Propensity Score Matching

James H. Steiger

Department of Psychology and Human Development
Vanderbilt University

Multilevel Regression Modeling, 2009

Propensity Score Matching

1. Introduction
2. Modeling the Covariates
3. Subclassification
4. Matching
   - Introduction
   - Why Match?
5. Balancing Scores
   - Definition

Matching and Subclassification

In previous discussions, we learned about selection bias and, in particular, the dangers of attempting to control for post-treatment covariates while assessing causality.

Near the end of Chapter 10, Gelman & Hill discuss the methods of *matching* and *subclassification* as aids to causal inference in observational studies.

The basic idea behind the methods is that, *if* you can identify relevant covariates so that ignorability is reasonable, you can assess causality by controlling for the covariates statistically. Such control can take several forms:

- You can examine conditional distributions, conditionalized on classifications on the covariate(s).
- You can match treatment and controls, and compare matched groups
- You can model the covariates along with the treatment

Matching and Subclassification

In previous discussions, we learned about selection bias and, in particular, the dangers of attempting to control for post-treatment covariates while assessing causality.

Near the end of Chapter 10, Gelman & Hill discuss the methods of *matching* and *subclassification* as aids to causal inference in observational studies.

The basic idea behind the methods is that, *if* you can identify relevant covariates so that ignorability is reasonable, you can assess causality by controlling for the covariates statistically. Such control can take several forms:

- You can examine conditional distributions, conditionalized on classifications on the covariate(s).
- You can match treatment and controls, and compare matched groups
- You can model the covariates along with the treatment

Matching and Subclassification

In previous discussions, we learned about selection bias and, in particular, the dangers of attempting to control for post-treatment covariates while assessing causality.

Near the end of Chapter 10, Gelman & Hill discuss the methods of *matching* and *subclassification* as aids to causal inference in observational studies.

The basic idea behind the methods is that, *if* you can identify relevant covariates so that ignorability is reasonable, you can assess causality by controlling for the covariates statistically. Such control can take several forms:

- You can examine conditional distributions, conditionalized on classifications on the covariate(s).
- You can match treatment and controls, and compare matched groups
- You can model the covariates along with the treatment

Modeling the Covariates

The problem with modeling the covariates is that, depending how influential the covariates are,with even minor model misspecification the estimate of the effect of the treatment may be seriously biased.

Since ignorability requires that all relevant covariates be accounted for, the “curse of dimensionality” quickly becomes a factor. A huge number of models is conceivable, and so the likelihood of misspecification is high.

Subclassification on a Single Covariate

Gelman & Hill (p. 204) illustrate subclassification with a simple example.

Suppose that the effectiveness of an educational intervention for improving kids' test scores was investigated in an observational setting where mothers chose whether or not to have their children participate, and randomization was not possible.

Selection bias is a fundamental problem in such a study.

Subclassification on a Single Covariate

Selection bias occurs when the treatment condition (e.g., experimental vs. control) of a participant is not independent of confounding covariates which are also correlated with the outcome.

For example, if mothers' high achievement motivation causes them to select into the experimental group, and also causes them to react to their children in a way that affects the outcome, then the results of the study will be biased.

Subclassification on a Single Covariate

Suppose, for the sake of argument, that there is only one confounding covariate in the study, and it is the level of education of the mother.

One way of controlling for the impact of this covariate is to create subclassifications, within which the covariate has the same value in experimental treatment and control groups.

Here are some illustrative data from Gelman & Hill .

Subclassification on a Single Covariate

Mother's education	Treatment effect estimate ± s.e.	N treated	N controls
Not a high school grad	9.3 ± 1.3	126	1358
High school graduate	4.0 ± 1.8	82	1820
Some college	7.9 ± 2.3	48	837
College graduate	4.6 ± 2.1	34	366

Gelman & Hill suggest computing an "overall effect for the treated" by using a weighted average *only over the treated*, i.e.

$$\frac{(126)(9.3) + (82)(4.0) + (48)(7.9) + (34)(4.6)}{126 + 82 + 48 + 34} = 7.0 \qquad (1)$$

Difficulties with Subclassification

Subclassification has advantages:

- It forces overlap
- It imposes roughly the same covariate distribution within subclasses

However, it has disadvantages as well:

- When categorizing a continuous covariate, some information will be lost
- The strategy is very difficult to implement with several covariates at once

Difficulties with Subclassification

Subclassification has advantages:

- It forces overlap
- It imposes roughly the same covariate distribution within subclasses

However, it has disadvantages as well:

- When categorizing a continuous covariate, some information will be lost
- The strategy is very difficult to implement with several covariates at once

Difficulties with Subclassification

Subclassification has advantages:

- It forces overlap
- It imposes roughly the same covariate distribution within subclasses

However, it has disadvantages as well:

- When categorizing a continuous covariate, some information will be lost
- The strategy is very difficult to implement with several covariates at once

Difficulties with Subclassification

Subclassification has advantages:

- It forces overlap
- It imposes roughly the same covariate distribution within subclasses

However, it has disadvantages as well:

- When categorizing a continuous covariate, some information will be lost
- The strategy is very difficult to implement with several covariates at once

Matching

Matching refers to a variety of procedures that restrict and reorganize the original sample in preparation for a statistical analysis. The simplest version is to do one-to-one matching, with each treatment observation being paired with a control that is as much like it as possible on the relevant covariates.

With multiple covariates, matching can be done on a "nearest neighbor" basis. For example, a treatment observation is matched according to the minimum Mahalanobis distance, which is, for two vectors of scores $\boldsymbol{x}^{(1)}$ and $\boldsymbol{x}^{(2)}$ on the covariates,

$$(\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)}) \tag{2}$$

Why Match?

Even if ignorability holds, imbalance of treatment and control groups can lead to misleading results and model dependencies.

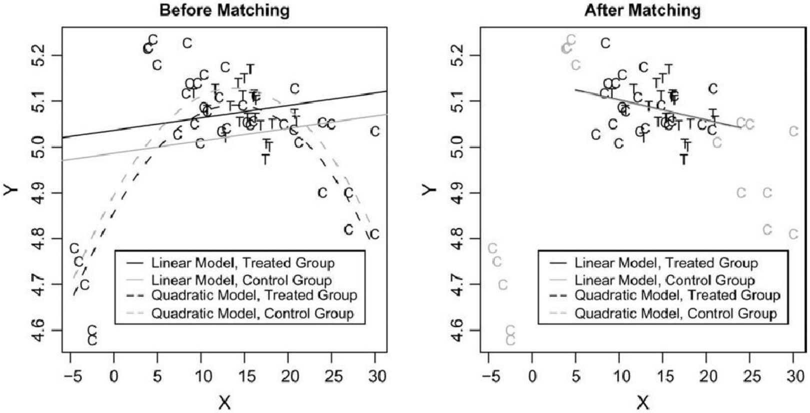
On the next slide, we look at an example from Ho, et al. (2007) which illustrates the problem.

Two models, a linear and a quadratic, are fit to the data.

Ultimately, these two models estimate the causal effect by the average vertical distance between the C's and T's. They differ only in how they compute this average.

In this case, the linear model estimates a causal effect of $+0.05$ the quadratic model estimates a causal effect of -0.04. The presence of control units far outside the range of the treated units creates this model dependence.

Why Match?


Before Matching
After Matching
Y
X
4.6
4.7
4.8
4.9
5.0
5.1
5.2
−5
0
5
10
15
20
25
30
C
T
Linear Model, Treated Group
Linear Model, Control Group
Quadratic Model, Treated Group
Quadratic Model, Control Group

Balancing Scores

Rosenbaum and Rubin(1983) introduced the notion of a *balancing score*, a function of the covariates that may be used in place of all the covariates to achieve balancing.

Balancing Score

Given treatment T and one or more covariates in X, a balancing score $b(X)$ satisfies the condition that the conditional distribution of X given $b(X)$ is the same for treated $(T = 1)$ and control $(T = 0)$, that is

$$X \perp T \mid b(X) \tag{3}$$

Balancing Scores

There are many possible balancing scores. For example, X itself is a balancing score.

Balancing scores can be characterized in terms of *coarseness* or *fineness*.

Coarseness — Fineness of a Balancing Score

A balancing score $a(x)$ is said to be coarser than balancing score $b(x)$ if $a(x) = f(b(x))$ for some function f. In such a case, we can also say that $b(x)$ is finer than $a(x)$.

The Propensity Score

Rosenbaum and Rubin (1983) concentrate on a particular balancing score, the *propensity score*.

The Propensity Score

Given a treatment T and a set of covariates X, the propensity score $e(x)$ is defined as

$$e(x) = \Pr(T = 1|X = x) \tag{4}$$

The Strong Ignorability Assumption

In deriving the key optimalit property of propensity scores Rosenbaum and Rubin (1983) assume *strong ignorability* of T given X.

Strong Ignorability

The property of *strong ignorability* of T given X holds if, for potential outcomes y^1 and y^0, the distribution of these potential outcomes is conditionally independent of T given X, *and* for any value of the covariates, there is a possibility of a unit receiving the treatment or not receiving the treatment. That is,

$$(y^1, y^0) \perp T | X \tag{5}$$

and

$$0 < \Pr(T = 1 | X = x) < 1 \quad \forall x \tag{6}$$

Mathematical Properties

Rosenbaum and Rubin (1983, p. 43–44) proved the following theorems:

1. The propensity score is a balancing score
2. Any score that is “finer” than the propensity score is a balancing score; moreover, X is the finest balancing score and the propensity score is the coarsest
3. If treatment assignment is strongly ignorable given X, then it is strongly ignorable given any balancing score $b(X)$
4. At any given value of a balancing score, the difference between the treatment and control means is an unbiased estimate of the average treatment effect at that value of the balancing score if treatment assignment is strongly ignorable. Consequently, with strongly ignorable treatment assignment, pair matching on a balancing score, subclassification on a balancing score and covariance adjustment on a balancing score can all produce unbiased estimates of treatment effects,
5. Using sample estimates of balancing scores can produce sample balance on X.

Mathematical Properties

Rosenbaum and Rubin (1983, p. 43–44) proved the following theorems:

1. The propensity score is a balancing score
2. Any score that is "finer" than the propensity score is a balancing score; moreover, X is the finest balancing score and the propensity score is the coarsest
3. If treatment assignment is strongly ignorable given X, then it is strongly ignorable given any balancing score $b(X)$
4. At any given value of a balancing score, the difference between the treatment and control means is an unbiased estimate of the average treatment effect at that value of the balancing score if treatment assignment is strongly ignorable. Consequently, with strongly ignorable treatment assignment, pair matching on a balancing score, subclassification on a balancing score and covariance adjustment on a balancing score can all produce unbiased estimates of treatment effects,
5. Using sample estimates of balancing scores can produce sample balance on X.

Mathematical Properties

Rosenbaum and Rubin (1983, p. 43–44) proved the following theorems:

1. The propensity score is a balancing score
2. Any score that is "finer" than the propensity score is a balancing score; moreover, X is the finest balancing score and the propensity score is the coarsest
3. If treatment assignment is strongly ignorable given X, then it is strongly ignorable given any balancing score $b(X)$
4. At any given value of a balancing score, the difference between the treatment and control means is an unbiased estimate of the average treatment effect at that value of the balancing score if treatment assignment is strongly ignorable. Consequently, with strongly ignorable treatment assignment, pair matching on a balancing score, subclassification on a balancing score and covariance adjustment on a balancing score can all produce unbiased estimates of treatment effects,
5. Using sample estimates of balancing scores can produce sample balance on X.

Mathematical Properties

Rosenbaum and Rubin (1983, p. 43–44) proved the following theorems:

1. The propensity score is a balancing score
2. Any score that is "finer" than the propensity score is a balancing score; moreover, X is the finest balancing score and the propensity score is the coarsest
3. If treatment assignment is strongly ignorable given X, then it is strongly ignorable given any balancing score $b(X)$
4. At any given value of a balancing score, the difference between the treatment and control means is an unbiased estimate of the average treatment effect at that value of the balancing score if treatment assignment is strongly ignorable. Consequently, with strongly ignorable treatment assignment, pair matching on a balancing score, subclassification on a balancing score and covariance adjustment on a balancing score can all produce unbiased estimates of treatment effects,
5. Using sample estimates of balancing scores can produce sample balance on X.

Mathematical Properties

Rosenbaum and Rubin (1983, p. 43–44) proved the following theorems:

1. The propensity score is a balancing score
2. Any score that is “finer” than the propensity score is a balancing score; moreover, X is the finest balancing score and the propensity score is the coarsest
3. If treatment assignment is strongly ignorable given X, then it is strongly ignorable given any balancing score $b(X)$
4. At any given value of a balancing score, the difference between the treatment and control means is an unbiased estimate of the average treatment effect at that value of the balancing score if treatment assignment is strongly ignorable. Consequently, with strongly ignorable treatment assignment, pair matching on a balancing score, subclassification on a balancing score and covariance adjustment on a balancing score can all produce unbiased estimates of treatment effects,
5. Using sample estimates of balancing scores can produce sample balance on X.

Key Implications

If strong ignorability holds, and treatment and control groups are matched perfectly on their propensity scores, then the difference in means between treatment and control groups is an unbiased estimate of treatment effects.

Moreover, subclassification or covariance adjustment can also yield unbiased treatment effects.

Key Questions

Propensity scores have some nice properties that, in principle, seem to solve a very vexing problem. However, before jumping on the very large propensity score bandwagon, we need to recall

1. The propensity score is a parameter, i.e., a probability. We never know it precisely. We only know sample estimates of it.
2. Propensity scores are guaranteed to yield unbiased causal effects only if strong ignorability holds.

For now, let's move on to a discussion of the practical aspects of calculating and using sample estimates of propensity scores.

Key Questions

Propensity scores have some nice properties that, in principle, seem to solve a very vexing problem. However, before jumping on the very large propensity score bandwagon, we need to recall

1. The propensity score is a parameter, i.e., a probability. We never know it precisely. We only know sample estimates of it.
2. Propensity scores are guaranteed to yield unbiased causal effects only if strong ignorability holds.

For now, let's move on to a discussion of the practical aspects of calculating and using sample estimates of propensity scores.

Introduction

Many approaches to matching are possible, and quite a few are automated in R packages.

A key aspect of all of them is that you never use the outcome variable during matching!

We shall briefly discuss several methods, then illustrate them with a computational example.

MatchIt and Zelig

Gary King and his associates have been actively involved in developing software called *MatchIt* and *Zelig* to facilitate matching and the modeling process.

Zelig subsumes a number of modeling procedures under a common framework, making modeling a more user-friendly exercise. It automates a number of useful methods for analyzing model fit. MatchIt, likewise, automates matching procedures, and provides methods for evaluating their success.

After installing these two packages, you will have a number of matching procedures, and related analytic methods, at your disposal.

Exact Matching

The simplest version of matching is exact.

This technique matches each treated unit to all possible control units with exactly the same values on all the covariates, forming subclasses such that within each subclass all units (treatment and control) have the same covariate values.

Exact matching is implemented in MatchIt using `method = "exact"`.

Subclassification

When there are many covariates (or some covariates can take a large number of values), finding sufficient exact matches will often be impossible.

The goal of subclassification is to form subclasses, such that in each the distribution (rather than the exact values) of covariates for the treated and control groups are as similar as possible.

Nearest Neighbor Matching

Nearest neighbor matching selects the r (default=1) best control matches for each individual in the treatment group (excluding those discarded using the `discard` option). Matching is done using a distance measure specified by the `distance` option (default=logit). Matches are chosen for each treated unit one at a time, with the order specified by the `m.order` command (default=largest to smallest). At each matching step we choose the control unit that is not yet matched but is closest to the treated unit on the distance measure. Nearest neighbor matching is implemented in MatchIt using the `method = "nearest"` option.

Optimal Matching

The default nearest neighbor matching method in MatchIt is "greedy" matching, where the closest control match for each treated unit is chosen one at a time, without trying to minimize a global distance measure. In contrast, "optimal" matching finds the matched samples with the smallest average absolute distance across all the matched pairs.

Overview

Using propensity scores, in practice, involves several steps:

1. Decide on the relevant covariates X
2. Develop a model for predicting $\Pr(T = 1)$ from X
3. Obtain sample propensity scores $\hat{e}(x)$ from the model
4. Use a matching procedure to obtain samples with $T = 1$ and $T = 0$ that are matched on $\hat{e}$.
5. Assess the success of the matching procedure
6. If the matching procedure has not been successful, go back to step 2 and update the model, otherwise proceed
7. Perform the desired parametric analysis on the preprocessed (matched) data

Overview

Using propensity scores, in practice, involves several steps:

1. Decide on the relevant covariates X
2. Develop a model for predicting $\Pr(T = 1)$ from X
3. Obtain sample propensity scores $\hat{e}(x)$ from the model
4. Use a matching procedure to obtain samples with $T = 1$ and $T = 0$ that are matched on $\hat{e}$.
5. Assess the success of the matching procedure
6. If the matching procedure has not been successful, go back to step 2 and update the model, otherwise proceed
7. Perform the desired parametric analysis on the preprocessed (matched) data

Overview

Using propensity scores, in practice, involves several steps:

1. Decide on the relevant covariates X
2. Develop a model for predicting $\Pr(T = 1)$ from X
3. Obtain sample propensity scores $\hat{e}(x)$ from the model
4. Use a matching procedure to obtain samples with $T = 1$ and $T = 0$ that are matched on $\hat{e}$.
5. Assess the success of the matching procedure
6. If the matching procedure has not been successful, go back to step 2 and update the model, otherwise proceed
7. Perform the desired parametric analysis on the preprocessed (matched) data

Overview

Using propensity scores, in practice, involves several steps:

1. Decide on the relevant covariates X
2. Develop a model for predicting $\Pr(T = 1)$ from X
3. Obtain sample propensity scores $\hat{e}(x)$ from the model
4. Use a matching procedure to obtain samples with $T = 1$ and $T = 0$ that are matched on $\hat{e}$.
5. Assess the success of the matching procedure
6. If the matching procedure has not been successful, go back to step 2 and update the model, otherwise proceed
7. Perform the desired parametric analysis on the preprocessed (matched) data

Overview

Using propensity scores, in practice, involves several steps:

1. Decide on the relevant covariates X
2. Develop a model for predicting $\Pr(T = 1)$ from X
3. Obtain sample propensity scores $\hat{e}(x)$ from the model
4. Use a matching procedure to obtain samples with $T = 1$ and $T = 0$ that are matched on $\hat{e}$.
5. Assess the success of the matching procedure
6. If the matching procedure has not been successful, go back to step 2 and update the model, otherwise proceed
7. Perform the desired parametric analysis on the preprocessed (matched) data

Overview

Using propensity scores, in practice, involves several steps:

1. Decide on the relevant covariates X
2. Develop a model for predicting $\Pr(T = 1)$ from X
3. Obtain sample propensity scores $\hat{e}(x)$ from the model
4. Use a matching procedure to obtain samples with $T = 1$ and $T = 0$ that are matched on $\hat{e}$.
5. Assess the success of the matching procedure
6. If the matching procedure has not been successful, go back to step 2 and update the model, otherwise proceed
7. Perform the desired parametric analysis on the preprocessed (matched) data

Overview

Using propensity scores, in practice, involves several steps:

1. Decide on the relevant covariates X
2. Develop a model for predicting $\Pr(T = 1)$ from X
3. Obtain sample propensity scores $\hat{e}(x)$ from the model
4. Use a matching procedure to obtain samples with $T = 1$ and $T = 0$ that are matched on $\hat{e}$.
5. Assess the success of the matching procedure
6. If the matching procedure has not been successful, go back to step 2 and update the model, otherwise proceed
7. Perform the desired parametric analysis on the preprocessed (matched) data

Deciding on Relevant Covariates

All variables in X that would have been included in a parametric model without preprocessing should be included in the matching procedure.

To minimize omitted variable bias, these should “include all variables that affect both the treatment assignment and, controlling for the treatment, the dependent variable.” (Ho, Imai, King, & Stuart,2007, p. 216) Keep in mind that, to avoid posttreatment bias, we should exclude variables affected by the treatment.

Deciding on Relevant Covariates – A Caution

As Ho, Imai, King, and Stuart (2007, p. 216–217) point out, the emphasis in the literature has been to include a virtual grab-bag of all covariates deemed even slightly relevant, and users need to be aware this point of view may be incorrect. The view is that this will decrease bias more than it will increase error variance.

Deciding on Relevant Covariates – A Caution

A Caution

“However, the theoretical literature has focused primarily on the case where the pool of potential control units is considerably larger than the set of treated units. Some researchers seem to have incorrectly generalized this advice to all data sets. If, as is often the case, the pool of potential control units is not much larger than the pool of treated units, then always including all available control variables is bad advice. Instead, the familiar econometric rules apply about the trade-off between the bias of excluding relevant variables and the inefficiency of including irrelevant ones: researchers should not include every pretreatment covariate available.”

The Lalonde Data

Lalonde(1986) constructed an observational study in order to compare it with an actual randomized study that had already been done. A homework question deals with this study in considerable detail.

Gelman & Hill have one set of Lalonde data, while the MatchIt library has another. The MatchIt demos will use their version of the data.