Causal Inference Notes

Alex Mansourati

December 2019

1 Preface

These notes were compiled while reading *Causal Inference* by Hernán MA and Robins JM (2019). Specifically, these notes are from Part I: Chapters 1, 2, 3, 6, 7, 8, Part II: Chapters 11-16, Part III: 19, 20, 21. There are some supplementary figures and definitions throughout from *Elements of Causal Inference* by Bernhard Schölkopf, Dominik Janzing, and Jonas Peters (2017).

2 Introduction

Advances in machine learning have produced increasingly spectacular results in creation of datasets and efficiency of computation of the best results. "With enough data, we can get arbitrarily close to the lowest risk eventually".

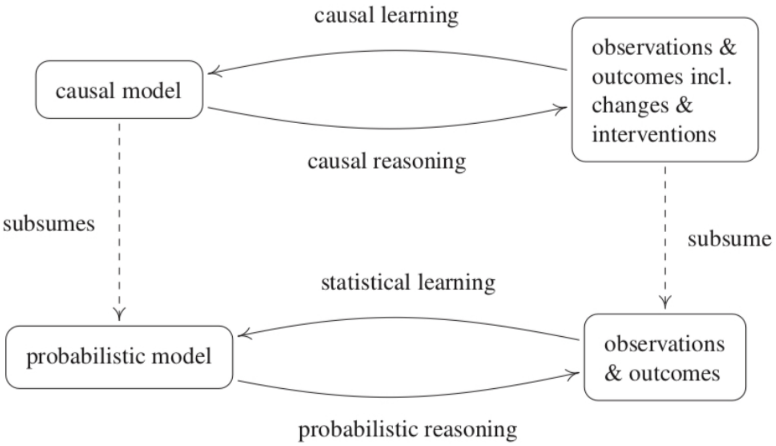

Figure 1: Relationship between **probabilistic inference** and **causal inference**, via *Elements of Causal Inference* (Bernhard Schölkopf, Dominik Janzing, and Jonas Peters).

Causal modelling *entails* a probability model, but it contains additional information *not* contained in the latter. Causal models contain more information that probabilistic ones do, because causal reasoning allows us to analyze the effect of interventions on distribution changes (see Figure 1).

3 Exchangeability and Randomization

In causal inference, we typically want to test the impact of a treatment A on an outcome Y. For example, we may want to understand the impact of a treatment like smoking ($A = 1$ for smokers who quit and $A = 0$ for those who continued smoking) on an outcome lung cancer ($Y = 1$ for individuals who contract lung cancer and $Y = 0$ for those who did not). Causation and association between Y and A mean two different things. Causation is concerned with *what if* questions about a *counterfactual* world, like ”what would be the outcome if everyone had been treated?”. Meanwhile, association is concerned with questions in the actual world, like ”what is the outcome of those treated?”. We of course have notation to match this discrepancy in meaning. We use $P[Y = 1|A = a]$ for association, i.e. what was the probability of contracting lung cancer of the individuals who received treatment a? Meanwhile, we use variables $Y^{a=1}$ and $Y^{a=0}$ to represent potential outcomes, or counterfactual outcomes. $P[Y^a = 1]$ is unconditional, it represents the outcome of the entire population under treatment $a = 1$. Causation is concerned with entire population under different potential worlds while association is concerned with disjoint sets within the given population.

A foundational concept in causal inference is the notion of *exchangeability*. Exchangeability is the condition that Y^a is independent of A for all a. In a dichotomous treatment, exchangeability holds under the following conditions,

$$P(Y^a = 1|A = 1) = P(Y^a = 1|A = 0) = P(Y^a = 1) \tag{1}$$

Randomization is important because it is expected to inherently produce exchangeability. This is so because randomization ensures that the independent variables predicting the outcome are equally distributed between the treated and untreated.

In a simple randomized control trial, the causal effect is written as $P(Y^{a=1}) - P(Y^{a=0})$. Randomization does not have to follow a 50-50 split of treated and untreated. We can also imagine that randomization can occur within different subsets of our population. For example, we may give a treatment to those in critical condition at a higher rate than those in stable conditions. We refer to these different subsets as the **strata** of population and are defined by covariates, L. If randomization is maintained within subsets of population ($Y^a \perp\!\!\!\perp A|L = l$) then we have **conditional exchangeability**.

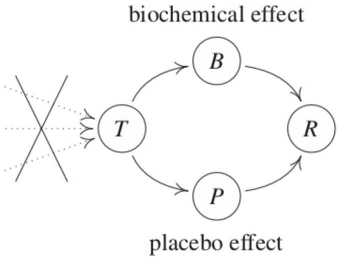

Figure 2: Simplified description of randomization study. T is treatment and R indicates recovery. The randomization over T removes the influence of any other variable on T and so removes any common influence between T and R, producing exchangeability. via *Elements of Causal Inference* (Bernhard Schölkopf, Dominik Janzing, and Jonas Peters).

4 Observational Studies

In observational studies, we lack the exchangeability offered by randomized control trials. However, we can attempt to analyze our data as if treatment had been randomly assigned, by conditioning on a set of measured covariates, L. Our observational data can be conceptualized as a randomized trial with the following conditions (referred to as **identifiability conditions** since they allow us to "identify" causal effects):

1. Consistency: The values of treatment under comparison are well-defined interventions that correspond to versions of treatment in data (i.e. well defined interventions).

2. Exchangeability: We assume that our data is conditionally randomized (i.e. we have exchangeability within levels of variables L).

3. Positivity: The probability of receiving every value of treatment conditional on L is greater than zero (i.e. within each strata of L, each treatment is observed).

In ideal randomized experiments, these identifiability conditions hold by design. In observational studies, we need to assume that identifiability conditions hold. These are considered "heroic" assumptions, explaining why observational studies are often viewed with suspicion.

When these assumptions break down, there are other options for identifying causal effects. For example, another option is using an instrumental variable (see Section 11). This has its own set of assumptions but is better suited for situation where gathering all covariates L is not possible (i.e. economics).

5 Causal Diagrams

A **DAG**, or directed acyclic graph, is a simple way to encode our subject-matter knowledge and our assumptions about the qualitative causal structure of a problem.

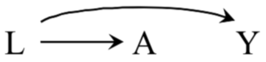

Figure 3: A simple DAG.

Two nodes on a DAG are marginally associated if one causes the other (arrow from one node to the other) or if they share common causes, otherwise two nodes are marginally independent. Conditional independence can be seen in Figure 3, A and Y are marginally dependent but conditionally independent given L.

The following defines a set of rules for understanding how conditioning changes how variables depend on each other:

> **Pearl's D-separation**: In a DAG G, a path between nodes i_1 and i_m are **blocked by a set** L (with neither i_1 nor i_m in L) whenever there is a node i_k such that one of the following two possibilities holds:
>
> (i) $i_k \in L$ and
>
> $i_{k-1} \to i_k \to i_{k+1}$ or
> $i_{k-1} \leftarrow i_k \leftarrow i_{k+1}$ or
> $i_{k-1} \leftarrow i_k \to i_{k+1}$
>
> (ii) neither i_k nor any of its descendants is in L and
>
> $i_{k-1} \to i_k \leftarrow i_{k+1}$ (aka ”collider”)
>
> Furthermore, in a DAG G, we say that two disjoint subsets of vertices A and B are d-separated by a third (also disjoint) subset L if every path between nodes in A and B is blocked by L. We then write $A \perp\!\!\!\perp B | L$.
>
> Via *Elements of Causal Inference* (Bernhard Schölkopf, Dominik Janzing, and Jonas Peters), Page 83

The first possibility listed in Pearl's D-separation rules shows us how we can use confounders to remove paths from treatment to outcome, thereby isolating the effect of our treatment on outcome. The second possibility shows us how we can inadvertently introduce bias into our model by conditioning on **colliders**, nodes that share common causes. When we condition on a collider, or one of its descendants, we open a path of dependence that didn't previously exist. This results in selection bias.

6 Confounding

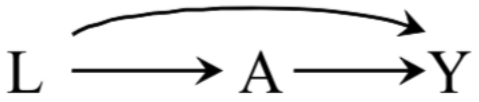

Figure 4: DAG with three nodes.

Consider the DAG in Figure 4. There are two paths of association between A and Y. The path through L is a "backdoor path". If L didn't exist, the entire association between A and Y would be causal effect of A on Y. In that simpler world, $E[Y^{a=1}] - E[Y^{a=0}]$ is our causal effect and in the world with L, $E[Y^{a=1}|L] - E[Y^{a=0}|L]$ is our causal effect since we have conditional exchangeability, $A \perp\!\!\!\perp Y^a|L$.

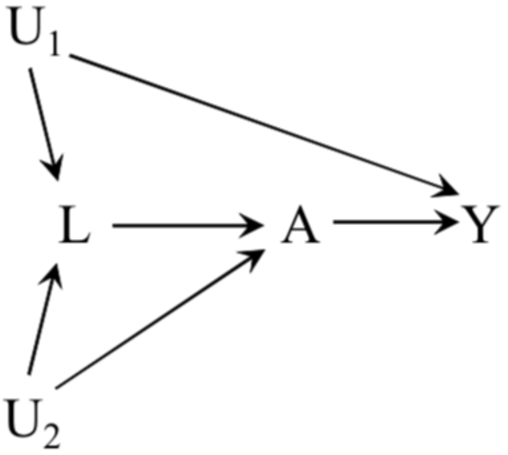

Figure 5: DAG with a collider opening backdoor path between A and Y.

In the DAG represented by Figure 5 above, the backdoor path is *not* successfully blocked by conditioning on L (assuming our DAG is correct!). We have *opened* a backdoor path by controlling on L. As described in Section 5, if we control on a collider, we can open a backdoor path, introducing new bias.

Traditionally, to identify confounder, researchers just looked for variables that were associated with both the treatment and outcome. Then, if the confounder-adjusted and unadjusted estimates differed, this process declared the existence of confounding. However, it is clear that this traditional method is harmful since it does not discount colliders.

DAG's are our a priori representation of how we believe causal relationships exist. It makes assumptions about confounders explicit and therefore open for criticism by other investigators.

There are modern techniques to aid in producing a DAG from relationships in observational data. See Chapter 7 of *Elements of Causal Inference* (Bernhard Schölkopf, Dominik Janzing, and Jonas Peters) for structure identification strategies. Even with these techniques, uncertainty about the set of possible causal structures is unavoidable.

We will discuss techniques to remove confounding. First, there are **G-methods** which try to "delete" the arrow from L to A. These techniques include standardization and IP weighting. They exploit the conditional exchangeability given L on entire population. Second, there are **stratification based methods** which do not "delete" arrow but rather compute conditional effect in subset of population defined by L. These techniques include stratification, matching and conventional outcome regression.

7 Why Model?

An **estimator** is some function of the data that is used to estimate unknown population parameter and a consistent estimator is one that moves closer to the population value as the sample size increases. In the case when treatment is not dichotomous, we may need some way of knowing how a treatment performs at various dosage levels. In this case, "the data does not speak for itself" and we often need to supplement the data with a model.

First example is a parametric estimator of the conditional mean, $E[Y|A] = \theta_0 + \theta_1 A$. This equation is a restriction on the shape of the conditional mean function $E[Y|A]$. We can now access the mean outcome for individuals at various treatment levels but this does not come for free. When using a parametric model, the inferences are correct only if the restrictions encoded in the model are correct (aka "correctly specified").

Most of book is devoted to model-based causal inference - relies on the condition of "no model misspecification". Some model misspecification is almost always expected, this can by partially rectified by using non-parametric estimator.

In the categorical treatment case, building a linear model would be called a **saturated model** since it just lets the data speak for itself, doesn't restrict anything. But we refer to them as models anyways because they look like models. More generally, a model is saturated when number of parameters in a conditional mean model is equal to the number of unknown conditional means in population.

When a model has only a few parameters but is used to estimate many population quantities then we say the model is **parsimonious**.

Standardization and IP weighting mentioned later in notes are examples of nonparametric estimators - they are based on a saturated model, no prior restrictions on the value of effect esimates.

Smoothing: Suppose that linear model from before was not as best as we could do, what if small change in low dose made a larger change on outcome than small change in high dose. We could introduce a squared term to the model, this would still be a linear model but would have a squared term.

Bias-Variance Tradeoff: more parameters in the model, less bias, but higher variance (bigger 95% confidence interval). Fewer parameters in the model, more bias but lower variance (smaller 95% confidence interval).

8 Parametric G-Methods

8.1 IP Weighting

Inverse probability weighting is a tool to balance the covariates, or control for the covariates, in each of the treatment groups. IP weights get used in the weighted least squares procedure.

IP weights effectively create a pseudo-population in which the arrow from the covariates L to treatment A is removed. This means that in this pseudo-population, A and L are independent and $E_{ps}[Y|A = a] = \sum_l E[Y|A = a, L = l]Pr[L = l]$. The distribution of the variables in L are the same in treated and untreated in this pseudo-population.

We estimate $E_{ps}[Y|A = 1]-E_{ps}[Y|A = 0]$ in the pseudo-population by fitting the (saturated) linear mean model $E[Y|A] = \theta_0+\theta_1 A$ by weighted least squares, with individuals weighted by their estimated IP weights: $\hat{W} : 1/\hat{Pr}[A = 1|L]$ for those that took treatment and $1/(1 - \hat{Pr}[A = 1|L])$ for those that did not take treatment. Here, the value of theta tells us the impact of the treatment on outcome.

We have both stabilized and non-stabilized IP weights. Both methods result in the same value of theta but stabilized weights will typically result in narrower 95% confidence intervals than non-stabilized weights.

8.2 Standardization

An alternative to IP weighting is standardization. While in IP weighting we built a "treatment model" (modelling $Pr[L = l]$), in standardization, we build an "outcome model" (modelling $E[Y|A = a, L = l]$). We model the conditional means for each strata, $E[Y|A = a, L = l]$. In the common case where our estimate is over millions of strata, we can fit a linear regression model for the mean outcome under treatment A and all confounders, included as covariates. Our objective is still to calculate $\sum_l E[Y|A = a, L = l]Pr[L = l]$ but fortunately, here, we do not need to estimate $P[L = l]$, we can equivalently calculate, $\frac{1}{n}\sum_l E[Y|A = a, L = l]$ since $\sum_l E[Y|A = a, L = l]Pr[L = l]$ can be equivalently written as the double expectation $E[E[Y|A = a, L = l]]$. This technique is also referred to as the *parametric g-formula.*

The following is a simple 4 step computational method to use Standardization with dichotomous outcome and confounder, without ever explicitly estimating $Pr[L = l]$:

1. Expand dataset: duplicate dataset twice and hardcode treatments on first duplicate to 0 and hardcode treatments to 1 on the second.

2. Outcome modelling: Use the original dataset to fit outcome regression.

3. Prediction: Predict outcome on the two new datasets that do not have unknown outcomes.

4. Average: Take the average of all outcomes predicted in second dataset and, separately, the average of all outcomes predicted in third dataset. These are the treatment effects in treated and untreated. The difference is the average causal effect.

Large differences between IP weighting and standardization can alert us of serious model misspecification. Both models yield the same result when no models are used to estimate them. Otherwise, they are expected to differ. No need to choose between the two methods, use both when possible.

Generally speaking, the parametric g-formula uses estimates of any function of the distribution of the outcome within levels of A and L to compute its standardized values.

The two g-methods described thus far can be used to estimate the ACE (average causal effect, also referred to as the average treatment effect) in the entire population of interest. If we were interested in a sub-population, we would restrict the calculation to that subset (also knowing as CATE, conditional average treatment effect).

Again, the validity of the methods of causal inference requires the following conditions: exchangeability, positivity, consistency, no measurement error and no model misspecification.

8.3 G-Estimation of Nested Models

G-estimation uses *structural nested models.* Under exchangeability, structural nested models (*structural nested mean model* in this case) can be written as:

$$E[Y^a - Y^{a=0}|A = a, L] = \beta_1 a + \beta_2 a L \tag{2}$$

Here we are calculating $E[Y^a|L] - E[Y^{a=0}|L]$ directly. To understand how this technique let us (re)consider three topics:

1. Exchangeability: Conditional exchangeability means the outcome distribution in treated and untreated would be the same if both groups received same treatment level. Or, in other words, $Pr[A = 1|Y^{a=0}, L] = Pr[A = 1|L]$. Knowing the counterfactual outcome in a given strata does not add any new information.

2. Structural Nested Mean Models: $E[Y^{a=1}|L] - E[Y^{a=0}|L] = E[Y^{a=1} - Y^{a=0}|L] = \beta_1 a + \beta_2 aL$. Removing the last effect-measure modification term, β_1 would be the ACE in each stratum and in entire population. This model will be estimated via g-estimation (described below). Standardization used parametric models for mean outcome of Y which like structural nest models here, rely on treatment A and covariates L (recall, marginal structural model looks like $E[Y^a|V] = \beta_0 + \beta_1 a + \beta_2 aV + \beta_3 V$. In contrast, these models are semi-parametric since they are agnostic of both the intercept and main effect of L (i.e. no β_0 and no β_3 for a term $\beta_3 L$). Therefore, fewer assumptions here, more robust to model misspecification.

3. Rank Preservation: Imagine we could rank individuals by both of their counterfactual outcomes. If both lists are identically ordered, we have *rank preservation* and moreover, models maintaining rank preservation within strata of L have *conditional additive rank preservation*: $Y_i^a - Y_i^{a=0} = \psi_1 a + \psi_2 aL_i$ for individuals i. This rank preservation is unlikely and is not used in this book but it is easier to describe G-estimation using rank preservation.

Suppose the goal is to estimate the parameters of structural nested mean model. For simplicity, start with, $E[Y^a - Y^{a=0}] = \beta_1 a$ (i.e. assuming ACE is constant across strata of L).

Assume additive rank-preserving model $Y_i^a - Y_i^{a=0} = \psi_1 a$ is correctly specified. We can re-write this as $Y_i^{a=0} = Y - \psi_1 A$ where Y and A represent what we have observed. How do we find ψ_1? Consider $Y^{a=0}$ as a function $H(\psi_1)$.

Consider the following equation:

$$logitPr[A = 1|H(\psi_1), L] = \alpha_0 + \alpha_1 H(\psi_1) + \alpha_2 L \tag{3}$$

From our earlier discussion on exchangeability, we know that the correct value of $H(\psi_1)$ in the following equation will produce a fitted model such that $\alpha_1 = 0$ (again, since $Pr[A = 1]$ is independent of $Y^{a=0}$ in each strata L).

So, the strategy here is to search over the space of ψ, calculate $H(\psi)$ and fit logistic regressions.

Now consider structural nested models with 2+ parameters, a model less like to be misspecified. Introduce, for example, $\beta_2 aL$ since we believe some variation in outcome based on covariates. The same brute force technique of searching over ψ_1 and ψ_2 is used here, with equations: $Y_1^a - Y_i^{a=0} = \psi_1 a + \psi_2 aV$ and $logitPr[A = 1|H(\psi_1, \psi_2), L] = \alpha_0 + \alpha_1 H(\psi_1) + \alpha_2 H(\psi_2)V + \alpha_3 L$. The brute force method search will be more involved but do-able. Structural nested models are not often used partly because of lack of user-friendly software and partly because extension to survival analysis requires additional considerations.

9 Stratification Methods

These methods are the most popular but are mentioned last because in general, they don't work well.

9.1 Outcome Regression

Similar to technique in Section 8.2, we fit an *outcome regression* model using *ordinary least squares*,

$$E[Y|A = a, L = l] = \alpha_0 + \alpha_1 A + \alpha_2 AL + \alpha_3 L \tag{4}$$

If the variables L are enough to account for confounding and model is correctly specfied, no further adjustment is needed. This is the end of the procedure, where as in Section 8.2 we had one more step.

9.2 Propensity Scores

Propensity scores, $\pi(L)$, measure the propensity of individuals to receive treatment given information available in covariates. We can generate propensity scores by fitting a logistic model for the probability of treatment A conditional on the covariates L. Individuals with the same propensity score will generally have different values of some covariates L, but the distribution of L will be the same in the treated and untreated, that is $A \perp\!\!\!\perp L|\pi(L)$. In other words, conditional exchangeability $Y^a \perp\!\!\!\perp A|L$ implies $A \perp\!\!\!\perp L|\pi(L)$. Because the propensity score is a continuous one-dimensional summary of multidimensional L, we can fit flexible models, like cubic splines rather than a single linear term.

Propensity matching attempts to form a matched population in which the treated and untreated are exchangeable because they have the same distribution of $\pi(L)$. For example, each treated individual is paired with one untreated individual with same propensity score value. The subset of original population comprised of these matched pairs is the *matched population* and under exchangeability and positivity given $\pi(L)$, association measures are consistent estimates of effect measures. This technique restricts analysis to treatment groups with overlapping distributions of estimated propensity score which is useful but simply restricting study population to that range is a lazy way to ensure positivity. We lose the transportability of results when restriction is solely based on value of propensity scores.

10 Summarizing Why We Use These Techniques

It is important to note that we do not necessarily need to predict our outcome well, we just need to include the variables with causal interpretation to guarantee exchangeability. If other covariates are included, we can end up with very high variance. Besides variance inflation, a predictive approach to causal inference, may also result in self-inflicted bias (i.e. colliders). All causal inference models require no misspecification of the functional form of the covariates (why we use cubic splines rather than linear terms) as well as the usual assumptions of exchangeability, positivity and consistency.

11 Instrumental Variables

Techniques up to this point (IP weighting, standardization, g-estimation, stratification, matching) have all worked to ensure conditional exchangeability. Unfortunately, all these models will fail if confounders are unmeasured, imperfectly measured or misspecified in model. IV methods avoid this confounder issue (economists like this technique for this reason).

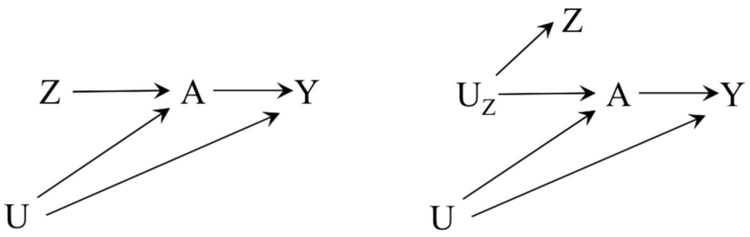

Figure 6: DAGs illustrating typical instrumental variables Z.

As an example, in Figure 6, consider Z as the price of cigarettes, A as smoking cessation and Y as weight loss.

IV methods rely on 4 conditions: (i) Z is associated with A, (ii) Z does not effect Y except through A, (iii) Z and Y do not share causes, (iv) monotonicity (details below). The assumptions for instrumental variables are very difficult to meet and so we typically refer to the IV as a *proposed IV* or *candidate IV* since conditions (ii) and (iii) are not empirically provable.

Historically, the fourth assumption was *effect homogeneity*. The definition of effect homogeneity has evolved through time but the earliest and most extreme being the assumption of constant effect of treatment A on outcome Y across individuals. This is of course very hard to acheive. Fortunately, in the early 1990s, an alternative assumption was proposed: *Monotonicity*. This alternative allows us to endow the usual IV estimand (Equation 5) with a causal interpretation, even though it does not suffice to identify the average causal effect in population. Let's explain. Imagine a double-blind randomized trial with Z (treatment assignment), A (treatment) and Y (outcome). If we knew values of the 2 counterfactual treatment variables, $A^{z=1}$ and $A^{z=0}$, for each participant, 4 disjoint subpopulations appear:

1. Always-takers: $A^{z=1} = A^{z=0} = 1$
2. Never-takers: $A^{z=1} = A^{z=0} = 0$
3. Compliers: $A^{z=1} = 1, A^{z=0} = 0$
4. Defiers: $A^{z=1} = 0, A^{z=0} = 1$

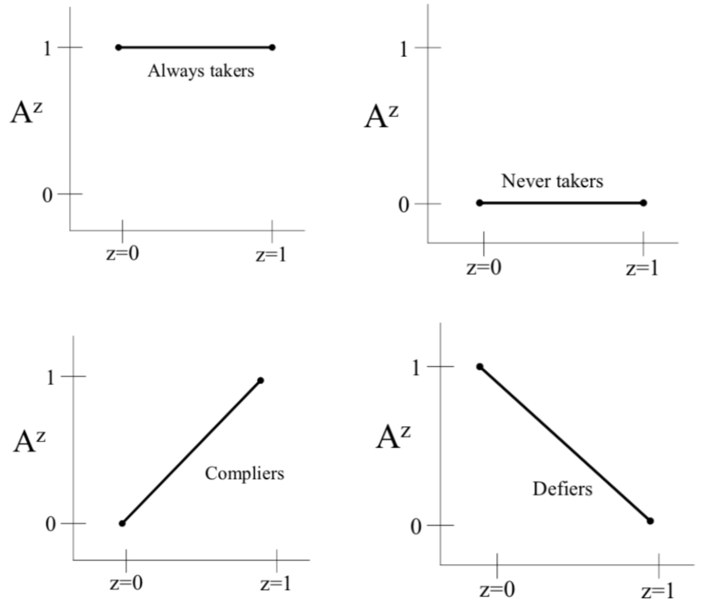

Figure 7: Illustration of the four compliance types.

These subpopulation are often referred to as *compliance types* or *principal strata*. Monotonicity occurs in the absence of **defiers**. Under monotonicity, the following is referred to as the *IV estimand* and is the ACE in the compliers:

$$E[Y^{a=1}] - E[Y^{a=0}] = \frac{E[Y|Z=1] - E[Y|Z=0]}{E[A|Z=1] - E[A|Z=0]} \tag{5}$$

The numerator here is the ACE of Z on Y, also referred to as the "intention to treat effect", and the denominator is the ACE of Z on A, also referred to as the measure of "adherence". The IV estimand under monotonicity represents the ACE of compliers only because the numerator is the effect of Z on Y but Z has no effect on never-takers and always-takers. So, while homogeneity seemed impossible to meet, monotonicity often seems plausible. The difficulty with monotonicity assumption becomes whether policymakers should prioritize treatments that have been shown to be impactful for only compliers:

> "Rather than arguing that the effect of the compliers is of primary interest, it may be more honest to accept that interest in the estimand is not the result of its practical relevance, but rather of the (often erroneous) perception that it is easy to identify." (Section 16.4, p. 202)

> "If the effect of compliers is considered of interest, relying on monotonicity is a promising approach in double-blind randomized trials

> with two arms and all-or-nothiing compliance, especially when one arm will exhibit full-adherence by design” (Section 16.4, p. 203)

An important topic in IV estimation is in the use of ”weak instruments”, which occurs when the true value of the Z-A association, the denominator of IV estimand, is small. This causes the following problems:

1. Yields effect estimates with wide 95% confidence interval

2. Amplifies bias from conditions (ii) and (iii). Weak instrument creates small denominator, amplifying numerator.

3. Weak instrument can be due to chance and not actual effect, yielding underestimated variance.

A strong proposed instrument that slightly violates conditions (ii) and (iii) may be preferable to a less invalid, but weaker proposed instrument. IV methods vs. other methods:

1. Requires modelling assumptions even with infinite data

2. Source of bias often easier to think about in other methods, this is why IV method should be used cautiously by novices

IV estimation is best used in a situation with a lot of unmeasured confounding, dichotomous and time-fixed A, strong/causal Z, one is genuinely interested in compliers and monotonicity holds.

12 Time-Varying Treatments

Treatment history at time k is represented as $\bar{A_k} = (A_0, A_1, ..., A_k)$, where A_i represents the treatment at time i. $\bar{A}$ is used when referring to entire history. The ACE of time varying treatment can no longer be defined as $E[Y^{a_k=1}] - E[Y^{a_k=0}]$ since this would only represent impact of treatment at a single time. The ACE of time varying treatment is only well-defined if treatment strategies of interest are specified.

It is helpful to differentiate between *dynamic treatment strategies* and *static treatment strategies.* A dynamic strategy is a rule in which treatment a_k depends on an individual's time-varying covariate(s), $\bar{L_k}$. A static strategy is the opposite, a strategy like ”always treat”, ”never treat” or ”treat every other day”.

For a static strategy, $E[Y^{\bar{a}}]$ is simply the mean outcome $E[Y|\bar{A} = \bar{a}]$ among those who followed $\bar{a}$. This is not true for dynamic strategies (let this strategy be g). $E[Y^g]$ depending on variables L requires application of g-methods on $\bar{L}, \bar{A}$ and Y to achieve conditional exchangeablity at each timepoint (i.e. *sequential exchangeability).*

$$Y_g \perp\!\!\!\perp A_0 | L_0 \tag{6}$$

$$Y_g \perp\!\!\!\perp A_1 | A_0 \tag{7}$$

A *sequentially randomized experiment* is an experiment in which treatment is randomly assigned at each time k to each individual – this is not often used but useful to understand key conditions for valid estimation.

For any strategy g, the treated and untreated at each time k are exchangeable for Y^g conditional on prior covariate history L_k and any observed treatment history $\bar{A}_{k-1} = g(\bar{A}_{k-2}, \bar{L}_{k-1})$ compatible with strategy g. This holds in sequantially randomized experiments and observational studies where probasbiiltiy of receivnig treatment at each time depends on treatment and measured covariate history $\bar{A}_{k-1}$, $\bar{L}_k$.

A SWIG is a DAG that represents a counterfactual world under a particular intervention. This along with d-separation can allow us to check for conditional exchangeability.

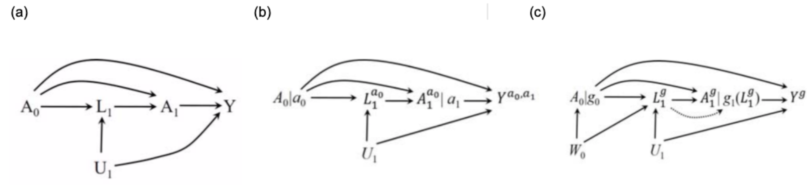

Figure 8: A DAG and SWIGs representing static/dynamic treatment. (b) is a corresponding static SWIG to DAG (a) and (c) has an added node W_0.

DAG (b) in Figure 8 is the DAG (a) world if all individuals received static strategy (a_0, a_1), where a_0 and a_1 can take values 0 or 1. Constants are automatically conditioned, so paths are blocked here.

Static sequential exchangeability for $Y^{\bar{a}}$ looks like the following, $Y^{\bar{a}} \perp\!\!\!\perp A_k | \bar{A}_{k-1} = a_{k-1}, \bar{L}_k$ for $k = 0, 1, ..., k$. This is weaker than sequential exchangeability for Y^g becuase it only requires conditional independence between counterfactual outcomes $Y^{\bar{a}}$ indexed by static strategies $g = \bar{a}$ and treatment A_k.

In a SWIG like (c) in Figure 8, the dotted line indicates when intervention relies on outcome of L, this relationship is shown differently but acts like any other arrow in classifying d-separation. If the W_0 node did not exist in (c), we can verify here that $Y^g \perp\!\!\!\perp A_0$ and that $Y^g \perp\!\!\!\perp A_1 | A_0 = g_0, L_1^g$. So, we can identify the mean counterfactual outcome under all strategies g.

Including W_0, missing confounders, in (c), $Y^g \perp\!\!\!\perp A_0$ no longer holds because of open path through W_0.

We can imagine the very long causal diagram (extending Figure 8) that contains all time points $k = 0, 1, 2, ...,$ and in which L_k affects subsequent $A_k, A_{k+1}, ...$ and shares unmeasured causes U_k with the outcome Y. At each time k we can use the covariate history L_k with A_{k-1} to block the backdoor paths between treatment A_k and the outcome Y (including the unmeasured U_k!). We can call these $\bar{L}_k$, *time-varying confounders*.

13 Treatment-Confounder Feedback

Treatment-confounder feedback occurs when confounder affects the treatment and the treatment affect the confounder. See Figure 9 for example. Note that if there are confounders on A_0 and L_1 then there is still feedback without a direct causal relationship between A_0 and L_1. In this figure, $E[Y^{a_0=1,a_0=1}] - E[Y^{a_0=0,a_1=0}]$ should be equal to zero (null case based on DAG below) but the traditional methods so far will fail us here.

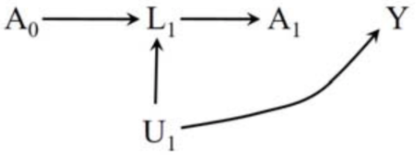

Figure 9: Basic DAG illustrating treatment confounder feedback

The problem is that methods like stratification conditions on our confounder L (a collider) which will in fact open a path from A_0 to Y, introducing a non-causal association. In other words, stratification eliminates confounding for A_1 at the cost of introducing selection bias for A_0. Again, note that A_0 and L_1 can be connected indirectly, through unmeasured confounders, for this to take place.

One modelling assumption to deal with this could be to use cumulative treatment, $cum(\bar{A}) = A_0 + A_1$, which in the dichotomous treatment case would take values 0, 1, 2. We expect the difference between "always treat" and "never treat" strategies, $E[Y^{cum(\bar{a}=2)}] - E[Y^{cum(\bar{a})=0}]$ to be zero based on our DAG but again we are conditioning on L and will be biased away from our expected outcome.

14 G-Methods for Time-Varying Treatments

With a_0, l, a like last section, we can compute the g-formula but need to condition on $(A_0 = a_0, A_1 = a_1, L = l_1)$. We calculate the join distribution of the counterfactuals $(Y^{\bar{a}}, L^{\bar{a}})$ under strategy $\bar{a}$.

The corrected g-formula in the general case over all possible l histories:

$$\sum_{l} E[Y|\bar{A} = \bar{a}, \bar{L} = l] \prod_{k=0}^{K} f(l_k|\bar{a}_{k-1}, \bar{l}_{k-1}) \tag{8}$$

Recall that there are 2^k possible treatement histories for dichotomous treatment.

$E[Y|\bar{A} = \bar{a}, \bar{L} = \bar{l}]$ and $\hat{f}(l_k|\hat{a}_{k-1}, \hat{l}_{k-1})$ will need to be estimated if the data are high-dimensional (expected in observational studies). So, estimates will need to be plugged into the g-formula. This is referred to as a "plug-in g-formula" and if based on parametric models, referred to as the "parametric g-formula".

Suppose we want to measure effect of time-fixed treatment A_1 only (i.e. $E[Y^{a_1=1}) - E[Y^{a_1=0}])$. Earlier in text we defined IP weights (propensity scores) as $W^{A_1} = \frac{1}{f(A_1|L_1}$. For time-varying treatment and confounders, IP weights need to be generalized. For $\bar{A} = (A_0, A_1)$, $\bar{L} = (L_0, L_1)$:

$$W^{\bar{A}} = \frac{1}{f(A_0|L_0)} * \frac{1}{f(A_1|A_0, L_0, L_1)} = \prod_{k=0}^{1} \frac{1}{f(A_k|\bar{A}_{k-1}, \bar{L}_k)} \tag{9}$$

with $A_{-1} = 0$ by definition. If IP weighting and the g-formula explained thus far differ, this is because of model misspecification, regardless of identifiability assumption.

Marginal structural mean model is fit on pseudo population (i.e. using IP weights):

$$E[\bar{Y}^{\bar{a}}] = \beta_0 + \beta_1 * cum(\bar{a}) \tag{10}$$

$$ACE = E[Y^{\bar{a}}] - E[Y^{\bar{a}=\bar{0}}] = \beta_! * cum(\bar{a}) \tag{11}$$

This is misspecified iif outcome depends on somethting otther htan the cumulattive treattement ($cum(a)$). We could alternatively fit:

$$E[Y|\bar{A}] = \theta_0 + \theta_1 * cum(\bar{A}) + \theta_2 * cum_{-5}(\bar{A}) + \theta_3 * cum(\bar{A})^2 \tag{12}$$

In this case, we've specified the outcome to depend on the lastt 5 months of non-linearly cumulative treatment.

Next, lets combine these two methods together to create a doubly robust estimator. Lets begin with simple, non-time varying case using both outcome model (g-estimation) and treatment model (IP weighting). As long as one of these models is correctly specified, we get valid estimation (bias cancels). We use the following steps:

1. Compute IP weights
2. Compute outcome model including IP weights as a parameter (+W if A=1, -W if A=0).
3. Standardize under A=1 and A=0

To extend to time-varying case, we follow the same steps with modification to step 2. We must fit a separate outcome regression model at each time m starting from last time k and going down towards m=0. For example, with treatment history summarized in $cum(\bar{A}_m)$,

$$E[\hat{T}_{m-1}|\bar{A}_m, \bar{L}_m] = \theta_{0,m} + \theta_1 * cum(\bar{A}_m) + \theta_2 * cum(L_m) + \theta_3 * \hat{W}_m^A \tag{13}$$

Under conditional exchangeability and positivity given L_m, this estimator validly estimates the AVE if one of the 3 following statements holds:

1. Treatment model is correct at all times.
2. Outcome model is correct at all times.
3. Treatment model is correct for time 0 to time K and outcome correct for $k+1$ to k for an $k < K$ (aka "k+1 robustness" condition)