SOME METHODS FOR CAUSAL INFERENCE, WITH APPLICATION TO AN OBSERVATIONAL EPILEPSY DATA SET

by

Mengyao Wang

Submitted in partial fulfillment of the requirements
for the degree of Master of Science

at

Dalhousie University
Halifax, Nova Scotia
May 2021

Contents

List of Tables

List of Figures

Abstract

Causal inference attempts to attribute a causal mechanism to a treatment in an observational study. Attributing cause is a major focus of research in bio-statistics and application to observational biomedical studies. There have been a number of different proposals as to how causal inference should be carried out. In this thesis, two methods - propensity score adjustment and graphical causal modeling – are explored through application to an observational data set.

The data set concerns several hundred pediatric patients with epilepsy collected over thirty plus years at the IWK hospital. The primary goal of the thesis is to identify which factors are important in determining whether patients remain on anti-epileptic medication at the end of follow up. The basic non-causal model – logistic regression, with or without stepwise selection - identifies a number of significant predictors in addition to the nominal treatment variable, which is the indicator of normal neurological status.

In an observational study the primary difficulty in attributing a causal mechanism to the treatment is that treatment groups are typically unbalanced with respect to possible confounders. One approach is to rectify the imbalance through the use of so-called balancing scores. The most generally used balancing score is the propensity score. Matched estimates of the treatment effect are obtained after matching treatment and control groups with respect to the propensity score.

More recently, directed acyclic graphical models (DAGs) have been proposed as a means to carry out causal inference. Among other things, the DAG provides information on which variables should not be used in the conditioning set. Several methods have been proposed to estimate the DAG, or an equivalence class of DAGs, underlying an observational data set. Some basic methods for estimating the DAG are introduced, and the "backdoor criterion" - a sufficient condition for identifying confounders which should not be included in the model – is applied. The resulting model is dramatically simplified as compared to the model selected using backwards elimination but has some commonality with a model estimated using propensity score matching.

Acknowledgements

I would like to thank my supervisor Dr.Smith for his consistent support with patience.

I would like to thank Camfield's for letting me use their data which they spent decades to collect.

Finally yet importantly, I would like to thank my family and friends for their encouragements and enthusiastic support.

Chapter 1

Introduction

1.1 Background

What is causation and why is it important in statistics? The essence of causal analysis is in identifying causal mechanisms and estimating the magnitude of causal effects. We would like to answer questions such as "Why are males better at engineering than females ?" "Is smoking the major cause of lung cancer?" "How effective is a given treatment in preventing disease?" These types of questions all focus on causal mechanisms.

One definition of a causal relationship is "a relationship between two events where one event is affected by another", in words, causation indicates that an event affects an outcome. We desire to understand the mechanism leading to the outcome. In this thesis we will explore several methods to assess causation.

It is often difficult to distinguish between correlation and causation. An association or correlation among variables simply indicates that their values shift together. It does not necessarily suggest that changes in one variable cause changes in the other variable. The expression often used is "correlation does not imply causation." On the other hand a causal relationship will lead to correlation.

Why is estimating a causal effect so important? Is correlation not good enough to identify relationships? Correlation may describe some of the relationship between variables. However, there are situations when we want to intentionally affect an outcome variables by changing a treatment variable. For example if studying does not cause an increase in test scores, then there's no point in studying. If a medicine does not cause an improvement in your health or ward off disease, there is no reason to take it. Therefore, understanding causal mechanisms is quite important, and correlations are insufficient to identify causal relationships.

Causal analysis plays a prominent role in a variety of fields, and particularly important in biostatistics, epidemiology, and medical research.

A concept that needs to be introduced is that of a **causal model**. It is a visualized graphical way to illustrate the causal relationship among a set of variables. It uses arrows and nodes to show how variables affect each other.

In next chapter, we will start with some new terminologies and concepts associated with causal effects.

1.2 Data set

The data set being used in the thesis concerns epilepsy in pediatric patients. The goal of the study is to develop causal models whether or not subjects followed and treated for epilepsy will be in remission at the end of followup, where remission is defined as not being on anti-epileptic medication. More specifically, we would like to determine what factors determine the status of taking anti-epileptic drugs, and if

one can control the outcome by altering related covariates. The unfiltered original data set contains 463 variables with 554 subjects. The predictor variables cover wide range of types, including binary, nominal, and ordinal variables.

A subset of the dataset consisting of 16 variables was analysed in Camfield et al. (2016). We further reduced that dataset by removing 3 more variables containing a substantial number of missing values, so our final analysis was based on 13 variables of primary interest.

The data set was collected over a number of years, and is observational in nature, with no well defined treatment or outcome variable. We have focused attention on a particular outcome (remission at end of followup), and have identified neurological status (normal or abnormal) as a treatment variable.

In a randomized controlled experiment the treatment would be randomized to subjects, after which the outcome is observed. In this case the effect of treatment could be estimated as the difference in proportion of subjects in remission between treatment (neurological status=normal) and control (neurological status=abnormal) groups. The most important distinction between a controlled, randomized trail and an observational study is that the treatment is not randomized in an observational study. The advantage of a randomized, controlled study is that the process of randomization balances the treatment and control groups on average with respect to all other covariates, and in this way leads to an estimated relationship between treatment and outcome which is causal. In an observational study the treatment is observed,

not randomized, typically with many other covariates. The association between treatment and outcome may be due to the treatment causing the outcome, or it may be due to some other covariates which influence both treatment and outcome. For example, a particular genetic cause for the epilepsy might lead to an abnromal neurological status. The epilepsy might be successfully treated with anti-epileptic medicine, but never cured as it is genetic in origin, so a patient carrying the disease causing gene will typically never be in remission. In this case the association between treatment and outcome is not due to direct causation.

Camfield et al. (2016) used logistic regression and ordinal logistic regression to model a number of epilepsy outcomes, with an emphasis on social outcomes. Without further adjustment, logistic regression does not estimate a causal model. The goal of the present study is compare causal and non-causal estimates by suitably adjusting the logistic regression model.

Predictor Variables		**Variable Name in data set**	**Description**
Surgery		surgery	have had surgery for epilepsy
Intelligence Level		intellnormal	normal intelligence level: 1/0
Neurology Level		neuronormal	normal neurological status: 1/0
Number of AEDs		naed	number of antiepileptic drugs
Age at first Seizure		agefirst	age at first seizure
Sex		sex	patient gender
Poor Level		poor	indicator of low income:1/0
Adequate Income		adequate	indicator of adequate income
Generalized tonic Seizure		nevergtc	never generalized tonic seizures
Secondary Seizure		neversec	never secondary seizures
Seizure Type	focal	focal	Indicator: 1/0
	generalized	generalized	
	symptomatic	symptomatic	

Table 1.1: Data Set Variable List

The outcome variable of interest is the binary indicator of whether or not a subject is in remission at the end of followup, which takes the value 1 if subject was on anti-epileptics drugs at end of followup, and 0 if the subject was not on anti-epileptic drugs at the end of followup. Table 1.1 shows the predictor variables involved in the study. The data consist of the triples $(Y_i, Z_i, X_i), i = 1, ..., n$. We assume the outcome variable Y and the treatment assigned Z are random, while X is fixed covariate. While Z is also a covariate, it distinguished from X in that Z represents the treatment variable of interest, while X denotes a vector of covariates which will be used to adjust the relationship between outcome Y and treatment Z. Let Y_{1i} and Y_{0i} be the responses of subject i after having received the treatment and control, respectively. The causal effect of the treatment on subject i is $Y_{1i} - Y_{0i}$, which will typically depend on the covariates X_i. Of course, the causal effect on subject i cannot be estimated, as only one of Y_{1i} or Y_{0i} is observed for subject i, depending on whether or not $Z_i = 1$ or $Z_i = 0$. The two outcomes are referred to in the causal analysis literature as *counterfactual* outcomes. The average causal effect can be estimated under appropriate assumptions.

There are several issues involved in causal inference that are not encountered in non-causal analysis. Most importantly, what distinguishes a causal from a non-causal effect, and how can the causal effect be estimated. Causal graphs are central to much of modern causal inference.

1.3 Thesis Structure

In Chapter 2, we review the use of logistic regression analysis for estimating non-causal effects. In Chapter 3 we introduce an algorithm called propensity score matching that has been use for causal inference, and apply the method to the seizure data. Chapter 4 provides an introduction to graphical methods of causal inference, including some important terminology, rules and algorithms, and these ideas are applied to the neurological data. Chapter 5 summarizes and generalizes some results, and makes suggestions for further work.

Chapter 2

Non-causal modeling using logistic regression

Logistic regression(Wright (1995)) is an appropriate method when people have large data set by hand, the dependent variable is dichotomous, and there are one or more nominal or interval level predictor variables. The types of questions that logistic regression can answer is something like how does the probability of getting lung cancer (yes vs. no) change for every additional pound a person is overweight and for every pack of cigarettes smoked per day? Do body weight and age have an influence on the probability of having a heart attack? Logistic regression seems a reasonable starting point to examine the effects of predictor variables on a binary dependent variable.

Logistic regression has two major assumptions: (1) the outcome variable should be dichotomous in nature and (2) there should ideally be no high correlations among the predictor variables. The second assumption is quite influential and this restriction limits the usefulness of the results in assessing causal structure. Table 2.1 gives the logistic regression output for the epilepsy data set.

The logistic regression model is

$$log(p/(1-p)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$$

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.9586167  0.8412147   3.517 0.000436 ***
surgeryno    -0.9992324  0.4916976  -2.032 0.042132 *
intellnormal -1.1871888  0.2409249  -4.928 8.32e-07 ***
neuronormal  -0.8809114  0.2525424  -3.488 0.000486 ***
nevergtc     -0.9572158  0.4806967  -1.991 0.046447 *
neversec     -1.4069285  0.3183354  -4.420 9.89e-06 ***
agefirst      0.0096086  0.0021008   4.574 4.79e-06 ***
followup     -0.0006243  0.0010073  -0.620 0.535395
sexmale      -0.0260248  0.2056637  -0.127 0.899304
focal        -0.7274160  0.6478507  -1.123 0.261516
generalized   0.2588468  0.5702930   0.454 0.649912
symptomatic   0.8360784  0.5933779   1.409 0.158831
poorinc      -0.5259572  0.2605484  -2.019 0.043523 *
adeqincome   -0.2752386  0.2530388  -1.088 0.276713
```

Table 2.1: Logistic regression output using remission as the outcome variable

The estimated coefficients are estimates of the β_j's, can be used to estimate $log(p/(1-p))$, and by transforming, to estimate the probability of success (remission), $p = P(aedend = 1)$.

These estimates tell the amount of increase in the predicted log of odds for a unit increase in the predictors. They are often difficult to interpret, so they are often converted into odds ratios. In this thesis we are primarily interest in knowing which covariates should be included in the model, and so we just work on the logit scale.

Table 2.1 lists all predictor variables with their estimated coefficients. We can observe that *intellnormal*, *neuronormal*, *neversec* and *agefirst* are all very significant, but the model also includes a number of non-significant variables. For example, the effect of *neuronormal*, measured on the logit scale, is -.8809, and the effect is judged to be highly significant (p=.000486).

As it is used here the logistic model is not a causal model. The p-values compare

two models - a full model with all predictors included, and a reduced model with all but the predictor of interest included. As such, the logistic regression estimates the importance of each variable when entered last in the regression equation.

We would like to reduce the model by removing predictor variables which are not important. There are a variety of methods to do this. As will be seen in chapter 4, causal graphical models provide one way to determine which predictors should be kept. Here I have used *backwards stepwise elimination* beginning with the full model in table 2.1 to do model selection. Table 2.2 shows the estimated effects after using backwards elimination for model reduction.

Comparing table 2.1 and table 2.2, the estimated coefficients in the reduced model are similar to those from the full model. *intellnormal*, *neuronormal*, *neversec* and *neversec* are all still highly significant, as is now **focal**.

Model reduction shows similar results of which variables we should include when doing analysis.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.041258   0.635409   4.786 1.70e-06 ***
surgeryno    -0.881596   0.487142  -1.810 0.070338 .
intellnormal -1.230574   0.233241  -5.276 1.32e-07 ***
neuronormal  -0.993226   0.242021  -4.104 4.06e-05 ***
nevergtc     -0.643573   0.413184  -1.558 0.119330
neversec     -1.364748   0.316265  -4.315 1.59e-05 ***
agefirst      0.009167   0.002021   4.536 5.73e-06 ***
focal        -1.406823   0.394942  -3.562 0.000368 ***
poorinc      -0.434769   0.228927  -1.899 0.057543 .
```

Table 2.2: Estimated coefficients after stepwise regression using backwards elimination

Backwards stepwise regression(Austin (2008)), also known as *backwards elimination regression*, is a stepwise regression approach that begins with a full model and at

each step gradually eliminates variables from the regression model to find a reduced model that best explains the data. The *stepwise* procedure in R uses the Akaike Information Criterion (AIC) to assess the fit, removing at each step that variable whose removal least affects the model fit, until no further variables can be removed. AIC gives an estimate of the quality of each model relative to other models, with smaller values of AIC indicating better model fit.

```
Step:  AIC=607.7
aedend ~ surgery + intellnormal + neuronormal + nevergtc + neversec +
    agefirst + focal + symptomatic + poorinc

               Df Deviance    AIC
<none>              587.70 607.70
- symptomatic   1   590.18 608.18
- focal         1   590.73 608.73
- poorinc       1   590.90 608.90
- surgery       1   591.15 609.15
- nevergtc      1   591.99 609.99
- neuronormal   1   600.50 618.50
- neversec      1   609.79 627.79
- agefirst      1   611.42 629.42
- intellnormal  1   613.56 631.56
```

Table 2.3: Importance of variables after backwards elimination

The AIC for the unreduced model was 613.56. After stepwise elimination, the model indicated in table 2.2 has AIC=607.7. Table 2.3 shows the AIC for the collection of models found by removing a single predictor from the model in table 2.2 According to the AIC criterion, no additional variable should be removed. Furthermore, the AIC values show the relative importance of the retained values. *Intellnormal* is the most important predictor in the reduced model, as its removal increases AIC the most. The ordering of importance based on AIC in table 2.3 is similar to the ordering using p-values in Table 2.2.

Note that reduced model still contains non-significant variables based on z-tests.

For example, *nevergtc* (p=.119), *surgeryno* (p=.070) and *poorinc* (p=.057).

I will show in chapter 4 how graphical models provide an alternative method to do model reduction. In chapter 3, I will show how matched logistic regression can remove the need to include all but the treatment variable in the logistic regression.

The marginal model considers only the treatment (here *neuronormal*) and outcome variables. The logistic regression output for this model is included in figure 2.4. The estimated effect size (-1.39) is larger in absolute value than for the full and reduced models considered in tables 2.1 and 2.2, which suggests that part of the marginal effect size is accounted for by correlation with other predictors which have been included in the larger models. This would not the the case if the design was orthogonal, that is, when the predictors are uncorrelated.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.5002     0.1679   2.980  0.00288 **
neuronormal  -1.3886     0.2009  -6.913 4.75e-12 ***
```

Table 2.4: Logistic regression estimates for marginal model using only “neuronormal”

Chapter 3

Causal modeling using propensity score matching

The concept of propensity score matching (PSM) was first introduced by Rosenbaum and Rubin (1983) in a paper entitled "The Central Role of the Propensity Score in Observational Studies for Casual Effects", and it has become very popular recent years. Initially propensity score methods were developed because researchers in many fields were only able to do observational studies. Random samples of subjects would be drawn and a number of variables of interest measured. Investigators had no control of the various independent variables and covariates, and subjects with different covariate profiles were unlikely to have equal probabilities of receiving the treatment. It is very likely that there are large differences on corresponding covariates between control and treatment group that may cause inaccurate and misleading results in assessing the effect of treatment on outcome.

If we want to correctly estimate treatment effects on a binary outcome in observational studies, how do we control the assignment to treatment group and control group? How do we alter other covariates to keep them constant so that we can simply compare outcomes among units that received the treatment versus those that did not. Some details and implementations will be addressed in this chapter.

3.1 Implementation

In an observational study, with outcome Y and binary treatment variable Z (1 for treatment, 0 for control), and covariates X, the propensity score, here denoted as $\phi(x)$, is defined as

$$\phi(x) = P(Z = 1|X = x)$$

This was first defined by Rosenbaum and Rubin (1983) as the conditional probability that the subject receives the treatment given the covariates. The average treatment effects (ATE) is the difference between the outcomes of treated and control observations. Where Y_1 is a subject's response under the treatment ($Z = 1$), Y_0 is the response under the the control ($Z = 0$), the treatment effect is $\Delta = Y_1 - Y_2$, and the average treatment effect for subjects with covariate pattern X is

$$ATE(X) = E(\Delta) = E[Y_1|X, Z = 1] - E[Y_0|X, Z = 0]$$

If the outcome variable Y is binary (which is true for the present example where Y denotes the indicator that the subject is in remission at the end of followup), then $E(Y) = 0 \times P(Y = 0) + 1 \times P(Y = 1) = P(Y = 1)$, and the the average treatment effect is the difference of the probabilities of remission ($Y = 1$) in the treatment ($Z = 1$) and control ($Z = 0$) populations.

In a randomized experiment, the average treatment effect can be estimated because the covariates of the experimental units in the two populations corresponding to $Z = 1$ and $Z = 0$ are expected to be similar, on average, due to random allocation of treatment to experimental units. For example say a researcher wants to test the

effect of a drug on lab rats. He randomly divides the rats in two groups and tests the effect of the drug in the treatment group and in the control group. As it is an experiment, everything is controlled by the experimenter. For example, all of the rats can be genetically the same and grown in the same environment, so she knows that any differences on the outcome between the groups will be due to the drug that she has given them. However this cannot be done with observational study where the experimental units in treatment and control groups are highly likely differ due to the distributions of covariates in the treatment and control groups.

Dawid (1979) developed the notion of a balancing score $b(x)$, which is a function of the observed covariates, and for which the conditional distribution of covariates X, given $b(x)$, is the same for treated and control units. David (1979) used the following notation for this:

$$x \perp z|b(x)$$

It is clear that x itself is a balancing score, as given x, the distribution of x is fixed for both treatment and control units, that is, conditional on $b(x) = x$, the distribution of X is independent of z. Rosenbaum and Rubin(1983) prove that the propensity score $\phi(x) = P(Z = 1|x)$ is a balancing score under the assumption that treatment assignment is *strongly ignorable*, and furthermore, that any other balancing score must be a function of the propensity score. In a randomized experiment it is assumed that X includes all covariates used to assign treatments, and possible related to the response (Y_1, Y_0). Formally, in a randomized trial, the treatment assignment and

responses are assumed to be conditionally independent given $X = x$, that is

$$(Y_1, Y_0) \perp Z|x$$

where

$$0 < P(Z = 1|X = x) < 1$$

for all x

When treatment assignment is strongly ignorable given covariates x, we will just say it is strongly ignorable. In plain words, strong ignorability given x means that (a) all the possible confounding phenomena (in the sense that they influence both Y and Z) are measured in X, so that conditioning on X removes the direct dependence between Y and Z. (b) there is a nonzero probability of a unit with covariates X receiving either treatment. If a treatment assignment is strongly ignorable given x, then it is strongly ignorable given any balancing score $b(x)$; thus, in particular, the treatment assignment is strongly ignorable given $\phi(x)$.

Combining these ideas, gives a justification for the idea of propensity score matching. For a given propensity score $\phi(x)$, suppose that we randomly sample two units from the entire population, one of which is a treatment unit and the other which is a control unit. This is called a matched pair. Strongly ignorable treatment assignment implies that

$$E[Y_1|\phi(x), z = 1] - E[Y_0|\phi(x), z = 0] = E[Y_1|\phi(x)] - E[Y_0|\phi(x)] = E[Y_1 - Y_0|\phi(x)]$$

Then by the law of iterated expectations,

$$E_{\phi(x)}E[Y_1|\phi(x), z = 1] - E_{\phi(x)}[Y_0|\phi(x), z = 0] = E_{\phi(x)}E[Y_1 - Y_0|\phi(x)] = E[Y_1 - Y_0]$$

Hence, the mean of the differences between the effects in matched pairs is an unbiased estimator for the treatment effect. This is also one of the properties of a balancing score derived by Rosenbaum and Rubin (1983). A consequence is that pair matching on a balancing score or covariate adjustment using a balancing score will produce unbiased estimates of the treatment effect, and also that using sample estimates of balancing scores will generally lead to sample balance on x.

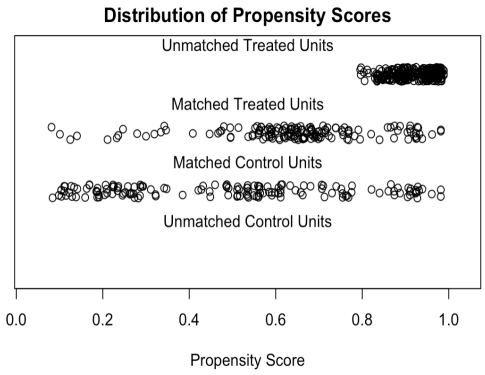
Rosenbaum and Ruben(1983) advocated for the use of the propensity score, but actually there are four different propensity score methods are used for removing the effects of confounding when estimating the effects of treatment on outcomes: propensity score matching, stratification (or sub-classification) on the propensity score, inverse probability of treatment weighting (IPTW) using the propensity score, and covariate adjustment using the propensity score. Rosenbaum and Ruben(1983) did not specify one particular way in which it was to be used. This indeterminacy continues to exist. D'Agostino (1998) provides an excellent review of the use of propensity scores in bio-statistics, but gives no guidelines for which adjustment to use. The focus here is to choose well-matched samples of the original treatment and control groups, thereby reducing bias due to the covariates. I will illustrate the use of the propensity score using matching only. One issue which arises is that most of the methods will drop the unmatched observations in the analysis, thereby disregarding the units that can not get matched, and losing some information on observational data. There is a *full matching* method that includes all units after matching. A benefit of full matching is that no units are discarded, which has the potential to improve precision and prevent

bias due to incomplete matching, but full matching applies weights to observations, which can be a disadvantage. Another method , *nearest neighbor matching*, has the disadvantage of disregarding unmatched treatment or control observations.

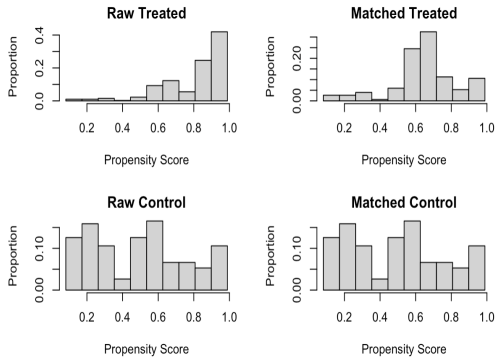
3.2 Propensity score matching using the *MatchIt* package

We use *"MatchIt"* package in R to do the match according to the propensity score. There are few matching method contain within MatchIt package - "exact", "genetic", "nearest", "full", "optimal" and "subclass", with the different methods designed for dealing with different data sets. For example exact matching is better off to implement with stratified data, and genetic matching is less specific in specifying the distance measure use to do the matching. The most commonly used matching method is nearest neighbor matching, also called "greedy matching". However, most of the methods will drop the unmatched observations when carrying out the analysis. I use both nearest matching and full matching to choose the data for analysis, and then used logistic regression with single predictor neuronormal to estimate the effect size on the logit scale.

Nearest matching is also known as "greedy matching". It runs through the list of treated units and selects the nearest eligible control unit that can be pair matched with each treated unit. It uses a distance measure to define which control unit is the nearest to the treated unit. The most commonly used distance is the absolute value of the propensity score difference between each treated and control units. Another popular distance method uses "Mahalanobis distance matching". It tends to work



(a) Jitter plot for nearest neighbor matching



(b) Histogram of matched observations

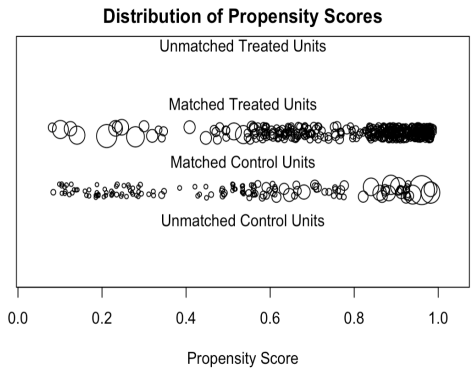
Figure 3.1: Logistic regression using nearest neighbor matched data

better with continuous covariates than with categorical covariates. In our case, we will stay with the absolute propensity score distance.

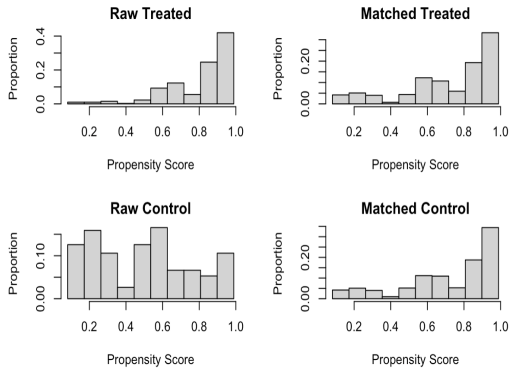
Full matching assigns every treated and control unit in the sample to one subclass each. It makes use of all individuals in the data set. Each subclass either contains one treated individual and multiple control individuals, or one control units and multiple treated individuals. The benefit of full matching is that no units are disregarded, which gives the potential to improve precision and prevent bias due to incomplete samples. It also has been shown to be particularly effective at reducing bias due to observed confounding variables. However, the effective sample size may be different from the original sample size, as the former is determined by matching weights. One cannot conclude that full matching offers more precise results than other matching methods.

Figure 3.1 is the output from using the nearest matching method. The histogram shows that matched treated and control group do not agree well, even though after matching. The raw treated is data shift to left a bit after matching, but there are

still substantial differences from the matched control distribution. The differences between matched treated and matched control cannot be ignored, and so we then move to the full matching method.



(a) Jitter plot based on full matching



(b) Histogram of matched observations

Figure 3.2: Logistic regression using full matched data

Figure 3.2 shows that the full matching method gives much better matching when compared with nearest matching. The left tail of the matched treated units in the full match jitter plot have several large circles. The circles represent the relative number of control units which a particular treatment unit gets matched to. Note that in Figure 3.1(a), there are very few treated units less than 0.4, but there are many control units in this range, so matching of each treatment unit to a control, generates a one to many matching.

Figures 3.3 and 3.4 show the result of logistic regression of *aedend* on *neuronormal* using the two matched data sets. Figures 3.3 and 3.4 shows that both nearest matching and full matching give different estimated coefficients as compared to the logistic models fit in chapter 2.

Note that with nearest neighbor matching, the standard error for the coefficient

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.5002     0.1679   2.980  0.00288 **
neuronormal  -1.0860     0.2388  -4.548 5.41e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3.3: Nearest Matching test logistic outcome

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.2963     0.1648  -1.797   0.0728 .
neuronormal  -0.4052     0.1964  -2.064   0.0395 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3.4: Full Matching test logistic outcome

of *neuronormal* is larger than for the unmatched analysis in table 2.4, which might reflect the unmatched treatment units that have been ignored in the matched analysis. The estimated coefficient using full matching is considerably smaller than for any of the other models considered so far, and is only marginally significant (p=.0395). This may reflect the replication arising from many to one matching, or it may suggest that *neuronormal* is not a particularly important predictor of *aedend*.

We need to be cautious as in an observational study one may not be able to clearly define the treatment variable, and perhaps, in most cases, there is no well defined treatment, so the experiment may vary depend on subjective choice. Based on the histograms of the matched distributions, it seems like full matching does a better job than nearest neighbor matching. However, both methods have their own pros and cons, and it is not reasonable to conclude which one is better based on these histograms or the logistic regression outputs. We demonstrated both methods here in order to illustrate differences. However, in most cases of observational study, researchers may only subjectively choose one of the methods to do the matching. One

matching method shows (*neuronormal)* to be a very significant predictor, while the other method shows it to be only weakly predictive. In the next chapter we will give another way to identify causal effect, which somes give results similar to those which propensity score matching gave us here.

Both nearest matching and full matching use case weights when estimating the treatment effect. With full macthing, the weights do the entire work of balancing the covariates across the treatment groups. These weights are computed based on subclass membership, and the weights function like propensity score weights, which can be used to estimate a weighted treatment effect. Ignoring weights essentially ignores the entire purpose of matching. When doing matching with replacement, which is what we did with full matching above, weights must be included to ensure control units matched to multiple treated units are weighted accordingly. The only time weights can be omitted after pair matching is when performing 1:1 matching without replacement. That is what we did for 1:1 nearest neighbor matching. Excluding weights in this case will not affect the analysis since all work was done by using the propensity score matching distance to do pair matching. On the other hand, nearest neighbor matching often excludes unmatched observations, as seen with the unmatched treatment units in figure 3.1(a).

So in conclusion, why does propensity score matching work for causal inference? In a strict sense, propensity score adjustment has no more to do with causal inference than regression modeling does. The only real difference with propensity scores is that they make it easier to adjust for more observed potential confounders than that

sample size may allow regression models to incorporate. Propensity score adjustment can be thought of as a data reduction technique where the reduction is along an important axis - confounding. Propensity score matching can exclude many observations and thus be terribly inefficient. Any method that excludes relevant observations should be viewed as problematic. The real problem with matching is that it excludes easily matched observations due to some perceived need for having 1:1 matching, and most matching algorithms are observation order-dependent. We may how for some improvement on propensity score matching in the future.

Chapter 4

Graphical Causal Models

"Causal models are mathematical models representing causal relationships within an individual system or population."(Zalta et al. (1995)). Causal model facilitate the causal relationships from data, and it not only can show the causation, but also express the dependencies among variables. Figure 4.1 shows a very simple graph from our daily life example.

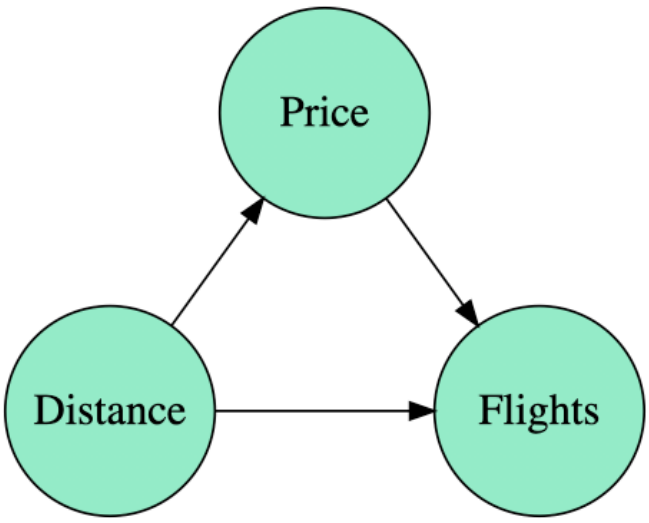


Figure 4.1: flight causal model

With intuitive common sense, both distance and price have effect on flights, and also distance has an effect on price as well. The arrows indicate causal relationships. If one alters the variable that an arrow begins from, that will cause a change in the

variable arrow goes in to. Graphs like this are causal models.

A graph is a collection of vertices X together with a collection of edges E. In statistical applications, the vertices represent random variables, and an edge between two vertices implies that two variables are associated. In a directed graph, the edges are directed, and the direction of the arrow denotes the direction of causation. For example $X_i \to X_j$ indicates that variable X_i causes X_j. A directed acyclic graph (DAG) is a directed graph without directed cycles, it only contains vertices and edges, with each edge going from one vertex to another. An acyclic graph has no closed loops such as $X_1 \to X_2 \ldots \to X_k \to X_1$. A parent of X is a node $Pa(X)$ for which $Pa(X) \to X$. A child of X is a node $Ch(X)$ for which $X \to Ch(X)$. These notions are generalized to ancestors and descendants, which may include intervening points.

The causal graph is the most important tool people use when doing causal analysis, and methods for causal inference on graphs typically assume the data arise from a DAG. Such graphs not only illustrate the causal relationship but also indicate the dependencies among variables, and what confounders should be including in fitting a relationship between treatment and outcome variables. Much of this work has been developed by Pearl (2009). Pearl, Glymour, and Jewell (2016) provides a more readable introduction to many of the associated issues. One important point is that every causal DAG implies a set of independence relationships that can be read off from the DAG by using a concept called d-separation.

4.1 d-separation

Figure 4.2 shows a very simple graph which gives some intuitive understanding of principles which will apply with more complicated models. This example comes from Pearl et al. (2016). In the figure, U_X, U_Y and U_Z are assumed to be independent random variables.

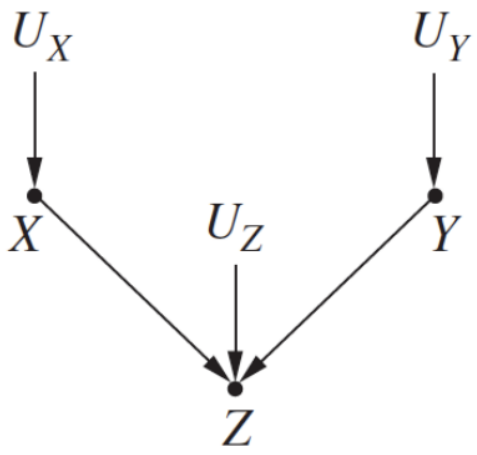


Figure 4.2: Simple collider

In the figure we can observe that variables X and Y both cause Z and each of the three variables have external causations which are independent of one other. Ignoring the external variables we can summarize the dependence relationships from the graph as following:

1. X and Z are dependent
2. Y and Z are dependent
3. X and Y independent
4. X and Y are dependent conditional on Z.

The first two points are obvious and the third point is fairly straightforward as well since there is no arrow or path connect X and Y. Point four is the key of this graph, and the question is why X and Y become dependent when conditioning on Z. If we condition on Z, that means we restrict consideration to cases in which Z takes the same fixed value. It is easiest to consider the linear case where $Z = X + Y$. Then with Z fixed, then an increase (decrease) in X must be accompanied by a decrease (increase) in Y, so conditioning on Z generates a dependence between X and Y.

We now introduce a term called d-separation, d-separation is a criterion for deciding, from a given causal graph, whether a set X of variables is independent of another set Y, given a third set Z.

The idea is to associate "dependence" with "connected" (i.e., the existence of a connecting path) and "independence" with "separation". The only twist on this simple idea is to define what we mean by "connecting path", and on the contrary, what we mean by "blocked path", and under what circumstance we can define the dependence and independence. To account for the orientations of the arrows we use the terms "d-separated" and "d-connected".

4.1.1 Unconditional separation

Rule 1: x and y are **d-connected** if there is an unblocked path between them.

By a "path" we mean a sequence of connected edges, disregarding their directionalities. By "unblocked path" we mean a path that can be traced without traversing a pair of arrows that collide "head-to-head", that is, a directed path in the graph

theoretic terminology. If two arrows collide into a variable from difference directions, for example $X_i \to X_j \leftarrow X_k$, the path connecting X_i and X_k is "blocked". In other words, arrows that meet head-to-head do not construct a connection for the purpose of passing, and such a meeting will be called a "collider".

Example 4.1:

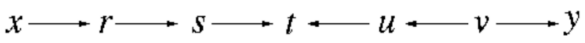


Figure 4.3: Rule 1 example

The graph in example 4.1 contains one collider at t. The path x-r-s-t is unblocked, hence x and t are d-connected. t and y are also d-connected, as well as the pairs u and y, t and v, t and u, x and s, etc. It might be confusing since it seems like there is no directed path from t to y or from y to t due to the structure $\leftarrow v \rightarrow$. However, the value of v will affect the values of t and y, so that t and y will likely be dependent, so are d-connected.

Strictly follow the definition of a collider at X as a structure $\rightarrow X \leftarrow$, there is no collider blocking the path connecting t and y, and we conclude that this is an unblocked path.

Moreover, x and y are not d-connected; there is no way of tracing a path from x to y without traversing the collider at t. Therefore, we conclude that x and y are d-separated, as well as x and v, s and u, r and u, etc.

4.1.2 Blocking by Conditioning

Rule 2: x and y are said to be d-connected, conditioned on a set of nodes Z, if there is a collider-free path between x and y that traverses no member of Z.

If no such path exists, we say that x and y are d-separated by Z, We also say then that every path between x and y is "blocked" by Z.

In example 4.2, $Z = \{r, v\}$.

Example 4.2:

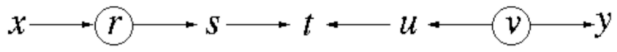


Figure 4.4: Rule 2 example

X and y are d-separated conditioned by Z, because there is no collider-free path between x and y that can travel without a node that contains in set Z, and so are also x and s,u and y. Conditioned on Z, the only d-connected paths here are s-t and u-t. s-t-u is blocked by Rule 1 since t is a collider.

4.1.3 Conditioning on Colliders

Rule 3: If a collider is a member of the conditioning set Z, or has a descendant in Z, then it no longer blocks any path that traces this collider.

In figure 4.5, let $Z = \{r, p\}$ be the nodes of the conditioning set. then s and y are d-connected because based on Rule 3, p is the descendant of collier t and if a collider itself or has a descendant in conditioning set Z, then this collider no longer blocks the path. Intuitively, conditioning on p generates information on t, and information on t

provides information on s and y. To see this, recall the discussion of Figure 4.2.

Example 4.3:

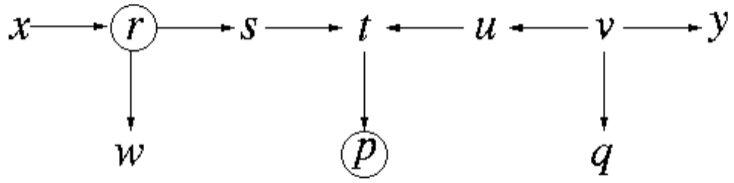


Figure 4.5: Rule 3 example

This means that s and y are likely to be dependent when conditioned on Z. They are conditionally d-connected. However, x and u are d-separated by Z using rule 2, as the path from x to u is blocked by $r \in Z$.

4.2 The *pc* algorithm for estimating the causal graph

While the causal graph implies important conditional independence relationships, in general, without suitable subject matter knowledge, the casual graph is unknown. Recently a number of algorithms have been developed which attempt to infer the structure of the causal graph from observational data. A survey of methods is given by Glymour, Zhang and Spirtes (2019). A number of these algorithms are available in the R library *pcalg* (Kalisch et al., 2012), The first widely used such algorithm was the *pc* (parent-child) algorithm (Spirtes et al., 2000). The *pc* algorithm requires a test of independence between two variables, conditional on other variables, in order to remove an edge between the two variables in question. For jointly Gaussian variables this is a test of zero partial correlation. For categorical variables, the R library *pcalg* uses the G^2 test of association, which is related to the likelihood ratio test. The G^2

test is used to test conditional or unconditional independence of discrete variables X and Y given the discrete set S, S can be the null set.

The following indicates the primary steps of the *pc* algorithm.

1. Form the complete undirected graph, with all nodes connected.

2. Eliminate edges between variables which are unconditionally independent.

3. For each pair of variables, say (X_i, X_j) with a connecting edge, and each other variable X_k with an edge to one of them, eliminate the edge between X_i and X_j if $X_i \perp X_j | X_k$. This requires a test of conditional independence.

4. For each pair (X_i, X_j) with a connecting edge, and each other pair of variables (X_k, X_l) with edges both connected to X_i or both connected to X_j, eliminate the edge between X_i and X_j if $X_i \perp X_j | (X_k, X_l)$.

5. Continue checking independence conditional on subsets of variables of increasing size, and remove edges where the test for conditional independence does not reject the null hypothesis.

6. For each triple (X_i, X_j, X_k) of variables with edges joining (X_i, X_j) and (X_j, X_k), but not (X_i, X_k), orient the undirected triple $X_i - X_j - X_k$ as $X_i \rightarrow X_j \leftarrow X_k$ provided X_j was not in the conditioning set on which X_i and X_k became independent

7. For each triple $X_i \rightarrow X_j - X_k$ where there is no edge joining X_i and X_k, orient $X_j - X_k$ as $X_j \rightarrow X_k$.

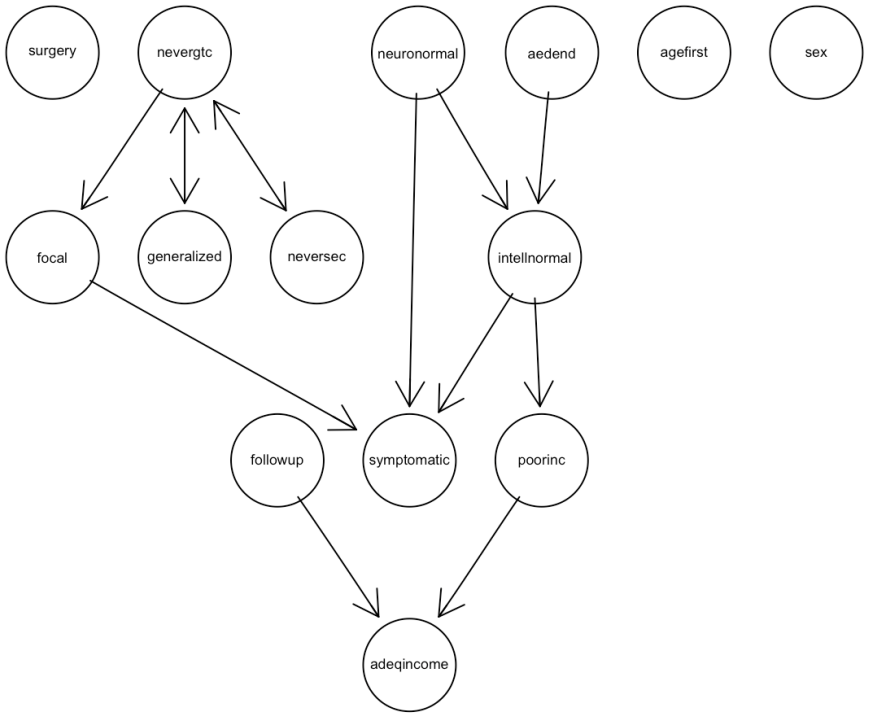


Figure 4.6: pc() function

Steps 1-5 generate the so-called *skeleton*, an undirected graph. Steps 6 and 7 are *edge propagation steps*. Different versions of the *pc* algorithm may contain several other edge propagation steps. One requirement is that edges are not allowed to be generated if they lead to a cycle in the graph.

When applied to the neurology dataset, using level $\alpha = .01$ for tests of independence, the *pc* algorithm estimates the causal graph as in Figure 4.6.

A number of operations have been developed which allow one to manipulate causal graphs to determine dependence relationships, assess effect size, and so on. In the following we discuss the two most important manipulations, the so-called *do calculus* and the *backdoor criterion*.

4.3 do-calculus

For a causal graph to be of use, it should lead to an estimate of the size of the causal effects between pairs of variables. With a single DAG, this can be done using Pearl's do-calculus, as described in Pearl et al. (2016).

By the notation $do(X = x)$, we can think of do as meaning some manipulation on X which fixes $X = x$ in the population. Suppose X is the only cause of Y then if we apply an intervention to change X and record change on Y, then the recorded change would be the causal effect of X on Y.

Suppose by external intervention, we first set the variable $X = x$ and then to the value $x + 1$. The effect of the change in X on Y can be expressed as

$$C(Y, X, x) = E(Y|do(X = x + 1)) - E(Y|do(X = x))$$

or, more generally, as

$$C(Y, X, x) = \frac{\delta}{\delta x} E(Y|do(X))|_{X=x}$$

.

Pearl's do-calculus Pearl et al. (2016) gives a numerical means of numerically evaluating the effect of intervention on X. In evaluating this, it is important that we condition on an appropriate set of control variables in order to block paths linking X and Y other than those which would exist in the graph where all paths into X have been removed. If other variables with unblocked paths to both X and Y exist, then there will be some confounding of the causal effect of X on Y.

The goal of causal inference is to estimate the causal effect of X on Y. In previous chapters we used logistic regression to estimate the effect of independent variables on a dependent variable. However, the results of regression and result of applying an intervention can be different, and that also indicates the differences between causal and non-causal relationship. The following two examples from Maathuis et al. (2009) illustrate the difference between a causal effect and a non-causal effect. Figure 4.7 gives the graphical models corresponding to the examples.

Example 4.1

$$Y = -X_1 + 2X_2 - X_3 + \epsilon_Y$$

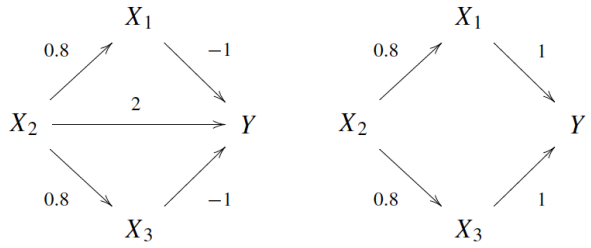
where $X_2 = \epsilon_2$, $X_1 = 0.8X_2 + \varepsilon_1$, $X_3 = 0.8X_2 + \varepsilon_3$. ϵ_1, ϵ_2, ϵ_3, and ϵ_Y are mutually independent normal random variables with mean zero and variances σ_1^2= 0.36, σ_2^2= 1, σ_3^2= 0.36 and $\sigma_Y^2 = 1$. Note that X_1, X_2 and X_3 all have variance 1.

In a regression analysis, X_2 is the most important predictor due to its having largest coeffiecent β=2. Then we move to intervention calculus, and let $\theta = (\theta_1, \theta_2, \theta_3)$ denote the causal effects of X_1, X_2, X_3 on Y. Where ϕ denotes the emply set, because $Pa(X_1) = X_2$, $Pa(X_2) = \phi$, and $Pa(X_3) = X_2$, we have $\theta_1 = \beta_{1|X_2} = -1$, $\theta_3 = \beta_{3|X_2} = -1$, and $\theta_2 = \beta_{2|\phi} = .8(-1) + 2 + .8(-1) = 0.4$, so the variable X_2 is the least important variable in terms of causal effect.

Example 4.2 As in example 4.1, but with

$$Y = X_1 + X_3 + \varepsilon$$

In the regression context, the least important variable is X_2 since it has coefficient



(a) Example 2.1 (b) Example 2.2

Figure 4.7: Graphical representation of the models used in Examples 4.1 and 4.2

$\beta_2 = 0$ in the regression. But in terms of intervention calculus, we have $\theta_1 = \beta_{1|X_2} = 1$, $\theta_3 = \beta_{3|X_2} = 1$, and $\theta_2 = \beta_{2|\phi} = .8(1) + .8(1) = 1.6$. Variable X_2 is now the most important variable. So the intervention calculus and regression analysis may give different results, even opposite results, since the set of variables that are controlled for may be different.

4.4 Backdoor Criterion

*"Given an ordered pair of variables (X, Y) in a directed acyclic graph G, a set of variables Z satisfies the backdoor criterion relative to (X, Y) if no node in Z is a descendant of X, and Z blocks every path between X and Y that contains an arrow into X."*Pearl et al. (2016)

That is, when estimating the effect of X on Y, we simply must make sure to keep all direct paths intact while blocking off any and all spurious paths to X, that is, all so-called **back-door paths**. We block the backdoor paths by conditioning.

If the backdoor criterion is satisfied, then Pearl et al. (2016) shows that the causal

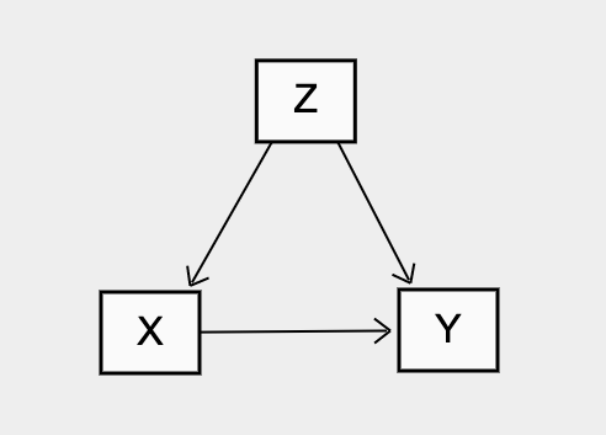


Figure 4.8: X is the cause of Y, but can not simply alter X and test on Y because both X and Y have a common parent Z and that means (1) X is affected by Z and X represents information from Z. (2) Z gives information to its child Y as well, we can not separate the causal effects from the indirect inference.

effect of X on Y is given by:

$$Pr(Y|do(X = x)) = \sum_z Pr(Y|X = x, Z = z)Pr(Z = z)$$

The proof begins with Pearl's *causal effect rule* (see Pearl et al. (2016)) which states that:

$$Pr(Y|do(X = x)) = \sum_t Pr(Pa(X) = t)Pr(Y|X = x, Pa(X) = t)$$

where t ranges over all the combinations of values that the parents $Pa(X)$ can take. This rule arises because it is the influence of the parents of X that is nullified when X is fixed by external manipulation.

The idea behind the backdoor criterion is that, in general, we would like to condition on a set of nodes Z such that:

- We block all spurious paths between X and Y.

- We leave all directed paths from X to Y unaltered.
- We create no new spurious paths from X to Y by conditioning on Z.

When trying to find the causal effect of X on Y, we want the nodes we condition on to block any "backdoor" path in which one end has an arrow into X, because such paths may make X and Y dependent, but they do not transmit causal influences from X. If we do not block them, they will confound the effect that X has on Y.

We condition on backdoor paths so as to fulfill our first requirement of blocking spurious paths. However, we don't want to condition on any nodes that are descendants of X. Descendants of X would be affected by an intervention on X and might themselves affect Y. Conditioning on them would block those pathways. Therefore, we don't condition on descendants of X so as to fulfill our second requirement. Finally, to comply with the third requirement, we should refrain from conditioning on any collider that would unblock a new path between X and Y.

Suppose that there is another set of conditioning variables Z in addition to $Pa(X)$, but which satisfies the backdoor criterion. Following from the causal effect rule, for any such set of variables Z, we have

$$Pr(Y|do(X=x)) = \sum_t Pr(Pa(X)=t) \sum_z Pr(Y, Z=z|X=x, Pa(X)=t)$$

Then

$$Pr(Y, Z=z|X=x, Pa(X)=t)$$

$$= Pr(Y|X=x, Pa(X)=t, Z=z) Pr(Z=z|X=x, Pa(X)=t)$$

so we have

$$Pr(Y|do(X=x)) =$$

$$\sum_t Pr(Pa(X)=t) \sum_z Pr(Y|X=x, Pa(X)=t, Z=z) Pr(Z=z|X=x, Pa(X)=t)$$

Now we use the fact that if Z satisfies the back-door criterion, then $Y \perp Pa(X)|X, Z$. Therefore,

$$Pr(Y|do(X=x)) =$$

$$\sum_t Pr(Pa(X)=t) \sum_z Pr(Y|X=x, Z=z) Pr(Z=z|X=x, Pa(X)=t)$$

Then as Z contains no descendants of X, $X \perp Z|Pa(X)$, so that

$$Pr(Y|do(X=x)) =$$

$$\sum_t Pr(Pa(X)=t) \sum_z Pr(Y|X=x, Z=z) Pr(Z=z, Pa(X)=t)$$

Since $\sum_t Pr(Pa(X)=t) Pr(Z=z|Pa(X)=t) = Pr(Z=z)$, we finally have

$$Pr(Y|do(X=x)) = \sum_z Pr(Y|X=x, Z=z) Pr(Z=z)$$

This says that we can estimate the causal effect of an intervention by manipulating conditional probabilities which can be estimated from observational data.

Observing Figure 4.6, the outcome variable *aedend* has only one arrow which goes out of the node. This means that there are no back-door paths which need to be blocked. The non-causal estimate and the causal estimates are the same for these data, and this choice of treatment (*neuronormal*) and outcome (*aedend*) variables. In fact, with (*aedend*) as the outcome, there is no possible treatment variable in the dataset for which adjustment is necessary to estimate a causal effect.

4.5 Equivalence classes of DAGS

Each DAG implies a set of conditional independence results which can be inferred from the structure of the graph. However, several DAG's can describe the same set of conditional independence results. Such a collection of DAGS forms an equivalence class. For example, with three variables $\{X_1, X_2, X_3\}$, the graphs $X_1 \rightarrow X_2 \rightarrow X_3$, $X_1 \leftarrow X_2 \leftarrow X_3$, and $X_1 \leftarrow X_2 \rightarrow X_3$ all have exactly one independence result, namely $X_1 \perp X_3 | X_2$. A second equivalence class, with only one member, is $X_1 \rightarrow X_2 \leftarrow X_3$, where X_2 is a collider. In this case the only conditional independence relation is the non-conditioned relationship $X_1 \perp X_3$.

An equivalence class can be described by a completely partially directed acyclic graph (CPDAG) Colombo et al. (2012) The CPDAG can be learned from conditional independence information if it can be assumed that the conditional independence relations are exactly those implied by the DAG via d-separation. For example, if it is known that the only conditional Independence relation is $X_1 \perp X_3 | X_2$, then the equivalence class consists of $X_1 \rightarrow X_2 \rightarrow X_3$, $X_1 \leftarrow X_2 \leftarrow X_3$, and $X_1 \leftarrow X_2 \rightarrow X_3$.

Under reasonable assumptions, the *pc* algorithm is able to identify the equivalence class, and has been shown to have an asymptotic consistency property(Kalisch and Bühlmann (2008)).

A DAG as estimated by *pc()* may contain both directed and undirected edges, with the edges being of 3 types:

1. there is an (directed or undirected) edge between i and j if and only if variables i and j are conditionally dependent given S for all possible subsets S of the

remaining nodes

2. a directed edge $i \to j$ means that this directed edge is present in all DAGs in the Markov equivalence class

3. an undirected edge $i - j$ means that there is at least one DAG in the Markov equivalence class with edge $i \to j$ and there is at least one DAG in the Markov equivalence class with edge $i \leftarrow j$.

For the neurology data set, the graph estimated by *pc* is shown in figure 4.6.

4.6 Latent variables and the *fci* algorithm

The following example by Richardson, taken from page 234 of Jordan (1998), describes a graph with a hidden confounder (a latent variable), and also with a selection effect.

The example represents a randomized trial of an ineffective drug with unpleasant side-effects. Patients are randomly assigned to the treatment or control group using indicator variable Tr. Those in the treatment group suffer unpleasant side-effects (variable SE), the severity of which is influenced by the patient's general level of health (H), with sicker patients having more severe side-effects. Those patients who suffer sufficiently severe side-effects are likely to drop out of the study. The selection variable (Sel) is an indicator which records whether or not a patient remains in the study. Since unhealthy patients who are taking the drug are more likely to drop out, those patients in the treatment group who remain in the study tend to be

healthier than those in the control group. Finally, general health status (H), which is an unobserved confounder, influences how rapidly the patient recovers (R). The graphical model is:

$Tr \rightarrow SE \leftarrow H \rightarrow R$, with additional edge $SE \rightarrow Sel$.

The example shows the need to accommodate the possibility of latent confounders and selection variables. The variables of interest, Tr and R, are observed to be correlated while the causal graph indicates independence between them. The observed correlation between treatment and response is an association induced by design, whereby only those subjects that eventually stay in the study are considered. The observed correlation, which is in effect a correlation conditional on the selection variable Sel, is an example of a selection effect, or selection bias. H is typical of a latent confounder which contributes to the spurious correlation.

In observational studies there may be several unobserved latent variables which affect the relationship between the variables of interest. Latent variables are marginalized out when only measured variables are analysed. This may cause difficulty in estimating relationships among the observed variables, or may even lead to non-identifiability.

For example, consider three observed variables $\{X_1, X_2, X_3\}$ and two latent variables $\{L_1, L_2\}$, forming the graph $L_1 \rightarrow X_1$, $L_1 \rightarrow X_2$, $L_2 \rightarrow X_2$, and $L_2 \rightarrow X_3$. The only independence relationship among the observed variables is $X_1 \perp X_3$, which leads to the single member equivalence class $X_1 \rightarrow X_2 \leftarrow X_3$, suggesting that both X_1 and X_3 are causes of X_2, which is an incorrect conclusion.

Another issue is that with latent variables present, the space of DAG's is not closed under marginalization and conditioning. For example, consider $X_1 \rightarrow X_2 \leftarrow L_1 \rightarrow X_3 \leftarrow X_4$. As shown in Colombo et al. (2012), this implies a set of Independent relationships among the X_j's for which there is no DAG on the four variables that satisfies the same set of conditional independence via d-separation.

As indicated in Zhang (2008), these difficulties can be dealt with by introducing a new class of graphs called *ancestral graphs*, and more particularly, a sub-class of *maximal ancestral graphs* (MAG) on the observed variables. A full discussion of the theory is beyond the scope of this thesis, but it is useful to identify some of the key ideas. The following summary is taken from Kalisch et al. (2012).

It has been shown that DAG with latent variables can be transformed into a unique MAG on the observed variables Zhang (2008). Several DAGs give the same MAG, in fact infinitely many if there is no constraint on the number of latent variables.

As with DAG's, it is only an equivalence class of MAG's which can be estimated. The equivalence class is represented by an object called a partial ancestral graph (PAG) Zhang (2008). A PAG contains the following types of edges: $o - o$, o—, $o \rightarrow$, $\rightarrow$, $\leftrightarrow$ and —. Bidirected edges indicate the presence of hidden variables, and undirected edges indicate the presence of selection variables. There is an edge between X and Y if and only if X and Y are conditionally dependent given S for all sets of variables S consisting of all selection variables and a subset of the observed variables. A tail on an edge means that this tail is present in all MAG's in the equivalence class. An arrow on an edge means that this arrow is present in all MAG's in the equivalence

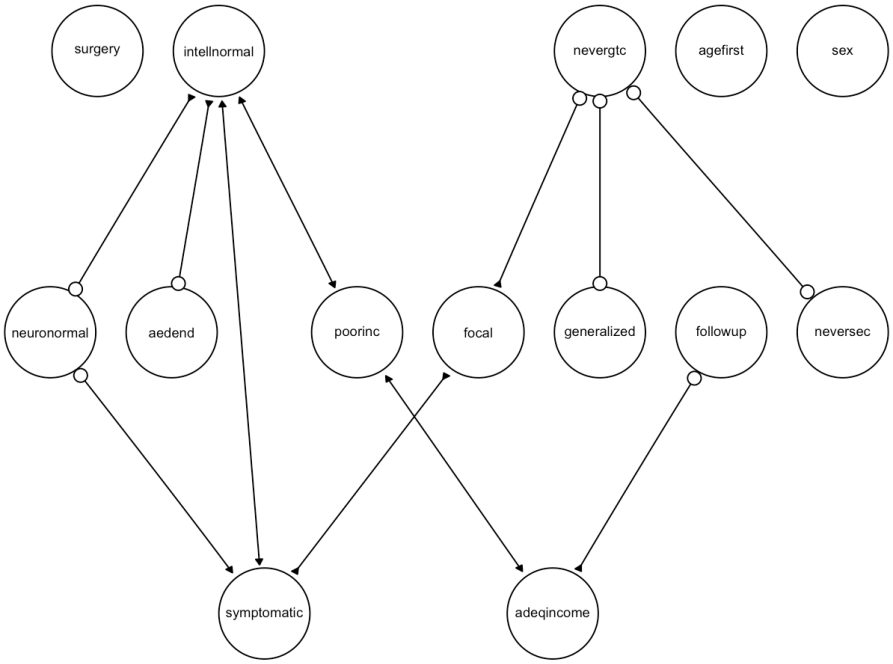


Figure 4.9: fci() function

class. An *o* edge mark means that there is a at least one MAG in the equivalence class where the edgemark is a tail, and at least one where the edgemark is an arrowhead.

A number of algorithms to estimate partial ancestral graphs were developed by Spirtes et al. (2000), Zhang (2008), and Colombo et al. (2012). Glymour et al. (2019) is a readable introduction to these and other algorithms for causal graph estimation, including the underlying assumptions.

Among other differences, the different algorithms typically have different orientation rules. Two popular algorithms in *pcalg* library which accommodate latent and selection variables are *fci* (fast conditional inference) and *rfci* (really fast conditional inference).

The graph in figure 4.9 was obtained by applying the *fci* function to the seizure data set.

	surgery	intellnormal	neuronormal	nevergtc	neversec	agefirst	followup
surgery	0	0	0	0	0	0	0
intellnormal	0	0	1	0	0	0	0
neuronormal	0	2	0	0	0	0	0
nevergtc	0	0	0	0	1	0	0
neversec	0	0	0	1	0	0	0
agefirst	0	0	0	0	0	0	0
followup	0	0	0	0	0	0	0
sex	0	0	0	0	0	0	0
focal	0	0	0	1	0	0	0
generalized	0	0	0	1	0	0	0
symptomatic	0	2	1	0	0	0	0
poorinc	0	2	0	0	0	0	0
adeqincome	0	0	0	0	0	0	1
aedend	0	2	0	0	0	0	0

	sex	focal	generalized	symptomatic	poorinc	adeqincome	aedend
surgery	0	0	0	0	0	0	0
intellnormal	0	0	0	2	2	0	1
neuronormal	0	0	0	2	0	0	0
nevergtc	0	2	1	0	0	0	0
neversec	0	0	0	0	0	0	0
agefirst	0	0	0	0	0	0	0
followup	0	0	0	0	0	2	0
sex	0	0	0	0	0	0	0
focal	0	0	0	2	0	0	0
generalized	0	0	0	0	0	0	0
symptomatic	0	2	0	0	0	0	0
poorinc	0	0	0	0	0	2	0
adeqincome	0	0	0	0	2	0	0
aedend	0	0	0	0	0	0	0

Figure 4.10: adjacency matrix

There is a known bug arising when the *plot* method is applied to the graph produced by *fci*. Specifically, the plot symbols can be misplaced, or the symbols inverted. The adjacency matrix shown in figure 4.10 corresponds to the plotted graph, and is necessary to clarify the structure of the graph. A 0 at coordinate (i, j) of the adjacency matrix means there is no edge between nodes i and j. A 1 at position (i, j) indicates the edge from i to j terminates in a "o", and a 2 at position (i, j) means that the edge from i to j terminates with an arrowhead.

Recall that for the PAG, an o edge mark means that there is a at least one MAG in the equivalence class where the edgemark is a tail, and at least one where the edgemark is an arrowhead. In comparing figures 4.6 and 4.9, it is clear that the graph in figure 4.6, estimated using *pc*, is one member of the equivalence class of

MAG's represented in figure 4.8, estimated using *fci*.

Based on what d-separation told us, *neuronormal* and *aedend* are dependent condition on *intellnormal*, because if we condition on *intellnormal*, that means we narrow down the interval of cases in which *intellnormal* can take value on. But what the value of *intellnormal* can take is dependent on *neuronormal* and *aedend*, so in order to fix the value of *intellnormal*, when we do some changes to *neuronormal*, we will have to adjust *aedend* as well to keep *intellnormal* stable, and that makes *neuronormal* and *aedend* has some connection in between. If *intellnormal* is fixed, then *neuronormal* and *aedend* will be correlated. In the same way, if we condition on symptomatic, that would make *neuronormal* and *intellnormal* are dependent.

Regardless of using *aedend* as dependent variable here, there are some other conclusions that can be drawn. Variables like *surgery*, *agefirst* and *sex* are not associated with any other variables. The variables determining the type of seizure, *nevergtc*, *generalized*, *neversec*, *focal*, and *symptomatic* are associated, but the only seizure variable related to the potential outcomes *intellnormal*, *neuronormal*, and *aedend* is *symptomatic* which is an indicator variable for brain injury. One interesting association has to do with income. Adequate income and low income are clearly related, being indicators associated with a broader income classification. There is a link between low income and intelligence level which may be due to families with low income not having access to educational opportunities, and thereby having an effect on measured intelligence level.

There is information which one causal graph can provide. What is different here

from the usual causal analysis is that there are no predefined treatment and control variables.

In seizure data example, due to the similarity of *pc()* graph and *fci()* graph, it is okay to conclude *pc()* graph is one member of the equivalence class estimated by *fci()*. They encode the same conditional independence information. In both *pc()* and *fci()* function, *pcalg* have preprogrammed versions of indepTest() for Gaussian data (gaussCItest()), discrete data (disCItest()), and binary data (binCItest()). Each of these independence test functions needs different arguments as input. We use *disCItest()* here and it is unnecessary move to the independence test for binary variables since both *disCItest()* and *binCItest()* is based on the G^2 statistic and takes as input a list containing the data matrix.

It is worth noting that there is a much simpler class of graphical model called a *conditional independence graph* (Lauritzen, 1996), in which two variables are not joined by an edge if and only if a test declares them to be conditionally independent given all other variables in the graph. In the case of jointly Gaussian variables, the conditional independence graph gives a so-called *partial correlation graph*, and the graphical structure gives the joint covariance matrix.

There is a function *backdoor()* in the *pcalg* package which can be used to identify backdoor paths. This function first checks if the total causal effect of one variable (x) onto another variable (y) is identifiable via the back-door criterion, and if this is the case, it explicitly gives a set of variables that satisfies the back-door criterion with respect to x and y in the given graph. We used the function to verify that there is

no backdoor path to outcome variable *aedend*.

```
> backdoor(fcip001.amat,3,14)
NULL
```

Figure 4.11: backdoor() output

Figure 4.11 shows result of *backdoor()* function to test the cause of *neuronormal* onto *aedend* in *fci()* graph. No back-door path is found, and therefore no adjustment is needed. Figure 4.9 indicates that in some members of the estimated equivalence class, including the model in figure 4.6, there are no variables in the data set which should be used to model *aedend*, while in other members of the equivalence class *intellnormal* is a useful predictor.

In terms of the models fit previously, the analysis after full matching with propensity scores comes closes to what the graphical models are telling us. Recall that in that case, the p-value for the significance of *neuronormal* was only marginally significant (p=.0395).

Chapter 5

Conclusion and Future Work

An important, but generally neglected issue when applying causal inference methods to observational data is in a truly observational study, there is often, and perhaps most often, not a well defined treatment variable. Thus there may be several of the measured covariates that may need to be considered as the treatment, with causal inference methods applied to each. Also, whether matching methods or causal graph are used, subjective choices may need to be made, for example, the choice of matching method, the algorithm for graph estimation and so on.

We chose *neuronormal* as treatment variable of interest, and *aedend* as the outcome, from among a number of possible choices.

Tthe propensity score matching method is also subjective. In this thesis, we both both nearest neighbor matching and full matching. Researchers might try different methods and make subject decisions as to which is better. Each method will have relative advantages and disadvantages.

When using propensity scores, one might use a causal model to develop the propensity score rather than a simple logistic regression, which is likely itself to be biased or to have excess variation due to the inclusion of inappropriate predictors.

In our observational seizure data we began with investigating the effect of neurological status on outcome. However, the estimated causal graph helps to identify useful relationships among many variables. For example in our estimated graph, the variables determining type of epilepsy form a sub-graph, and are associated with other variables only through symptomatic epilepsy. This dramatically simplifies the possible associations between epilepsy type and a number of possible outcomes.

In carrying out a causal analysis, it is clearly advantageous to have a causal graph based on subject matter knowledge at the beginning of the study. As is usual, there is no substitute for a properly designed experiment, and the controlled randomized trial is still the gold standard for assessing the causal effect of a treatment on an outcome.

Bibliography

Peter C Austin. Using the bootstrap to improve estimation and confidence intervals for regression coefficients selected using backwards variable elimination. *Statistics in medicine*, 27(17):3286–3300, 2008.

Carol Camfield, Peter Camfield, and Bruce Smith. Poor versus rich children with epilepsy have the same clinical course and remission rates but a less favorable social outcome: A population-based study with 25 years of follow-up. *Epilepsia*, 57 (11):1826–1833, 2016.

Diego Colombo, Marloes H Maathuis, Markus Kalisch, and Thomas S Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, pages 294–321, 2012.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

Michael Irwin Jordan. *Learning in graphical models*, volume 89. Springer Science & Business Media, 1998.

Markus Kalisch and Peter Bühlmann. Robustification of the pc-algorithm for directed acyclic graphs. *Journal of Computational and Graphical Statistics*, 17(4):773–789, 2008.

Markus Kalisch, Martin Mächler, Diego Colombo, Marloes H Maathuis, and Peter Bühlmann. Causal inference using graphical models with the r package pcalg. *Journal of Statistical Software*, 47(11):1–26, 2012.

Marloes H Maathuis, Markus Kalisch, Peter Bühlmann, et al. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A):3133–3164, 2009.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

Raymond E Wright. Logistic regression. 1995.

Edward N Zalta, Uri Nodelman, Colin Allen, and John Perry. Stanford encyclopedia of philosophy, 1995.

Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896, 2008.