**HOW TO CONDUCT PROPENSITY SCORE MATCHING: AN INTRODUCTION**

by Simon Moss

|  |
| --- |
| **Introduction** |

Many studies are designed to examine the effect of some intervention or treatment. If possible, to explore the impact of this intervention, researchers should conduct a randomized control trial in which participants are randomly allocated to conditions. However, in many circumstances, researchers cannot randomly allocate participants to conditions. In these circumstances, to address this problem, they can utilize a statistical technique called propensity score matching. Propensity score matching is often more effective than alternative techniques such as ANCOVA or multiple regression. The reason is that propensity score matching is effective even if the covariates are not linearly related to the outcome.

|  |
| --- |
| **Randomized control trials versus quasi experimental designs** |

**Randomized control trials**

To appreciate the importance of propensity score matching, you need to understand the distinction between randomized control trials, sometimes called experiments, and quasi experimental designs. To illustrate, suppose you wanted to examine whether caffeine improves marks on exams. If the researcher conducts a randomized control trial

* half the individuals would be randomly assigned to the condition in which they receive a caffeine pill
* the remaining individuals would be randomly assigned to the condition in which they receive a placebo pill instead, perhaps a sugar pill
* all individuals would then complete the exam, and their marks would be recorded

|  |
| --- |
| Receives caffeine then completes exam |
| Image result for silhouette personImage result for silhouette personImage result for silhouette personImage result for silhouette personImage result for silhouette personImage result for silhouette personImage result for silhouette personImage result for silhouette personImage result for silhouette personImage result for silhouette personImage result for silhouette person |
| Receives placebo then completes exam |

* because of this random assignment, the pre-existing characteristics of individuals—such as their age, intelligence, and motivation—should be the same in each condition
* to illustrate, in this example, the number of blue participants is the same in each condition
* consequently, if the individuals who received caffeine outperform the individuals who receive the placebo, we can quite certain this difference can be ascribed to the caffeine rather than other pre-existing differences between the conditions

|  |
| --- |
| Receives caffeine then completes exam: Average 85% |
| Image result for silhouette personImage result for silhouette personImage result for silhouette personImage result for silhouette personImage result for silhouette personImage result for silhouette personImage result for silhouette personImage result for silhouette personImage result for silhouette personImage result for silhouette personImage result for silhouette person |
| Receives placebo then completes exam: Average 73% |

**Quasi-experimental design**

Sometimes, researchers cannot conduct a randomized control trial or experiment. In particular, they cannot as randomly allocate participants to conditions. To illustrate

* perhaps some people are allergic to caffeine—and, therefore, these individuals need to receive the placebo
* or perhaps, rather than ask people to consume either caffeine or a placebo, the researchers simply ask the participants whether they have consumed caffeine—such as coffee—this morning

|  |
| --- |
| Receives caffeine then completes exam: Average 85% |
| Image result for silhouette personImage result for silhouette personImage result for silhouette personImage result for silhouette personImage result for silhouette personImage result for silhouette personImage result for silhouette personImage result for silhouette personImage result for silhouette personImage result for silhouette personImage result for silhouette person |
| Receives placebo then completes exam: Average 73% |

* as this figure shows, even before the study commenced, the participants who received caffeine differed from the participants who received placebos
* for example, the participants who received caffeine might have been older or more intelligent
* their higher score on this exam could thus be ascribed to age or intelligence rather than caffeine

Therefore, as this simple example demonstrates, the results derived from quasi-experimental designs are not as compelling as the results derived from randomized control trials.

|  |
| --- |
| **The rationale of propensity score matching** |

**Calculate probability each person is assigned to the intervention condition**

Propensity score matching is designed to partly correct this limitation of quasi-experimental designs. To conduct propensity score matching, most researchers will utilize statistical software. Nevertheless, to understand the rationale that underpins propensity score matching, you should first observe how the software completes this task. You do not need to understand the rationale yet: this rationale will emerge gradually as you read this document

In particular, the software first calculates the probability each person would be assigned to the intervention condition based on their other characteristics—sometimes called confounds or co-variates. To illustrate

* consider the following figure
* according to this schematic figure, if a black person arrives, this individual is likely to be assigned to the intervention and receive caffeine.
* if a yellow or green arrives, this individual is likely to be assigned to the control and receive the placebo

|  |
| --- |
| Receives caffeine then completes exam: Average 85% |
| Image result for silhouette personImage result for silhouette personImage result for silhouette personImage result for silhouette personImage result for silhouette personImage result for silhouette personImage result for silhouette personImage result for silhouette personImage result for silhouette personImage result for silhouette personImage result for silhouette person |
| Receives placebo then completes exam: Average 73% |

In practice, the difference between the intervention and control on these covariates is not usually as pronounced. Therefore, to decide the probability that each person would have been assigned to the intervention condition, the software

* conducts a technique called a logistic regression analysis
* this technique can generate a formula
* according to this formula, the probability an individual would be assigned to the intervention equals
* the precise formula is not relevant here
* the important issue is this formula can predict the likelihood a person will be assigned to the intervention condition from the other characteristics that were measured, generating this table
* to illustrate, from their age, IQ, and motivation, the formula indicates the probability the first person would be assigned to the intervention condition is .76. Indeed, this person was assigned to the intervention condition.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Person | Age | IQ | Motivation | Probability this person would be assigned to the intervention | Condition this person was actually assigned |
| 1 | 24 | 121 | 5 | .76 | Intervention |
| 2 | 31 | 97 | 8 | .12 | Control |
| 3 | 28 | 99 | 6 | .87 | Intervention |
| 4 | 19 | 104 | 4 | .17 | Control |
| 5 | 45 | 119 | 9 | .36 | Intervention |
| 6 | 30 | 108 | 3 | .36 | Control |

**Match the conditions**

So, how does the software utilize these probabilities—called propensity scores? To answer this question, consider the following table. In this table, the first column represents the propensity scores of individuals in the intervention group. The second column represents the propensity scores of individuals in the control group.

|  |  |
| --- | --- |
| Intervention condition: Received caffeine | Control condition: Received placebo |
| .31 | .36 |
| .84 | .82 |
| .43 | .45 |
| .59 | .55 |
| .15 | .17 |
| .65 | .63 |

As this table shows

* in each row, the propensity scores are similar
* in other words, propensity scores in the intervention condition match the propensity scores in the control condition
* therefore, a person with a specific age, IQ, and motivation should be equally likely to be assigned to the intervention condition or to the control condition
* consequently, when propensity scores are matched, the design slightly resembles a randomized control trial—because, in randomized control trials, each person is equally likely to be assigned to the intervention condition or to the control condition.

In short, if the propensity scores in the intervention condition match the propensity scores in the control condition, the design theoretically resembles a randomized control trial. Differences between the conditions can be more confidently ascribed to caffeine rather than to the other measured co-variates, such as age, IQ, and motivation.

**How to match the conditions**

But, how does the software match the conditions. The software can apply one of several principles

* First, the software could select only a subset of participants
* For example, the software might choose a participant from the intervention group with a propensity score of .50. The software would then identify a participant from the control group with a similar propensity score.
* The software would then continue this procedure until the maximum number of matched pairs is identified.

This method is straightforward in principle, but not especially helpful in practice. The problem is this method will exclude many participants, diminishing the likelihood of significant results. Instead, software can utilize other procedures instead, such as interval matching, Mahalanobis metric matching, and weighting. In principle, these methods are similar, but utilize more complex algorithms to increase statistical power.

|  |
| --- |
| **Software to conduct propensity score matching** |

So far, this document has merely introduced the rationale that underpins propensity score matching. To actually conduct this technique, you need to use the relevant software.

**R**

The free statistical package R can be used to conduct propensity score matching. If you have yet to download R, you should

* Go to <https://cran.rstudio.com>
* Click the “Download R” option that is relevant to your computer—such as Mac or Windows
* Click the options to download the latest version of R
* Install R on your computer by following the instructions—just like you would install any software

Then, you should

* Go to <https://www.rstudio.com>
* Click Download R studio
* Click the RStudio option associated with your computer under the heading “Installers”
* Install this software and follow the various prompts
* R Studio is designed to simplify the use of R

Finally, Google the web for instructions on how to utilize R to conduct propensity score matching.

**Stata**

If you use Stata, you can utilize the psmatch2 command. Specifically

* the command resembles psmatch2 condition, x1, x2, out(y).
* In this example, condition is a column that represents the condition in which participants were assigned
* x1 and x2 are covariates, such as age and IQ
* y is the outcome, such as exam performance

Alternatively, you can utilize the teffects command. In particular

* the command resembles teffects psmatch y (condition x1 x2, probit), atet
* condition, x1, x2, and y are defined as in the previous example
* the subcommand *atet* calculates the average treatment effect in the treated individuals—that is, the impact of this intervention on participants in the intervention condition
* if you add “, nn(2)”, the software matches each participant in the intervention condition with two participants in the control condition, where nn represents nearest neighbour.

**SPSS**

You can conduct propensity score matching in SPSS. However, before you start, you need to download R and an accompanying package, both of which are free. The following table illustrates how you can achieve this goal and then conduct propensity score matching

|  |  |
| --- | --- |
| Activity | Details |
| Install R on to your computer | First, decide which version of R to download   * For SPSS Version 21, install R.2.24.0 * For SPSS Version 22, install R.2.25.0 * For later versions of SPSS, usually you should install the latest versions of R   To install R   * Go to https://cran.rstudio.org * Click the “Download R” option that is relevant to your computer—such as the Mac or Windows version * Then, click the options to download R * Install R on your computer by following the instructions—just like you would install any software. * Usually, you can just press Next after each instruction. * If you plan to use R in the future, go to https://www.rstudio.com * Click Download R studio * Then click the RStudio option associated with your computer under the heading “Installers” * Install this software and follow the various prompts * R Studio is designed to simplify the use of R * To download R, visit cran.r-project.org/bin/ * Then press macosx or windows depending on whether you want to install on a Mac or PC respectively * Press install R for the first time and then Download… |
| Download IBM SPSS Statistics-Essentials for R | * Google “IBM Tools for SPSS products” * Then click the link that resembles IBM Tools for SPSS Products 2019/01/16 17:26:18 * Press *register here* to register an account |

|  |
| --- |
| **Some output to interpret** |

The output that you generate will depend on which statistical package you utilize. This section, however, outlines some principles you should consider to interpret the output.

**Measures of balance**

* Most packages will present some statistics that indicate the extent to which the intervention and control conditions were matched successfully—that is, the degree to which the conditions are similar on the co-variates.
* One example is the reduction in pseudo R2 from the logistic model. A high value indicates the conditions are significantly more similar to each other after the propensity scores were matched. Likewise, if the pseudo R2 from the logistic model is close to zero, the matching had been successful

**Differences between the conditions**

* Most statistical packages will generate a p value associated with the differences between conditions after propensity score matching. If this p value is less than .05, the differences between conditions is significant

|  |
| --- |
| **Further reading** |

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behavioral Research, 46, 399-424.

Austin, P. C. (2014) A comparison of 12 algorithms for matching on the propensity score. Statical Medicine, 33, 1057-1069.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. Journal of Economic Surveys, 22, 31-72.

Winger, D. G., & Nason, K. S. (2016). Propensity-score analysis in thoracic surgery: when, why, and an introduction to how. Journal of Thoracic Cardiovascular Surgery, 151, 1484–1487.