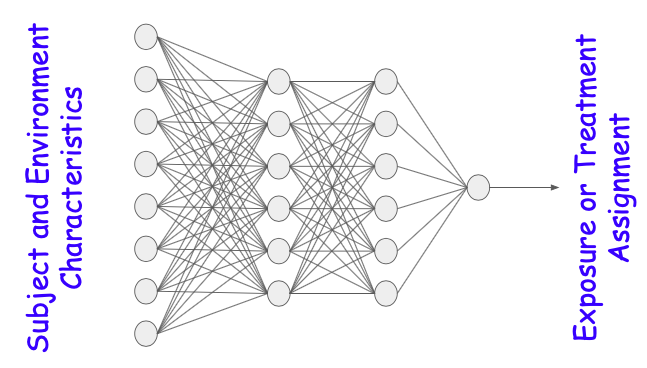
Propensity Score-Based Methods for Causal Inference

Module 5: Modeling the Assignment Mechanism



**I. Module Objectives**

Once we have defined the research question and associated causal effect, described the strengths and limitations of the available data, and drawn a causal graph to identify the important relationships and variables to include the propensity score, we need to estimate each subject propensity for treatment or exposure, i.e. the probabilities associated with each possible treatment or level of exposure. For this module, we will assume that the resulting assignment is binary (e.g. experimental versus standard treatment, or exposed versus not exposed).

Ideally, the assignment mechanism is random to guarantee that, at least on average, the groups are comparable and unaffected by confounding. However, in many cases, randomization may be impractical or unethical. Randomized trials may also be difficult to implement in generalizable populations. All of these factors motivate the need for observational studies. In observational studies, however, other processes or “sorting” to treatment or exposure levels may substantially bias results with the treated versus untreated, or exposed versus unexposed groups having very different characteristics.

The next step of propensity score-based methods, after formulating the causal question and deciding on which variables should be included in the propensity score, is to model the assignment mechanism. This is typically accomplished with a logistic model, but other methods (e.g. machine learning models) have substantial appeal in this setting because the focus is strictly on prediction rather than interpretation.

***For this module, the methods of interest are essentially the same as we would use for prediction of some outcome measure. Therefore, the focus of any “new information” for this module is more of a matter of perspective***, i.e. predicting an exposure or use of an intervention or treatment, versus not exposed or not using the intervention or treatment.

Further, our perspective in using propensity scores is to focus solely on prediction rather than interpretation or model building (with any issues specific to variable specification discussed in the previous module). We are not concerned with interpreting coefficients or variable significance. Rather, the model is specified to include potential confounders and predictors of the outcomes, and our single quantity of interest is the predicted probability of exposure for each subject. That predicted probability (or proportion or other measure of exposure dose in more general cases) will then be used in subsequent steps to create a pseudo population (as described in the next module). We refer to this current step as modeling the assignment mechanism.

For simplicity, we will focus on prediction of a binary exposure (e.g. smoking or not smoking, exposed to radiation or not exposed).

By the end of this module, you will be able to:

1. Describe the strengths and limitations of using standard methods, such as logistic regression, for the propensity score model
2. Describe the strengths and limitations of using machine learning methods, such as neural networks, for the propensity score model

**II. Module Assignments**

**Optional Assignment to gain further prerequisite knowledge (before starting the Required Assignment):**

There are many available resources for describing regression approaches for outcome prediction in general, and, more specifically, prediction of binary outcomes. For those without the prerequisite course in fundamental biostatistical methods, you may want to take either a for-credit course, or an online course, or study a textbook, such as *Applied Logistic Regression* (Hosmer and Lemeshow, 2000).

**Required Assignments: (~18 pages to read)**

For an overview of some different methods that might be considered for generating propensity scores, read the following article: Westreich, D., Lessler, J. and Funk, M.J., 2010. Propensity score estimation: machine learning and classification methods as alternatives to logistic regression. *Journal of clinical epidemiology*, *63*(8), p.826-833.

For additional discussion of different modeling mechanisms, and simulations for how they impact propensity score weighting and the final effect estimate, read the following article: Lee, B.K., Lessler, J. and Stuart, E.A., 2010. Improving propensity score weighting using machine learning. *Statistics in medicine*, *29*(3), pp.337-346. Although propensity score weighting and effect estimation using the outcomes model are not covered until the next two modules, the article is still useful to read now since it focuses on the assignment mechanism model. You might also want to re-read the article after finishing the next two modules.

**Optional Assignments:(~1 hour 17 minutes)**

For a presentation of results from the Lee, et al. paper, and related issues, watch this video from Elizabeth Stuart: <http://cepim.northwestern.edu/calendar-events/2015-12-01>. The slides for the presentation are in a separate window on the same webpage.

**III. Project Exercises**

Create a copy of this Google Doc or download the Module onto your computer and review the material offered above under Module Assignments before beginning these workbook exercises.

Thinking about what you learned in this module so far, begin developing the analysis plan for your project by answering the following questions:

1. For your binary exposure of interest, first consider using logistic regression and the covariates (that included potential confounders and outcome predictors) identified in the last module to model the exposure assignment mechanism. More specifically, answer the following questions:
   1. Based on your understanding of the subject matter and factors that drive exposure (or treatment or intervention status), will using only the main effects sufficiently capture the predicted exposure for a given subject?
   2. If not, are there specific interactions and/or nonlinear terms that you could specify to add to the logistic model?

* 1. After doing so, do you feel the model from step b will accurately predict the exposure for a given subject? Why or why not?

1. Now reconsider using neural networks for the assignment mechanism model; see the assigned reading by Westreich, et al. if you are not familiar with that method.

The advantage to using neural networks is that they inherently fit interactions and non-linear terms to predict (in this case) exposure. However, a disadvantage is that they are essentially a black box technique where the ‘user’ cannot specify the nature of that relationship. Considering that, answer the following questions:

1. Does your subject matter knowledge of the problem suggest that there are more complicated relationships in predicting exposure than we can explicitly specify with logistic regression?

This might, for instance, be true when considering how a range of different personal characteristics, insurance coverage and health system factors, physician preferences, geographic or regional differences, and various social factors dictate an individual’s propensity for receiving one type of surgery versus another.

If you believe this scenario is true for your research question and exposure of interest, describe (specific to your problem of interest) how the use of neural networks may benefit the ability to predict an individual’s propensity for being exposed. If not, go to part b of this question.

1. In contrast, does your subject matter knowledge of the problem suggest there is a clear relationship you can (at least approximately) specify through a logistic model?

This might, for instance, be true when comparing patients who elect to have surgery versus multidisciplinary care for knee injuries. In that case, we might be able to say that there are specific issues regarding insurance coverage, the level of pain they are experiencing, their age, and their prior history of surgery that could be used to predict propensity for surgery with an explicitly specified logistic model.

If you believe this scenario is true for your research question and exposure of interest, describe (specific to your problem of interest) how the use of neural networks may detract from your ability to predict an individual’s propensity for being exposed.

1. Now reconsider using classification trees for the assignment mechanism model; see the assigned reading by Westreich, et al. if you are not familiar with that method.

The advantage to using classification trees is that they recursively split the data into different subgroups and thus inherently fit interactions and non-linear terms to predict (in this case) exposure. However, they model the relationship through repeated binary splits, rather than weighted sums of the data, which can either be an advantage or a disadvantage depending on whether your exposure of interest is associated with a discrete number of high risk or low risk groups. Considering that, answer the following questions:

1. Does your subject matter knowledge of the problem suggest that repeated splitting the data (on confounders and outcome predictors) to predict exposure will be more effective than explicitly specifying linear relationships (in a logit scale) with logistic regression?
2. In contrast, does your subject matter knowledge of the problem suggest there is more of a continuous relationship between the covariates and exposure that could be better modeled with logistic regression or neural networks?

If you believe this scenario is true for your research question and exposure of interest, describe (specific to your problem of interest) how the use of classification trees may detract from your ability to predict an individual’s propensity for being exposed.

[Link to go back to the Course Overview Document](https://docs.google.com/document/d/1UDTkp3rbhqdun7jvSvktaZmTtoUWOz_VUDQw3HIsElg/edit?usp=sharing)