STAT 535: Statistical Machine Learning **Autumn 2024**

Lecture 14: Causal inference

Instructor: Yen-Chi Chen

14.1 Potential outcome model

In the previous lecture, we have seen that the DAG (directed acyclic graph) can be used to draw causal conclusion from the data. Here we introduce another framework for drawing causal conclusion called *potential outcome* model, a commonly used framework in medical research and social sciences.

Let $Y \in \mathbb{R}$ be the response variable/variable of interest and $A \in \{0, 1\}$ be the binary treatment. $A = 1$ refers to the case that the individual receive a treatment (treatment group) and $A = 0$ refers to the case that the individual receive a placebo (control group). You can think of Y as a measure of health condition (such as blood pressure) and the binary treatment A refers to weather this individual receives certain treatment or not. The goal is to study the causal effect of A on Y.

Under this scenario, our data consists of pairs

$$(Y_1, A_1), \cdots, (Y_n, A_n),$$

where Y_i is the outcome of the i-th individual and A_i is the treatment indicator of the i-th individual.

If the treatment A indeed has a causal effect on Y, then we should think of two versions of Y, denoted as $Y(0)$ and $Y(1)$. $Y(0)$ is the outcome variable if the individual does not receive any treatment ($A = 0$). $Y(1)$ is the outcome variable in the case that the individual receive a treatment ($A = 1$). The above model is called the potential outcome model. Here is a key concept of the potential outcome model:

$$\text{“}Y|A = 0\text{”} = \text{“}Y(0)|A = 0\text{”}, \quad \text{“}Y|A = 1\text{”} = \text{“}Y(1)|A = 1\text{”}.$$

Namely, given $A = 0$, we can replace Y by $Y(0)$ and given $A = 1$, we can replace Y by $Y(1)$.

In the potential outcome model, every individual has two outcomes variables. One is the observed outcome that we can observe and the other is the *potential outcome* that we do not observe. For instance, suppose that $A_i = 0$ (no treatment) then $Y_i = Y_i(0)$ is the observed outcome. The other outcome $Y_i(1)$ is the potential outcome of the response Y_i (if the individual receive the treatment). Thus, we only observe one of the two outcomes. Note that the observed response $Y_i = Y_i(A_i)$.

The causal effect can be viewed as the distributional difference between random variables $Y(1)$ and $Y(0)$. A simple summary of the difference is the mean difference, which is also known as the *average treatment effect (ATE)*:

$$\tau = \mathbb{E}(Y(1)) - \mathbb{E}(Y(0)).$$

So we will think of methods for estimating the ATE.

One may think of using the difference in conditional mean to estimate the ATE. Namely, we use the estimator

$$\widehat{\tau}_{\text{naive}} = \frac{\sum_{i=1}^n Y_i I(A_i = 1)}{\sum_{i=1}^n I(A_i = 1)} - \frac{\sum_{i=1}^n Y_i I(A_i = 0)}{\sum_{i=1}^n I(A_i = 0)} = \frac{\sum_{i=1}^n Y_i A_i}{\sum_{i=1}^n A_i} - \frac{\sum_{i=1}^n Y_i (1 - A_i)}{\sum_{i=1}^n (1 - A_i)}.$$

However, this estimator may be biased if the response $Y_i(0), Y_i(1)$ and A are dependent. For a concrete example, suppose that a doctor always give a treatment to patients that look very sick and the treatment

only have a small effect. Then even if the treatment is effective, the averaged outcome of those who received the treatment will still be lower than the averaged outcome of those without the treatment.

Thus, a common requirement to ensure that the native estimator converges to the ATE is

$$(Y(0), Y(1)) \perp A. \tag{14.1}$$

The independence of the two versions of outcomes and the treatment assignment. Under this assumption (and some other mild conditions such as the absolute mean exists and there is positive probability for an individual receiving/not receiving a treatment), we have

$$\widehat{\tau}_{\mathsf{naive}} \xrightarrow{P} \tau.$$

In clinical trial, one scenario to ensure $(Y(0), Y(1)) \perp A$ is the *randomized-control trial*: every individual is randomly assigned to treatment or control without using any additional information about this individual.

To see why randomization $(Y_i(0), Y_i(1)) \perp A$ makes the naive estimator work, note that $\mathbb{E}(Y|A = a) = \mathbb{E}(Y(a)|A = a)$ for $a = 0, 1$. Then under randomization

$$\begin{aligned} Y(1) \perp A &\Rightarrow \mathbb{E}(Y|A = 1) = \mathbb{E}(Y(1)|A = 1) = \mathbb{E}(Y(1)), \\ Y(0) \perp A &\Rightarrow \mathbb{E}(Y|A = 0) = \mathbb{E}(Y(0)|A = 0) = \mathbb{E}(Y(0)). \end{aligned}$$

Thus,

$$\tau = \mathbb{E}(Y(1)) - \mathbb{E}(Y(0)) = \mathbb{E}(Y|A = 1) - \mathbb{E}(Y|A = 0)$$

and the right-hand sided is what $\widehat{\tau}_{\mathsf{naive}}$ is consistently estimating.

14.1.1 Relaxing randomization

In practice, the total randomization on the treatment may be very challenging or even unethical (this basically requires that a doctor has to choose not to treat someone who is very sick when the randomized decision is $A = 0$). And randomization is often note the case in an observational study. So we would like to think of relaxing the condition $(Y_i(0), Y_i(1)) \perp A$.

One possible approach is to use the confounder (confounding variable) X. Namely, in our data, we not only observe the outcome Y and the treatment A but also some additional information about each individual, denoted as X (could be univariate or multivariate). So our data is

$$(Y_1, A_1, X_1), \cdots, (Y_n, A_n, X_n).$$

In a medical study, X is often the demographic variables (gender, educational level, ...etc) but it could also be a clinical variable of other diseases or health conditions.

We allow the outcomes $(Y_i(0), Y_i(1))$ and the treatment A to be dependent, but they are *conditionally independent* given the observed confounding variable X. Namely,

$$(Y(0), Y(1)) \perp A|X. \tag{14.2}$$

Under this assumption, we have

$$\begin{aligned} Y(1) \perp A|X &\Rightarrow \mathbb{E}(Y|A = 1, X) = \mathbb{E}(Y(1)|A = 1, X) = \mathbb{E}(Y(1)|X), \\ Y(0) \perp A|X &\Rightarrow \mathbb{E}(Y|A = 0, X) = \mathbb{E}(Y(0)|A = 0, X) = \mathbb{E}(Y(0)|X). \end{aligned}$$

- **Regression adjusted estimator.** By the law of total expectation, $\mathbb{E}(Y(a)) = \mathbb{E}(\mathbb{E}(Y(a)|X))$ so we can rewrite the ATE as

$$\begin{aligned}\tau = \mathbb{E}(Y(1)) - \mathbb{E}(Y(0)) &= \mathbb{E}(\mathbb{E}(Y(1)|X)) - \mathbb{E}(\mathbb{E}(Y(0)|X)) \\ &= \mathbb{E}(\mathbb{E}(Y|A=1,X)) - \mathbb{E}(\mathbb{E}(Y|A=0,X)).\end{aligned} \tag{14.3}$$

  Let $m_1(x) = \mathbb{E}(Y|A=1, X=x)$ and $m_0(x) = \mathbb{E}(Y|A=0, X=x)$ be the regression function of the treatment and the control groups. It is easy to see that they can be estimated using the group-specific data (observations with $A=1$ or $A=0$). Then Equation (14.3) implies that the ATE can be written as

$$\tau = \mathbb{E}(m_1(X) - m_0(X)).$$

  Thus, let $\widehat{m}_1(x)$ and $\widehat{m}_0(x)$ be the regression estimator (you may use a parametric estimator or a nonparametric estimator). Then we can estimate the ATE using

$$\widehat{\tau}_{\mathsf{RA}} = \frac{1}{n}\sum_{i=1}^{n}(\widehat{m}_1(X_i) - \widehat{m}_0(X_i)).$$

- **Inverse probability weighted (IPW) estimator.** The IPW uses an alternative property of (14.2) that the conditional expectation

$$\begin{aligned}\mathbb{E}(YI(A=a)|X) = \mathbb{E}(\mathbb{E}(YI(A=a)|A,X)|X) &= \mathbb{E}(\underbrace{\mathbb{E}(Y(a)|X)}_{=\omega(X)} I(A=a)|X) \\ &= \mathbb{E}(Y(a)|X)P(A=a|X).\end{aligned}$$

  The quantity $\pi_a(X) = P(A=a|X)$ is called the *propensity score*, which can be easily estimated (by treating A as the response variable and apply a regression with respect to X). The above equation implies

$$\mathbb{E}(Y(a)) = \mathbb{E}(\mathbb{E}(Y(a)|X)) = \mathbb{E}\left\{\frac{\mathbb{E}(YI(A=a)|X)}{\pi_a(X)}\right\} = \mathbb{E}\left\{\mathbb{E}\left\{\frac{YI(A=a)}{\pi_a(X)}|X\right\}\right\} = \mathbb{E}\left\{\frac{YI(A=a)}{\pi_a(X)}\right\},$$

  which implies that following estimator of $\mathbb{E}(Y(a))$:

$$\widehat{\mathbb{E}}(Y(a)) = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i I(A_i=a)}{\pi_a(X_i)}.$$

  With the estimated propensity scores $\widehat{\pi}_a(x)$, the ATE can be estimated using

$$\widehat{\tau}_{\mathsf{IPW}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i I(A_i=1)}{\widehat{\pi}_1(X_i)} - \frac{Y_i I(A_i=0)}{\widehat{\pi}_0(X_i)}\right).$$

  This estimator is called IPW because we inversely weighting each response Y_i according to the propensity score $\widehat{\pi}_a(X_i)$.

- **Doubly-robust estimator.** We may combine both RA and IPW estimators to form a doubly-robust estimator. The key insight is as follows. We can rewrite $\mathbb{E}(Y(a))$ as

$$\begin{aligned}\mathbb{E}(Y(a)) &= \mathbb{E}\left\{\frac{(Y - m_a(X))I(A=a)}{\pi_a(X)} + m_a(X)\right\} \\ &= \mathbb{E}\left\{\frac{1}{\pi_a(X)}\left[YI(A=a) + m_a(X)(\pi_a(X) - I(A=a))\right]\right\}.\end{aligned}$$

Her is an interesting property about this equality. If the regression function $m_a(X) = \mathbb{E}(Y|A = a, X)$, then even if the propensity score $\pi_a(x) \neq P(A = a|X = x)$, we still have $\mathbb{E}\left\{\frac{(Y-m_a(X))I(A=a)}{\pi_a(X)}\right\} = 0$ so the first equality gives $\mathbb{E}(Y(a)) = \mathbb{E}(m_a(X))$, which is still consistent. On the other hand, if the propensity score $\pi_a(x) = P(A = a|X = x)$ but the regression function is mis-specified $m_a(X) \neq \mathbb{E}(Y|A = a, X)$, we still have $\mathbb{E}\left(\frac{1}{\pi_a(X)} m_a(X)(\pi_a(X) - I(A = a))\right) = 0$ so the second equality leads to $\mathbb{E}(Y(a)) = \mathbb{E}\left\{\frac{1}{\pi_a(X)} Y I(A = a)\right\}$, again still consistent. Thus, either the regression function or the propensity score is correctly specified, we have a consistent estimator. This means that our estimator is doubly-robust to the models we are using and the corresponding estimator is called a doubly-robust estimator.

14.1.2 Local average treatment effect and instrumental variable

Some useful references:

- http://www.cedlas-er.org/sites/default/files/cer_ien_activity_files/miami_late.pdf
- http://ec2-184-72-107-21.compute-1.amazonaws.com/_assets/files/events/slides_late

In many medical research, although we randomized the treatment assignment to participants, they may not comply with what we ask them to do. This creates a problem when we are attempting to estimate the causal effect since the treatment assignment and the actual treatment are different.

One possible solution to this problem is to introduce the concept of instrumental variable (IV) and view the treatment assignment as an instrument and define a separate variable for the actual treatment that is being used. Again, let Y denote the outcome variable of interest and A denote the *actual treatment* and Z denote the *instrument* (the assigned treatment). For simplicity, assume that both A and Z are binary. $A = 1$ denotes the case where the individual receive a treatment and $Z = 1$ denotes the case where we assign the individual to receive a treatment (but the individual may refuse to take it, leading to the case $A = 0$ and $Z = 1$). On the other hand, $A = 0$ is the control case and $Z = 0$ is the assignment that the individual is assigned to be in the control group (it could happen that $A = 1$ and $Z = 0$–the individual still takes the treatment even if we assign him/her not to). So our data is

$$(Y_1, A_1, Z_1), \cdots, (Y_n, A_n, Z_n).$$

In this case, the actual treatment A has two potential outcome $A(0)$ and $A(1)$. According to the potential outcome, we can definite the individual to 4 categories: Note that we only get to observe $A(z)|Z = z$, namely,

$(A(0), A(1))$	
1,1	Always-taker
0,1	Complier
0,0	Never-taker
1,0	Defier

$$\text{“}A|Z = 0\text{”} = \text{“}A(0)|Z = 0\text{”}, \quad \text{“}A|Z = 1\text{”} = \text{“}A(1)|Z = 1\text{”}.$$

In this case, the outcome variable has 4 potential outcomes, depending on A, Z: $Y(a, z)$. We only have access to observe $Y(a, z)|A = a, Z = z$, namely, one of the four potential outcomes:

$$\text{“}Y|A = 0, Z = 0\text{”} = \text{“}Y(0, 0)|A = 0, Z = 0\text{”}, \quad \text{“}Y|A = 0, Z = 1\text{”} = \text{“}Y(0, 1)|A = 0, Z = 1\text{”},$$
$$\text{“}Y|A = 1, Z = 0\text{”} = \text{“}Y(1, 0)|A = 1, Z = 0\text{”}, \quad \text{“}Y|A = 1, Z = 1\text{”} = \text{“}Y(1, 1)|A = 1, Z = 1\text{”}.$$

In this case, we often assumed that

$$\text{(Exclusion Restriction)} \quad Y(a,z) = Y(a,z') \quad \text{for all } a,z,z'. \tag{14.4}$$

Namely, the IV has no effect on the potential outcomes–the difference is due to the actual treatment. This reduces the 4 potential outcomes into 2 potential outcomes $\{Y(a) : a = 0, 1\}$. Also, the randomization of Z can be viewed as the condition

$$\text{(Randomization)} \quad Z \perp Y(0), Y(1), A(0), A(1). \tag{14.5}$$

Due to the problem that the actual treatment and the potential outcomes may be dependent (we only randomized at the assignment Z), it is hard to identify meaningful causal effect without making assumptions. Identifying the ATE is not feasible with the above two conditions. However, we are able to identify the local average treatment effect (LATE)

$$\tau_{\mathsf{LATE}} = \mathbb{E}(Y(1) - Y(0)|\mathsf{complier}).$$

The LATE measures the causal effect on those who complied with our assignment. You can show that under Exclusion Restriction and Randomization[1], the LATE can be written as

$$\tau_{\mathsf{LATE}} = \frac{\mathbb{E}(Y|Z=1) - \mathbb{E}(Y|Z=0)}{\mathbb{E}(A|Z=1) - \mathbb{E}(A|Z=0)}, \tag{14.6}$$

and we can easily each expectation using conditional mean.

14.1.3 Dynamic treatment regime

The dynamic treatment regime is a popular approach to the precision medicine (also known as the personalized medicine). It has received a lot of attentions these days from the causal inference community. Here we briefly discuss its basic concept and give a high-level introduction about how the method works.

The dynamic treatment regime considers the problem where we have multiple time points that we need to make a decision on the treatment that an individual receive. Meanwhile, there will be new information coming up before we make a new treatment assignment.

Consider the simplest case where we have two time points so we have two possible treatment assignment $A_1, A_2 \in \{0, 1\}$ that are both binary (it can be easily generalized to multiple categories). When the individual enters the study, we collect their baseline information, denoted as X. Then we make the decision on the first treatment $A_1 = a_1(X)$ using the baseline information. After some time, the individual comes back and we measure his/her first outcome variable Y_1. Then we use all the information available (i.e., X, A_1, Y_1) to make a second treatment $A_2 = a_2(X, A_1, Y_1)$. After a while, the individual comes back and we collect the final information on the outcome variable Y_2. The goal is to maximizes the expected outcome Y_2 by choosing the optimal treatments a_1, a_2. In this case, (a_1, a_2) is called the treatment regime.

The techniques used in solving a dynamic treatment regime problem involve ideas from 1. classification, 2. Markov chains, and 3. dynamic programming.

To analyze this problem, we introduce an objective/utility function of a_1, a_2:

$$V(a_1, a_2) = \mathbb{E}(Y_2|A_1 = a_1, A_2 = a_2),$$

[1] See https://faculty.washington.edu/yenchic/short_note/note_IV.pdf

which is the expected (final) outcome of the study variable Y_2 under a treatment regime (a_1, a_2) (sometimes it is called a policy in the bandit problem). The best treatment regime is

$$(a_1^*, a_2^*) = \mathsf{argmax}_{a_1,a_2} V(a_1, a_2).$$

Note that $a_1^* = a_1^*(X), a_2^* = a_2^*(Y_1, A_1, X)$.

We can further expand $V(a_1, a_2)$ as

$$\begin{aligned} V(a_1, a_2) &= \mathbb{E}(Y_2|A_1 = a_1, A_2 = a_2) \\ &= \mathbb{E}(\mathbb{E}(Y_2|A_1 = a_1, A_2 = a_2, X, Y_1)) \\ &= \int\int \mathbb{E}(Y_2|A_1 = a_1, A_2 = a_2, X = x, Y_1 = y_1)p(y_1|A_1 = a_1, X = x)dy_1 p(x)dx. \end{aligned}$$

The above equality shows a very interesting feature–the only part that involves a_2 is in the conditional expectation of Y_2. Thus, the optimal treatment regime a_2^* will be the one that maximizes it, i.e.,

$$a_2^*(Y_1, a_1, X) = \mathsf{argmax}_{a_2} \mathbb{E}(Y_2|A_1 = a_1, A_2 = a_2, X, Y_1).$$

We can rewrite it as

$$a_2^*(Y_1, a_1, X) = \begin{cases} 1, & \text{if } \mathbb{E}(Y_2|A_1 = a_1, A_2 = 1, X, Y_1) \geq \mathbb{E}(Y_2|A_1 = a_1, A_2 = 0, X, Y_1) \\ 0, & \text{if } \mathbb{E}(Y_2|A_1 = a_1, A_2 = 1, X, Y_1) < \mathbb{E}(Y_2|A_1 = a_1, A_2 = 0, X, Y_1) \end{cases}. \tag{14.7}$$

With this, we can then rewrite the conditional expectation under the optimal treatment a_2^* as

$$\omega_2(a_1, Y_1, X) = \mathbb{E}(Y_2|A_1 = a_1, A_2 = a_2^*, X, Y_1).$$

To obtain the optimal treatment of a_1, we consider the case where the optimal a_2^* is used so the objective function becomes

$$\begin{aligned} V(a_1, a_2^*) &= \int\int \mathbb{E}(Y_2|A_1 = a_1, A_2 = a_2^*, X = x, Y_1 = y_1)p(y_1|A_1 = a_1, X = x)dy_1 p(x)dx \\ &= \int\int \omega_2(a_1, Y_1, X)p(y_1|A_1 = a_1, X = x)dy_1 p(x)dx. \end{aligned}$$

The only part that involves a_1 is the integral

$$\int \omega_2(a_1, Y_1, X)p(y_1|A_1 = a_1, X)dy_1 = \mathbb{E}(\omega_2(A_1, Y_1, X)|A_1 = a_1, X)$$

so the optimal treatment will be

$$a_1^* = \mathsf{argmax}_{a_1} \mathbb{E}(\omega_2(A_1, Y_1, X)|A_1 = a_1, X).$$

Namely,

$$a_1^*(X) = \begin{cases} 1, & \text{if } \mathbb{E}(\omega_2(A_1, Y_1, X)|A_1 = 1, X) \geq \mathbb{E}(\omega_2(A_1, Y_1, X)|A_1 = 0, X) \\ 0, & \text{if } \mathbb{E}(\omega_2(A_1, Y_1, X)|A_1 = 1, X) < \mathbb{E}(\omega_2(A_1, Y_1, X)|A_1 = 0, X) \end{cases}. \tag{14.8}$$

As you can see, equation (14.8) is essentially a classifier. The difference is that here we do not have a label so the loss is measured by a 'random loss $-Y_1$' (Y_1 is called the reward in the reinforcement learning literature).

With equations (14.8) and (14.7), we obtain the optimal treatment regime $a_1^*(X)$ and $a_2^*(Y_1, a_1^*(X), X)$. You can easily generalize this idea to more time points using a similar derivation.

Note that during our derivation, we start with solving the last time points. This idea is called *dynamic programming* in computer science, which is how the term 'dynamic' appears in the dynamic treatment regime problem.

This approach is popular in precision medicine because the optimal treatment incorporates both the individual's background information (X) and the information available along the study (Y_1). The traditional medicine is the case where the treatment only uses the clinical information of a disease without considering X (and may not include Y_1 in the future treatment). The dynamic treatment regime provides a new approach to better treat each individual.

In practice, we will replace every 'expectation' by an estimated quantity. The estimator often relies on a study where individuals have been assigned to different treatments at different time points. In a sense, these individuals who contribute to the estimation procedure may not receive the optimal treatment. Sadly, this is unavoidable to obtain an estimator.

Modeling strategies. The construction of an optimal treatment regime becomes estimating the conditional expectations. This would be challenging when X is large (or when the number of treatment is large). One common model people used in practice is to assume a parametric model on the conditional expectation. In the case of using a linear model, we often assume that

$$\mathbb{E}(Y_2|A_1 = a_1, A_2 = a_2, X, Y_1) = \omega_{a_1,a_2} + \gamma Y_1 + (A_1\delta_1 + A_2\delta_2 + \beta)^T X,$$

where $\omega_{a_1,a_2}, \gamma \in \mathbb{R}$ and $\delta_1, \delta_2, \beta, X \in \mathbb{R}^d$. γ is the factor that determines how the outcome from the previous time point is correlated with the current outcome. The two vectors δ_1, δ_2 are the change of slope due to the treatments; they measure the interaction effect between the treatments and the background information. Note that we allow the slope to change with respect to A_1, A_2. The fact that the slope can change implies that the optimal decision will use information from X (you can show that if $\delta_1 = \delta_2 = 0$, the optimal decision rule a_2^* will not involve X). For the conditional density of Y_1 given A_1 and X, a common model is to assume a normal distribution with the mean changing with respect to A_1 and X.

Many time points. When there are many time points, say T, time points, we have many random variables:

$$X, A_1, Y_1, A_2, Y_2, A_3, Y_3, \cdots, A_T, Y_T.$$

There will be a total of $d + 2T$ variables (d is the number of variables in X). Even if d is small, the final treatment $A_2 = a_2(X, A_1, Y_1, \cdots, Y_{T-1})$ still relies on many variables. Estimating the optimal treatment will be a challenging task. A possible remedy to this problem is to introduce some conditional independence such that only the outcomes (and treatments) in the recent time points will affect the outcome at a specific time point. Namely, Y_t only depends on X and $\{(A_{t-k}, Y_{t-k}) : k = 1, 2, \cdots, s\}$ for some s.

Here are some useful references about the dynamic treatment regime:

1. Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2), 331-355.
2. Moodie, E. E., Richardson, T. S., & Stephens, D. A. (2007). *Demystifying optimal dynamic treatment regimes.* Biometrics, 63(2), 447-455.
3. Chakraborty, B., & Murphy, S. A. (2014). *Dynamic treatment regimes. Annual review of statistics and its application*, 1, 447-464.

Interestingly, the dynamic treatment regime is closely related to two important topics in machine learning: the *bandit* problem and the *reinforcement* learning. They both share similar theoretical techniques as the dynamic treatment regime problems but the constraints and contexts are different.

14.2 Conditional average treatment effect (CATE)

The conditional average treatment effect (CATE) is a very popular topic due to its application in precision medicine (also known as personalized medicine). This concept is sometimes called heterogenous treatment effect. We will use the potential outcome model with a binary treatment A for the CATE.

In a conventional setup, the CATE is

$$\tau(x) = \mathbb{E}(Y(1) - Y(0)|X = x).$$

We will make the standard ignorability assumption in the binary treatment that

$$(Y(1), Y(0)) \perp A|X. \tag{14.9}$$

There are many ways for estimating the CATE. They rely on the following 'nuisances'

$$m_a(x) = \mathbb{E}(Y|A = a, X = x)$$
$$\pi_a(x) = P(A = a|X = x).$$

Here we describe some popular approaches.

14.2.1 S-Learner and T-Learner

S-Learner. The S-learner approach is the most straight foward approach. Under the ignorability assumption,

$$\begin{aligned}\tau(x) &= \mathbb{E}(Y(1) - Y(0)|X = x)\\ &= \mathbb{E}(Y(1)|X = x) - \mathbb{E}(Y(0)|X = x)\\ &= \mathbb{E}(Y|A = 1, X = x) - \mathbb{E}(Y|A = 0, X = x)\\ &= m_1(x) - m_0(x).\end{aligned}$$

Thus, S-learner's idea is to estimate the regression function

$$m_a(x) = \mathbb{E}(Y|A = a, X = x)$$

by regressing Y on both X and A to obtain $\widehat{m}_a(x)$ and then use the difference $\widehat{m}_1(x) - \widehat{m}_0(x)$ as the estimator.

T-Learner. The idea of T-learner is similar to S-learner. However, it is based on the fact that $m_a(x)$ is essentially the regression model of Y on X under the sample with $A = a$. Thus, we first split the data into $\mathcal{D}_1, \mathcal{D}_0$ where $\mathcal{D}_a$ is those observation with $A = a$. For $\mathcal{D}_1$, we regress Y on X and obtain $\widehat{m}_1(x)$ and similarly we obtain $\widehat{m}_0(x)$ by fitting another regression model on $\mathcal{D}_0$. The difference is then the estimator of CATE.

As you can see, the S-learner and T-learner are very similar. The only difference is that S-learner fit a single model jointly for both X and A while T-learner fit two separate models.

14.2.2 X-Learner

The X-learner is a modified version of the T-learner from the following paper:

[KSBY2019] Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. Proceedings of the national academy of sciences, 116(10), 4156-4165.

It is based on a localized version of the CATE:

$$\tau_a(x) = \mathbb{E}(Y - m_a(X)|X = x).$$

With the above quantity, you can show that the CATE can be written as

$$\tau(x) = \pi_0(x)\tau_1(x) + \pi_1(x)\tau_0(x).$$

The nuisance $\tau_a(x)$ can be obtained by regression $Y - m_a(X)$ versus X. The estimation procedure is summarized as follows.

1. Split the data into $\mathcal{D}_1, \mathcal{D}_0$ based on the treatment $A = a$.
2. Similar to T-learner, we obtain $\widehat{m}_a(x)$ by regressing Y to X under sample $\mathcal{D}_a$.
3. For observations in $\mathcal{D}_1$, we compute a pseudo-response $\tilde{Y}^{(1)} = Y - \widehat{\mu}_0(X)$. Similarly, we obtain $\tilde{Y}^{(0)} = \widehat{\mu}_1(X) - Y$ for $\mathcal{D}_0$. This avoids using the data twice.
4. We estimate $\widehat{\tau}_1(x)$ by regressing $\tilde{Y}^{(1)}$ on X in the sample $\mathcal{D}_1$ and similarly for $\widehat{\tau}_0(x)$
5. Estimate CATE by $\widehat{\tau}(x) = \widehat{\pi}_0(x)\widehat{\tau}_1(x) + \widehat{\pi}_1(x)\widehat{\tau}_0(x)$.

14.2.3 R-Learner and U-Learner

The R- and U-learners are built on a slightly different generative model. We start with a regression model on the potential outcomes:

$$Y(1) = m_1(X) + \epsilon_1(X), \qquad Y(0) = m_0(X) + \epsilon_0(X).$$

Also, using the fact that $Y = AY(1) + (1 - A)Y(0)$, we have

$$m(x) = \mathbb{E}(Y|X = x) = \pi_1(x)m_1(x) + (1 - \pi_1(x))m_0(x).$$

As a result, we obtain

$$\begin{aligned} Y - m(X) &= AY(1) + (1 - A)Y(0) - \pi_1(X)m_1(X) - (1 - \pi_1(X))m_0(X) \\ &= Y(0) - m_0(X) + A(Y(1) - Y(0)) - \pi_1(X)(m_1(X) - m_0(X)) \\ &= \epsilon_0(X) + (A - \pi_1(X))\tau(X) + A(\epsilon_1(X) + \epsilon_0(X)) \\ &= (A - \pi_1(X))\tau(X) + \epsilon(X, A), \end{aligned}$$

where $\epsilon(X, A) = \epsilon_0(X) + A(\epsilon_1(X) + \epsilon_0(X))$ is a conditional mean 0 quantity.

The equation

$$Y - m(X) = (A - \pi_1(X))\tau(X) + \epsilon(X, A) \tag{14.10}$$

is the key to the R- and U-learners. Both learners will first use the whole data to estimate $\widehat{m}(x)$ and $\widehat{\pi}_1(x)$, which can be done by any conventional approaches. Equation (14.10) is first appeared in

Robinson, P. M. (1988). Root-N-consistent semiparametric regression. Econometrica: Journal of the Econometric Society, 931-954.

R-learner. Because both $Y - m(X)$ and $A - \pi_1(X)$ can be estimated from the data. The R-learner tries to find $\tau(X)$ by minimizing the following R-learning loss function:

$$L_n(f) = \frac{1}{n}\sum_{i=1}^{n}\left[(Y_i - \widehat{m}(X_i)) - (A_i - \widehat{\pi}_1(X_i))f(X_i)\right]^2.$$

We can also add a penalization to the above loss function to regularize the estimator as well. See the following paper for more details:

> Nie, X., & Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. Biometrika, 108(2), 299-319.

As a concrete example, we may model f as a linear function and use L_1 penalty, which leads to

$$\widehat{\beta}_\lambda = \mathsf{argmin}_\beta \frac{1}{n}\sum_{i=1}^{n}\left[(Y_i - \widehat{m}(X_i)) - (A_i - \widehat{\pi}_1(X_i))\beta^T X_i\right]^2 + \lambda\|\beta\|_1.$$

The CATE estimator is $\widehat{\beta}_\lambda^T x$. Note that here we include the constant 1 in the covariate (intercept term of the linear model).

U-learner. U-learner is a method proposed in [KSBY2019]. Instead of viewing the problem as a minimization problem, the U-learner further define

$$U = \frac{Y - m(X)}{A - \pi_1(X)}$$

and use the fact that $\mathbb{E}(U|X) = \tau(X)$. Thus, based on the nuisance estimators $\widehat{m}(x)$ and $\widehat{\pi}_1(x)$, we then compute

$$\widehat{U}_i = \frac{Y_i - \widehat{m}(X_i)}{A_i - \widehat{\pi}_1(X_i)}$$

and then regress $\widehat{U}_i$ with X_i to obtain $\widehat{\tau}(x)$.

14.3 Graphical models

Some useful references:

- http://mlg.eng.cam.ac.uk/zoubin/tut06/cambridge_causality.pdf
- https://stat.ethz.ch/~mmarloes/meetings/slides3a.pdf

In the previous lecture, we have seen that the DAG (directed acyclic graph) can be used as an elegant tool for representing the underlying causal relationship among variables. Here we discuss an popular method to define the causal effect in a DAG using the *do operator*.

Suppose that we have several variables $V_1, \cdots, V_d$ of interest and we use the DAG to specify the underlying generating model (Bayesian network). To simplify the problem, suppose that we are interested in estimating the causal effect V_1 on the variable V_2. We often relabel the variables as $V_1 = X$, $V_2 = Y$, and make the rest of them as $Z_1, \cdots, Z_m$, where $m = d - 2$. With this notation, the parameter of interest is the causal effect from X on Y. Let $p(x, y, z_1, \cdots, z_m)$ be the joint PDF (it can be generalized to PMF as well) and

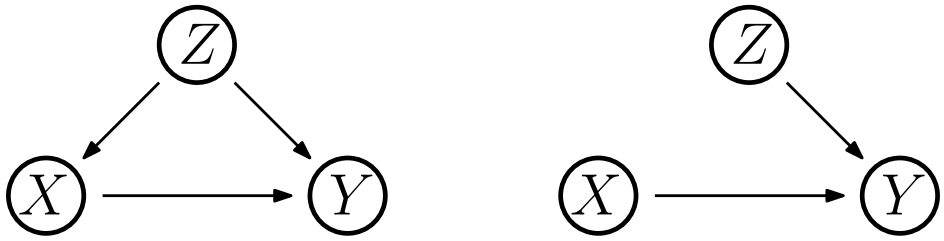

Figure 14.1: **Left:** original DAG G. **Right:** the DAG after the $\mathbf{do}(x)$ operation.

$G = (V, E)$ where $V = (V_1, \cdots, V_d) = (X, Y, Z_1, \cdots, Z_m)$ and $E_{ij} = 1$ if there is a directed arrow from node i to node j. The DAG implies

$$p(x, y, z_1, \cdots, z_m) = p(x|\mathsf{PA}_x)p(y|\mathsf{PA}_y)\prod_{j=1}^{m} p(z_j|\mathsf{PA}_{z_j}), \tag{14.11}$$

where PA_v denotes the set of parent nodes of variable v.

Defining the causal effect from X on Y is not easy because they may be interacting with variables $Z_1, \cdots, Z_m$. The *do operator* provides a solution to this. The do operator defines the causal effect using the conditional PDF

$$p(y|\mathbf{do}(x)) \equiv p(y|\mathbf{do}(X = x)). \tag{14.12}$$

We often define $\tau(x) = \frac{\partial}{\partial x}\mathbb{E}(Y|\mathbf{do}(X) = x) = \frac{\partial}{\partial x}\int yp(y|\mathbf{do}(x))dy$ as the causal effect on Y from X. Note that in general,

$$p(y|\mathbf{do}(x)) \neq p(y|x)$$

except for the simple case where there is only an arrow $X \to Y$ and no other arrows toward Y.

The conditional PDF $p(y|\mathbf{do}(X = x))$ is interpreted as: *we change system in a way that the variable X is set to x, this leads to a density function of Y and this density function is $p(y|\mathbf{do}(x))$.*

Given a DAG $G = (V, E)$ where $V = (X, Y, Z_1, \cdots, Z_m)$, the do operation defines a new DAG $G' = (V, E) = G(\mathbf{do}(x)) = (V, E(\mathsf{do}(x)))$ such that *all directed arrows to X is removed.* This leads to a new factorization of the joint PDF:

$$p(\mathbf{do}(x), y, z_1, \cdots, z_m) = p(\mathbf{do}(x))p(y|\mathsf{PA}_y)\prod_{j=1}^{m} p(z_j|\mathsf{PA}_{z_j}) \tag{14.13}$$

or the corresponding conditional density

$$p(y, z_1, \cdots, z_m|\mathbf{do}(x)) = p(y|\mathsf{PA}_y)\prod_{j=1}^{m} p(z_j|\mathsf{PA}_{z_j}). \tag{14.14}$$

Equation (14.14) is known as *g-formula* (by J. Robins), or *truncated factorization formula* (by J. Pearl).

If we use the DAG to interpret the result, the new DAG G' preserves all causal effects except for the ones that are affecting X. This is exactly how we (commonly) think about the causal effect due to X–we keep the entire system as is except we add an intervention at variable X that sets it to be x.

The power of equation (14.14) is that the left-hand-side $p(y, z_1, \cdots, z_m|\mathbf{do}(x))$ is the conditional density due to the do operation $\mathbf{do}(x)$, which is a theoretical entity, and the right-hand-side is what we can identify using the original DAG. With equation (14.14), we can identify equation (14.12) using

$$p(y|\mathbf{do}(x)) = \int p(y, z_1, \cdots, z_m|\mathbf{do}(x))dz_1 \cdots dz_m = \int p(y|\mathsf{PA}_y)\prod_{j=1}^{m} p(z_j|\mathsf{PA}_{z_j})dz_1 \cdots dz_m$$

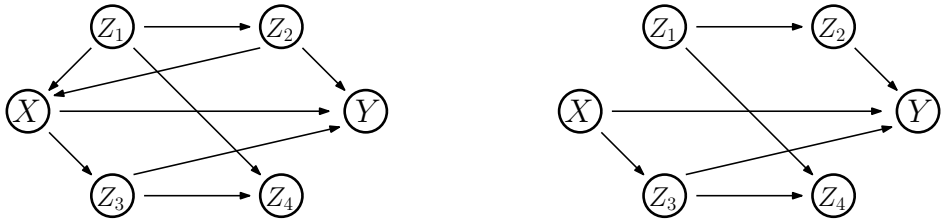

Figure 14.2: A more complicated DAG. **Left:** original DAG G. **Right:** the DAG after the $\mathbf{do}(x)$ operation.

and $\tau(x)$ accordingly.

Note that in this case, we often assume that the DAG is known so we can estimate all the conditional densities $p(z_j|\mathsf{PA}_{z_j})$ (and $p(y|\mathsf{PA}_y)$) using the data. Equation (14.14) shows that we can *identify* the causal effect from the data.

Example 1. In the example of Figure 14.1, we have an original DAG in the left and a new DAG due to $\mathbf{do}(x)$. The original DAG implies a factorization of the joint PDF

$$p(x, y, z) = p(x|z)p(y|x, z)p(z).$$

All the three conditionals can be estimated/identified from the data if we know this DAG in advance. Using the g-formula (equation (14.14)), the $\mathbf{do}(x)$ operation leads to a conditional density

$$p(y, z|\mathbf{do}(x)) = p(y|x, z)p(z),$$

which is still identifiable from the three conditionals provided in the original DAG.

Example 2. In Figure 14.2, we provide a more complicated example where there are 6 variables. The left panel displays the original DAG and the right panel displays the DAG after a $\mathbf{do}(x)$ operation. The original DAG implies the following factorization

$$p(x, y, z_1, z_2, z_3, z_4) = p(z_1)p(z_2|z_1)p(x|z_1, z_2)p(z_3|x)p(z_4|z_1, z_3)p(y|x, z_2, z_3).$$

All these conditionals are identifiable from the data. After the $\mathbf{do}(x)$ operation, the conditional density is

$$p(y, z_1, z_2, z_3, z_4|\mathbf{do}(x)) = p(z_1)p(z_2|z_1)p(z_3|x)p(z_4|z_1, z_3)p(y|x, z_2, z_3).$$

Each element in the right-hand-sided is identifiable so we can identify the entire conditional density. Note that if we are only interested in $p(y|\mathbf{do}(x))$, we can write it as

$$\begin{aligned} p(y|\mathbf{do}(x)) &= \int p(y, z_1, z_2, z_3, z_4|\mathbf{do}(x))dz_1dz_2dz_3dz_4 \\ &= \int p(y|x, z_2, z_3)p(z_1)p(z_2|z_1)p(z_3|x)\left(\int p(z_4|z_1, z_3)dz_4\right)dz_1dz_2dz_3 \\ &= \int p(y|x, z_2, z_3)p(z_1)p(z_2|z_1)p(z_3|x)dz_1dz_2dz_3. \end{aligned}$$

So we only need to estimate these 4 conditionals. In a sense, we do not need to consider estimating any effect of Z_4 (since it does not have a causal effect onto Y). Using the DAG induced by the do operator, we have

$$Z_3 \perp Z_1, Z_2|X, \quad Y \perp X, Z_1|Z_2, Z_3, \quad Z_2 \perp X|Z_1,$$

we can further write the above equality as

$$\begin{aligned}
p(y|\mathbf{do}(x)) &= \int p(y|x,z_2,z_3)p(z_1)p(z_2|z_1)\ \underbrace{p(z_3|x)}_{=p(z_3|z_2,x)}\ dz_1dz_2dz_3 \\
&= \int p(y,z_3|x,z_2)p(z_1)p(z_2|z_1)dz_1dz_2dz_3 \\
&= \int \underbrace{p(y|x,z_2)}_{=p(y|x,z_1,z_2)}\ p(z_1)\ \underbrace{p(z_2|z_1)}_{=p(z_2|z_1,x)}\ dz_1dz_2 \\
&= \int p(y,z_2|x,z_1)p(z_1)dz_2dz_1 \\
&= \int p(y|x,z_1)p(z_1)dz_1.
\end{aligned}$$

The conditional density due to the do operator is essentially the conditional density after *adjusting* $p(z_1)$ so variable Z_1 is called *adjustment set*; see Definition 3.6 of the following paper:

> Maathuis, M. H., & Colombo, D. (2015). A generalized back-door criterion. *The Annals of Statistics*, 43(3), 1060-1088.

The adjustment set has offers an elegant way to further simplify the g-formula–the conditional density of Y given the do operator $\mathbf{do}(x)$ is the same as we adjust the conditional density of Y given X and the variables in the adjustment sets. In a sense, the adjustment set represents the possible sources of interaction from other variables onto the causal effect from X onto Y. So we have to adjust for these variables to obtain the desired causal effect.

Note that all the above analysis is relied on the fact that we know the DAG in advance. This is possible if we have additional scientific knowledge about each variable. However, in a general observational study, all we can estimate (using the data) is the conditional independence, which is an undirected graph. We may have some partial knowledge about each edge, leading to a mixed graph (a graph with some directed and some undirected edges). Here are some papers related to finding the adjustment sets for different types of graphs (different types of graphs representing situations where we have different prior knowledge about the relations among variables):

> 1. Perković, E., Textor, J., Kalisch, M., & Maathuis, M. H. (2018). Complete graphical characterization and construction of adjustment sets in Markov equivalence classes of ancestral graphs. *The Journal of Machine Learning Research*, 18(1), 8132-8193.
> 2. Perković, E., Textor, J., Kalisch, M., & Maathuis, M. H. (2015). A complete generalized adjustment criterion. In *Uncertainty in Artificial Intelligence* (pp. 682-691). AUAI Press.

Remark (Structural Equation Modeling). A popular method that uses the DAG to make inference is the structural equation modeling (SEM). In the simplest form, the SEM assumes a linear effect for every arrow in the DAG. Suppose $X \to Y$ and $Z \to X$ and $Z \to Y$, then an SEM will be

$$Y = \alpha_Y + \beta X + \gamma Z + \epsilon_Y, \quad X = \alpha_X + \eta Z + \epsilon_X$$

and $\epsilon_X, \epsilon_Y \sim N(0, \sigma^2)$ with $Z \sim p(z)$. If we are interested in the causal effect from X onto Y, β will be the parameter of interest. So the question is: how do we properly apply the regression to obtain a consistent estimate of β. In general, we need to observe X, Y, Z to properly estimate β (using multiple linear regression). Note that Z is called the confounder for the causal effect from X to Y.

If Z is unobserved (unobserved confounder problem), then we cannot identify the causal effect β. However, IV (instrumental variable) offers a solution to this problem. Suppose that we do not observe Z (so we cannot identify the causal effect β) but we observe another variable U such that there is an arrow $U \to X$ and no other arrows related to U. This variable U is an IV (formally, it is called a valid IV). Suppose again the linear effect and modify X as

$$X = \alpha_X + \eta Z + \xi U + \epsilon_X.$$

Putting this into Y, we obtain

$$\begin{aligned} Y &= \alpha_Y + \beta(\alpha_X + \eta Z + \xi U + \epsilon_X) + \gamma Z + \epsilon_Y \\ &= \alpha' + \eta' Z + \beta\xi U + \epsilon'. \end{aligned}$$

Thus, regressing Y with U leads to a slope $\beta\xi$. Regressing X with U yields the slope ξ. So we can estimate β by the ratio of the two regression coefficient even if we do not observe the confounder Z.

14.4 Continuous treatment

When the treatment variable A is continuous or contains infinite amount of possible values, the causal inference becomes very challenging. In particular, the potential outcomes $Y(a)$ will be hard to characterize since there could be many of them,

Let $Y \in \mathbb{R}$ be the outcome of interest, $A \in \mathbb{R}$ be the treatment variable, and $X \in \mathbb{R}^d$ be the confounders. Similar to the conventional setup, our data is the collection of

$$(Y_1, A_1, X_1), \cdots, (Y_n, A_n, X_n).$$

In the continuous treatment problem, the parameter of interest is the *dose-response curve*

$$m(a) = \mathbb{E}(Y(a))$$

and we often work with the following conditions:

- **(C1: consistency)** Conditioned on $A = a$, $Y = Y(a)$.
- **(C2: ignorability)** $Y(a) \perp A|X$ for all a.

These two conditions are essentially the same conditions as in the binary treatment case.

Under (C1-2), we can rewrite the dose-response curve as the following form:

$$\begin{aligned} m(a) &= \mathbb{E}(\mathbb{E}(Y|A = a, X)) \\ &= \mathbb{E}(\mu(a, X)), \end{aligned} \tag{14.15}$$

where

$$\mu(a, x) = \mathbb{E}(Y|A = a, X = x)$$

is the regression function. To estimate the dose-response curve, there are three popular approaches.

14.4.1 Direct nonparametric estimation

Regression adjustment. Equation (14.15) suggests a plug-in approach to estimate $m(a)$:

$$\widehat{m}(a) = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mu}(a, X_i),$$

where $\widehat{\mu}(a, x)$ is a regression estimator, which is the nuisance parameter in this case. This is essentially a regression adjustment method.

Inverse probability weighting. While it is not easy to think about the IPW approach for continuous treatment because the conventional IPW uses the indicator function $I(A = a)$, which will almost always be 0, it is possible to use a *kernelized* method to construct an IPW estimator. The IPW (kernel smoothing) method is

$$\widehat{m}_h(a) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{A_i - a}{h}\right) \frac{Y_i}{\widehat{p}(a|X_i)},$$

where $\widehat{p}(a|x)$ is the estimated conditional PDF of $A = a$ given $X = x$ and K is a smoothing kernel and $h > 0$ is a smoothing bandwidth. Essentially, we replace the indicator function $I(A = a)$ by its kernelized version $\frac{1}{h}K\left(\frac{A_i - a}{h}\right)$. The probability weighting (propensity score) is replaced by the conditional PDF $p(a|x)$. More details about this idea can be found in

> Huber, M., Hsu, Y. C., Lee, Y. Y.,& Lettry, L. (2020). Direct and indirect effects of continuous treatments based on generalized propensity score weighting. Journal of Applied Econometrics, 35(7), 814-840

You can also combine both estimators to form a double-robust estimator as well.

Positivity conditions. In both regression adjustment and IPW, the positivity conditions become very crucial. The two nuisance $\widehat{\mu}(a, x)$ and $\widehat{p}(a|x)$ will be evaluated at every X_i with $A = a$. Thus, both estimators rely heavily on the uniform consistency of the two nuisances on the region $\{a\} \times \mathsf{supp}(X)$. It is possible to bypass the positivity condition using a method called the *integral estimator*, see the following paper:

> Zhang, Y., Chen, Y. C., & Giessing, A. (2024). Nonparametric Inference on Dose-Response Curves Without the Positivity Condition. arXiv preprint arXiv:2405.09003.

14.4.2 Pseudo-outcome approach

Another approach to estimate the dose-response curve is the pseudo-outcome approach introduced in the following paper:

> Kennedy, E. H., Ma, Z., McHugh, M. D., & Small, D. S. (2017). Non-parametric methods for doubly robust estimation of continuous treatment effects. Journal of the Royal Statistical Society Series B: Statistical Methodology, 79(4), 1229-1245.

A challenge of estimating the dose-response curve is that the parameter of interest $m(a)$ is not pathwise differentiable, making the analysis on its efficient influence function very challenging. However, it is still possible to construct an estimator with doubly-robustness property by the use of pseudo-outcomes.

Let $\widehat{\mu}(a,x), \widehat{p}(a|x)$ be the estimated regression and conditional PDF as in the previous sections. The pseudo-outcome of (Y,A,X) is

$$\widehat{Y} = \frac{Y - \widehat{\mu}(A,X)}{\widehat{p}(A|X)} \cdot \frac{1}{n}\sum_{j=1}^{n} \widehat{p}(A|X_j) + \frac{1}{n}\sum_{j=1}^{n} \widehat{\mu}(A, X_j).$$

Applying this to every observations, we obtain

$$\widehat{Y}_1, \cdots, \widehat{Y}_n.$$

Then we estimate $m(a)$ by applying a nonparametric regression of $\widehat{Y}$ versus A.

The pseudo-outcome approach enjoys the doubly-robust property–we just need either $\widehat{\mu}$ or $\widehat{p}$ to be correctly specified to obtain consistency. However, this method still heavily relies on the positivity condition.

14.4.3 Marginal structural modeling

The marginal structural modeling is a popular approach to deal with problems involving multiple or even infinite potential outcomes. The idea is very simple: we put a parametric model on the dose-response curve, i.e.,

$$m(a) \equiv \theta(a;\gamma),$$

where γ is the underlying parameter.

Examples:

1. *Linear Model.* $\mathbb{E}(Y(a)) = \theta(a;\gamma) = \gamma_0 + \gamma_1^T a$.
2. *Log-linear model.* $\log \mathbb{E}(Y(a)) = \gamma_0 + \gamma_1^T a$ or equivalently,
$$\theta(a;\gamma) = \exp(\gamma_0 + \gamma_1^T a).$$
3. *Logistic model.* For $Y \in [0,1]$, or a binary Y, we may use $\log\left\{\frac{\mathbb{E}(Y(a))}{1-\mathbb{E}(Y(a))}\right\} = \gamma_0 + \gamma_1^T a$ or equivalently,
$$\theta(a;\gamma) = \frac{\exp(\gamma_0 + \gamma_1^T a)}{1 - \exp(\gamma_0 + \gamma_1^T a)}.$$

To estimate the underlying parameter γ, there are two possible approaches–inverse probability weighting and regression adjustment.

Inverse probability weighting. The IPW approach estimates $\widehat{\gamma}$ via the following procedure:

1. Estimate $\widehat{p}(a), \widehat{p}(a|x)$ from $(A_1,X_1),\cdots,(A_n,X_n)$.
2. Construct the estimating equation
$$\widehat{\Psi}(\gamma; Y, A, X) = \frac{\widehat{p}(A)}{\widehat{p}(A|X)}(Y - \theta(A;\gamma))s(A;\gamma), \qquad s(a;\gamma) = \frac{\partial}{\partial\gamma}\theta(a;\gamma). \tag{14.16}$$
3. Find $\widehat{\gamma}_n$ by solving
$$0 = \frac{1}{n}\sum_{i=1}^{n} \widehat{\Psi}(\widehat{\gamma}_n; Y_i, A_i, X_i).$$

4. Obtain the estimator $\widehat{\theta}(a) = \theta(a; \widehat{\gamma}_n)$.

Essentially, the IPW is derived from the usual M-estimator when we assume a parametric model with inverse probability weighting. If we ignore $\widehat{p}(A)$ in equation (14.16), the whole estimating equation behaves just like the score equation. The multiplier $\widehat{p}(A)$ is to stabilize the estimation procedure, see the following note for more discussion: https://faculty.washington.edu/yenchic/short_note/note_msmc.pdf.

Regression adjustment (g-computation). Alternatively, we may put a parametric model of $m(a)$ indirectly from a model on $\mu(a, x)$ and estimate it accordingly. Specifically, we place a model $\mu(a, x; \beta)$ and the estimate β by a least square approach, leading to $\widehat{\beta}$. We can then obtain an estimator of $m(a)$ accordingly.

Here is what we will do in practice. We first compute the estimator

$$\widehat{\beta}_n = \mathsf{argmin}_\beta \frac{1}{n} \sum_{i=1}^{n} (Y_i - \mu(A_i, X_i; \beta))^2.$$

Then we construct the estimator of the MSM via

$$\widehat{\theta}(a) = \frac{1}{n} \sum_{i=1}^{n} \mu(a, X_i; \widehat{\beta}_n).$$