distributions for all the covariates. This means that if we divide the population into groups of constant propensity, then subjects in each group can be treated as if they had participated in a randomized experiment.
- Propensity scores can only balance the covariates included in the model and are based on the critical but untestable assumption that all the confounders are measured and included in the model. (Sensitivity analyses can be done.)

- Other assumptions assumed in this presentation are that the treatment applied to one subject does not affect the outcome of any other subject. This may not be the case in more advanced models including those with multi-level settings. We assume each person could have been exposed to either treatment.
- There are four general analysis steps:
  - *What is it that we want to estimate the causal effect of*—e.g., the whole population, those who receive the treatment, by gender, etc.?
  - *Select covariates and estimate propensity scores*.
    - Do not adjust for post-treatment variables as if they were confounders (anything that could have been affected by the treatment should not be in the list of confounders). Generally, confounders come from a list of baseline characteristics thought to be related to the treatment and outcome. Also, include any baseline variable thought to be correlated with the outcome.
    - Including unnecessary predictors is not going to bias your estimate of causal effect, but omitting important predictors will. Caveats: variables that are strongly related to the treatment but not the outcome (instrumental variables) will decrease precision; variables that are strongly related to the outcome, but not the treatment, will increase precision and not affect bias. Do not include instrumental variables in the model.

        - In terms of the relationship between theory and the predictors, this approach predicts who is going to receive the treatment. It is not about coming up with some parsimonious model that fits some theory.
        - The next step is to estimate the propensity scores; the common way to estimate the scores is to use logistic regression with the treatment indicator *t* as the outcome. The confounders are used as predictors.
    - *Propensity scores are then used to control for potential confounding.* Alternative methods for using propensity scores include matching, inverse-propensity weighting (IPW) and subclassification. IPW has been used for many years to adjust for unequal probabilities in selection.
        - The basic idea of IPW is that the treated and untreated groups are not simple random samples from the population. The treated group has an over sampling of people with a high propensity of selecting into the treatment and the untreated group has an over sampling of people with low propensities to select into the treatment.
        - To get an unbiased estimate of the mean for the whole population, we assign less weight to the over sampled cases and more weight to the under sampled ones.
        - The weighted average of the two groups is an estimate of our average *causal effect on the population*. (Weights should be stabilized for the groups.)
        - To estimate the *causal effect among the treated*, we need to weight the untreated sample to resemble the population of treated persons. The treated sample does not need to be weighted because it is already a random sample from the treated population.
        - Note—one way that propensity score methods are different from regression analysis is that with propensity score methods, you change your data in a way that makes the data mimic the data that you would have gotten from a randomized experiment (for the untreated group).
        - Features of IPW include that if the propensity model is incorrectly specified, the estimates will be biased; and even if the propensity model is correct, it may be inefficient. Outliers are given too much weight; slight misspecification of the model can have great impact on the largest weights; deleting cases with huge weights is not recommended because these cases are useful.

- **Summary of Presentation #2: Donna Coffman (Application of Propensity Score Methods)**
    - Diagnostics: box plots of the propensities or logit propensities will reveal the degree of overlap or common support (without overlap, causal inferences really are not supported by the data).
        - If no one in the treatment group has a propensity score similar to someone in the control group (multiple individuals can match an individual in the other group), then you would not be able to find a match for that person. If that happens, the whole thing is going to fall apart.

    - If you include lots of covariates, it can actually exacerbate the problem of common support. But excluding them can violate unconfoundedness.
    - The goal of diagnostics is to assess whether or not we've achieved balance on the covariates between the two groups. The goal is to have similar covariate distributions in the treated and control groups.
        - If you're creating groups of individuals with similar propensity scores, you need to check the balance in any subclass.
        - If you have not achieved the balance, go back to the propensity score model and include interaction terms or quadratic terms, etc., and fit the propensity score model again until you achieve balance. If you can't achieve balance, then causal inferences are not warranted.
- Most of these techniques are little more than linear and logistic regression. The analysis can be done in many different statistical packages. Stata and R have dedicated propensity score packages that will do the balance checks. SAS and SPSS have more limited macros and functions.
- An example from a paper currently in second review for the *Journal of Primary Prevention* (contact Donna for a copy). The data are from the Early Childhood Longitudinal Study, kindergarten cohort.
    - 1,701 children attended HS and 3,362 received parental care prior to entering kindergarten. The outcome is a standardized assessment of reading completed in the fall of kindergarten.
    - Causal questions include: What difference in reading development would we expect to observe if all the children in our study had enrolled in HS, compared to if all of them had received parental care? Alternatively, what difference in reading development would we expect to observe if all children who attended HS had instead received parental care? And a question we weren't able to answer: what is the causal effect of HS versus parental care among those who are eligible for HS?
    - We used inverse propensity weighting, sub-classification and ANCOVA to estimate the average causal effect for the whole population. And we used inverse propensity weighting, sub-classification and matching (two types), to estimate the average causal effect among those who went to HS. We looked at the mean difference on the reading skills for those who attended HS versus those who attended parental care.
    - Donna spent time highlighting tables and figures related to the study:
        - The first table (descriptive statistics) shows sample size and missing data on some of the confounders. With 40-50 confounders, they used multiple imputation to impute all of the confounders. And then, they did propensity score estimation and matching, with weighting or sub-classification in each of the imputations. The causal effect estimates across imputations were combined. The chart shows the standardized mean difference between the HS group and the parental care group on each of the confounders, e.g., socio-economic status is 0.61.
        - The next table examines balance and shows standardized mean differences for subclassification, weighting and matching. Weighting looks at the average treatment effect and average causal effect among the children who went to HS. (Average causal effect measures how much HS helped children who went

        to HS.) This results in standardized mean differences that are close to zero (0.2 is the rule of thumb but you can go to 0.15).
        - Matching was done two different ways: nearest neighbor and optimal. In both cases, matching was done on a one to one basis. Again the standardized mean differences dropped. (So in matching, the parent care group (3362) was reduced to the 1701 cases that best matched the HS children.)

- **Summary of Presentation #3: Liz Davis**
    - Given that there is a range of knowledge among attendees, Liz reviewed some of the key points of propensity score methods and used a child care specific example of the use of these methods.
    - *Why are we talking about these methods?* This is a problem of the counter factual--we want to know the impact of something, whether it's participating in a program or having some kind of treatment. "What would have happened in the absence of treatment?"
    - *Random control trials (RCTs)* are used, when possible, to ensure that participation in the treatment is the only difference between those who are subject to the intervention and those who are excluded from it. But in many situations, RCTs aren't feasible, and we have observational data that have not been randomly assigned to program participation (or to treatment).
        - This may result in biased estimates of the effect of the treatment/program on outcomes because the treated group and the control group differ on characteristics and some of those differences relate to the fact that some of them ended up in the treatment group and some of them did not, i.e., program participation is non-random (e.g., administrative issues or self-selection).
    - *When can we use propensity score methods?* There are two conditions/assumptions to keep in mind if you're going to use these methods.
        - The first is that *all relevant characteristics have to be observable to the researcher*, i.e., you have data on the characteristics of the variables that matter. This is called selection on observables or unconfoundedness.
        - The second key assumption is called the *common support condition*; for each value of X (the variables going into the logistic equation), there's a positive probability of being both treated and untreated, i.e., each person could have been in either the treatment or the control group.
    - *What is the propensity score?* It's the probability that a member of the population receives the treatment/program given a set of observed variables. The propensity score is usually estimated in a logistic model, but other methods are used as well.
        - We have variables (confounders) that we think are related to the probability of being in the treatment group. Thinking of them as confounders reminds us that they have to do with why individuals are in the treatment group. To deal with confounding, we need a way to estimate the propensity score.
    - *What's the key to success with propensity score methods?* The most important thing is the selection model. This model is what makes this approach different from regression methods. It causes you to stop and think about why some observations are in the treatment group and others are not.

  - It’s important to consider all the variables that might determine participation or make the treatment group different than the control group. Are there explicit criteria for determining participation, e.g., eligibility for child care subsidies depends on working, having a certain income level, using non-parental care, etc.?
  - An advantage of propensity score methods is that you can use a very flexible, functional form for estimating the logistic regression model. So, although the logit model itself is basically linear, you can use interaction terms and squared and higher order terms to capture potential nonlinearities.
- *When can we use propensity score methods in terms of data?* It helps to have a lot of data because: you want to be able to capture everything you think matters to being in the treatment group that’s also related to the outcome; and depending on the method you’re using, having a reasonably large sample size is also helpful. For instance, in Donna’s example, if you're matching one to one, you're reducing your sample size to whichever is the smaller group.
- *Using the same source of data for the treatment and control groups is important* to make sure the variables are defined and measured the same for both groups. Also, when you're dealing with missing data or data errors, you want to make sure you've handled them the same way for both the treatment and the control group.
- *Why would you use propensity score methods rather than regression?*
  - The key advantage of propensity score methods is that it forces you to think about selection into the treatment group and makes explicit the comparison between treatments and controls. This provides a way to model and think very explicitly about selection process.
  - Another advantage is that it allows for heterogeneous treatment effects, i.e., that treatment effects or the size of the effect on the outcome might differ for different subgroups of the population (boys versus girls).
  - In regression methods, you can also sometimes model heterogeneous treatment effects, but you have to have the right functional form to do that. In regression models we may include interaction terms. If you have the model right, know the functional form, or if treatment effects really are homogenous, then regression is better. Regression estimates are more efficient and have lower variance and smaller standard errors. Often however, the treatment effects are likely to differ across subgroups or we aren’t able to get the functional form correct.
  - *Are propensity score methods a magic bullet? Unfortunately, not.*
    - Unconfoundedness and common support assumptions are key to propensity score methods giving us valid estimates of the causal effects. It can be difficult to obtain balance in small samples.
    - Most importantly, *matching is not appropriate when selection into the treatment is based on unobservable characteristics that are correlated with the outcomes of interest.* This occurs when you have something else driving selection into treatment that you don't have the data on, e.g., in a lot of studies of women's labor force participation, things like ability and motivation and personality traits are considered important and we often don’t have measures of those.

        - In terms of matching, there are a lot of different options and you need to decide which of the matching algorithms to use. There's not a lot of guidance about this but it is growing.
        - With both regression and propensity score matching, if we have unobservable characteristics that are affecting the decision to be in the treatment and affecting the outcome of interest, neither is going to give you unbiased estimates of the causal impacts.
    - *Example*: a paper by Anna Johnson, Rebecca Ryan and Jeanne Brooks-Gunn was published in the Journal of Child Development in July 2012. The paper is, *Child Care Subsidies: Do They Impact the Quality of Care Children Experience?*
        - The question examined is, if you look at children who are receiving child care subsidies, and you have a measure of the quality of care they're receiving, do they receive higher or lower quality care than other children?
        - This is a good example for the use of propensity score methods for a couple of reasons. We're dealing with a self-selected group of parents who are actively using a child care subsidy. They must meet certain eligibility requirements, but we know that only a small percentage of eligible parents use child care subsidies. What makes a family that uses subsidies different from one that doesn't?
        - These researchers wanted to compare subsidy users to those who are eligible but didn't use a subsidy. And then they wanted to know, what is the effect of using a subsidy on the quality of child care the child is receiving?
        - For this study, the Early Childhood Longitudinal Survey-Birth Cohort is used. This survey includes observational data on quality of care received in preschool.
            - They started with the subsample of children they estimated were eligible for child care subsidies (the group they thought could have been treated).
            - The researchers ran a propensity score logit equation to estimate the probability of using a subsidy among this eligible sample (probability of subsidy take-up).
            - From earlier studies on take-up, they had some guidance about the variables to include in this model, the kinds of things likely to predict using a subsidy among those who were eligible, e.g., family characteristics and parental preferences.
            - This paper describes the process of estimating propensity scores and matching subsidy recipients with eligible non-recipients with similar propensity scores. A model of child care quality is then estimated using that matched group. Briefly, they found that subsidy recipients use better quality care compared to those who were eligible non-recipients who used non-parental care (but not HS or public prek). Subsidy recipients had lower quality care than eligible non-recipients who used HS or prek services.

3. Summary of Discussion and Q&A

- Question about children enrolled in parental care versus HS (Donna's study): for the entire sample (including children not eligible for HS), the study concludes that there is no evidence to expect better or worse reading ability if children who attended HS had received parental care instead. An obvious missing variable is the eligibility question (the treated are a subgroup of children in the study and there isn’t a matched sample on the eligibility question).

- Question about using regression versus propensity scores: propensity score methods try to improve your comparison; you change the data set (develop a single number summary of the covariates) rather than changing the model; in regression, there isn’t a distinction between the treatment variable and all the other variables. However, if you do both regression and propensity score analysis, and use all the same covariates (all literally related to the outcome), your answer should not differ. It is acceptable to use variables that were used to create a propensity score as controls as well, e.g., for theoretical reasons.
- What are the general parameters for how large a sample you need in propensity score methods? It depends on your data. If you have 30 children who went to HS and 30 who didn’t, and they were very similar on the propensity scores you used to match them, it would be acceptable to use these methods. Any power analysis that you would do for a t-test of regression still applies. The question is how many people must be removed from your sample to achieve matching.
    - What about number of variables to observations? You don't have to worry about multicollinearity in the propensity score model because multicollinearity relates to standard errors on the logistic regression coefficients and you’re not concerned with interpreting those. However, in order to fit the propensity model, you need to have more cases than covariates.