STAT 535: Statistical Machine Learning **Autumn 2019**

Lecture 12: Causal inference and missing data

Instructor: Yen-Chi Chen

Please read Section 12.3 on your own; it is from an undergraduate course (STAT 403).

12.1 Causal inference: 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

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

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

$$(Y(0), Y(1)) \perp A. \tag{12.1}$$

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

$$\widehat{\tau}_{\mathsf{naive}} \overset{P}{\to} \tau.$$

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.

12.1.1 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,

$$(Y(0), Y(1)) \perp A|X. \tag{12.2}$$

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

$$\begin{aligned}\tau = \mathbb{E}(Y(1)) - \mathbb{E}(Y(0)) &= \mathbb{E}(\mathbb{E}(Y(1)|X)) - \mathbb{E}(\mathbb{E}(Y(0)|X))\\ &= \mathbb{E}(\mathbb{E}(Y|A=1,X)) - \mathbb{E}(\mathbb{E}(Y|A=0,X)).\end{aligned} \tag{12.3}$$

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 (12.3) 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 (12.2) 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 $\mathbb{E}(Y(a))$:

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

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

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

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 $\mathbb{E}(Y(a)) = \mathbb{E}\left\{\frac{1}{\pi_a(X)}YI(A = a)\right\}$, 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.

12.1.2 Local average treatment effect

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

$$\text{(Exclusion Restriction)} \quad Y(a,z) = Y(a,z') \quad \text{for all } a, z, z'. \tag{12.4}$$

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

$$\text{(Randomization)} \quad Z \perp Y(0), Y(1), A(0), A(1). \tag{12.5}$$

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, the LATE can be written as

$$\tau_{\mathsf{LATE}} = \frac{\mathbb{E}(Y|Z=1) - \mathbb{E}(Y|Z=0)}{\mathbb{E}(A|Z=1) - \mathbb{E}(A|Z=0)}, \tag{12.6}$$

and we can easily each expectation using conditional mean.

12.1.3 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:

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

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

$$a_2^*(Y_1, a_1, X) = \begin{cases} 1, & \text{if } \mathbb{E}(Y_2|A_1 = a_1, A_2 = 1, X, Y_1) \geq \mathbb{E}(Y_2|A_1 = a_1, A_2 = 0, X, Y_1) \\ 0, & \text{if } \mathbb{E}(Y_2|A_1 = a_1, A_2 = 1, X, Y_1) < \mathbb{E}(Y_2|A_1 = a_1, A_2 = 0, X, Y_1) \end{cases}. \tag{12.7}$$

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,

$$a_1^*(X) = \begin{cases} 1, & \text{if } \mathbb{E}(\omega_2(A_1, Y_1, X)|A_1 = 1, X) \geq \mathbb{E}(\omega_2(A_1, Y_1, X)|A_1 = 0, X) \\ 0, & \text{if } \mathbb{E}(\omega_2(A_1, Y_1, X)|A_1 = 1, X) < \mathbb{E}(\omega_2(A_1, Y_1, X)|A_1 = 0, X) \end{cases}. \tag{12.8}$$

As you can see, equation (12.8) 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 (12.8) and (12.7), 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.

12.2 Causal inference: the "do" operator

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 $G = (V, E)$ 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

$$p(x, y, z_1, \cdots, z_m) = p(x|\mathsf{PA}_x)p(y|\mathsf{PA}_y)\prod_{j=1}^{m} p(z_j|\mathsf{PA}_{z_j}), \tag{12.9}$$

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

$$p(y|\mathbf{do}(x)) \equiv p(y|\mathbf{do}(X = x)). \tag{12.10}$$

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:

$$p(\mathbf{do}(x), y, z_1, \cdots, z_m) = p(\mathbf{do}(x))p(y|\mathsf{PA}_y)\prod_{j=1}^{m} p(z_j|\mathsf{PA}_{z_j}) \tag{12.11}$$

or the corresponding conditional density

$$p(y, z_1, \cdots, z_m|\mathbf{do}(x)) = p(y|\mathsf{PA}_y)\prod_{j=1}^{m} p(z_j|\mathsf{PA}_{z_j}). \tag{12.12}$$

Equation (12.12) 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 (12.12) 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 (12.12), we can identify equation (12.10) using

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

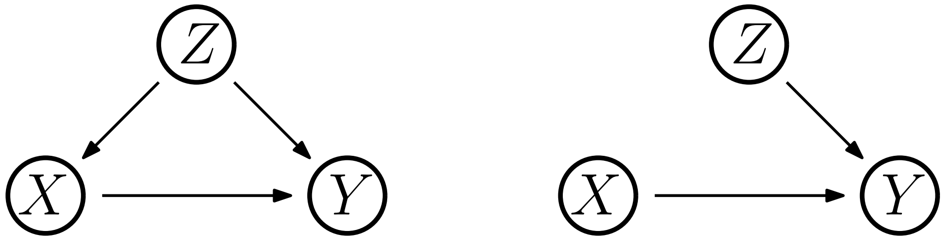

Figure 12.1: **Left:** original DAG G. **Right:** the DAG after the $\mathbf{do}(x)$ operation.

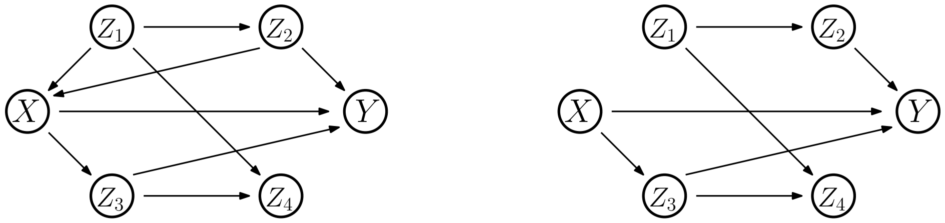

Figure 12.2: A more complicated DAG. **Left:** original DAG G. **Right:** the DAG after the $\mathbf{do}(x)$ operation.

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 (12.12) shows that we can *identify* the causal effect from the data.

Example 1. In the example of Figure 12.1, 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 (12.12)), 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 12.2, 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

$$\begin{aligned}
p(y|\mathbf{do}(x)) &= \int p(y, z_1, z_2, z_3, z_4|\mathbf{do}(x)) dz_1 dz_2 dz_3 dz_4 \\
&= \int p(y|x, z_2, z_3) p(z_1) p(z_2|z_1) p(z_3|x) \left(\int p(z_4|z_1, z_3) dz_4 \right) dz_1 dz_2 dz_3 \\
&= \int p(y|x, z_2, z_3) p(z_1) p(z_2|z_1) p(z_3|x) dz_1 dz_2 dz_3.
\end{aligned}$$

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

$$\begin{aligned}
p(y|\mathbf{do}(x)) &= \int p(y|x, z_2, z_3) p(z_1) p(z_2|z_1) \underbrace{p(z_3|x)}_{=p(z_3|z_2,x)} dz_1 dz_2 dz_3 \\
&= \int p(y, z_3|x, z_2) p(z_1) p(z_2|z_1) dz_1 dz_2 dz_3 \\
&= \int \underbrace{p(y|x, z_2)}_{=p(y|x,z_1,z_2)} p(z_1) \underbrace{p(z_2|z_1)}_{=p(z_2|z_1,x)} dz_1 dz_2 \\
&= \int p(y, z_2|x, z_1) p(z_1) dz_2 dz_1 \\
&= \int p(y|x, z_1) p(z_1) dz_1.
\end{aligned}$$

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.

12.3 Missing data: introduction and simple cases

Missing data is a very common problem in every scientific research. In a survey sample, it occurs when there are individuals who refuse to answer some questions. In a medical research, it happens when participants drop out of the study.

There are three common strategies that practitioners are using to handle missing data:

- **Complete-case analysis (ignoring observations with missing entries).** The complete-case analysis removes any observations that contain one or more missing entries. When the proportion of missing is small (and the missingness is irrelevant to any variables, including the one that can be missing), this is an okay procedure. But in general, this would lead to a biased estimate.

  To see this, think about estimating the average income of a city from a social survey. Many rich people would refuse to provide their incomes (it will be easy to identify them), leading to missing entries. In this scenario, if we ignore those individuals whose income is missing, we will get a biased estimate of the average income.

  In an observation study in medical research, sometimes people would perform analysis by adjusting the *inclusion criteria*: the criteria that determines which individual will be included in our analysis. In case that they require individuals to be fully observed, this is essentially a complete-case analysis.

- **Ignorable missingness (missing at random).** Another common approach is to make assumptions and choose a good model so that the missingness is ignorable. Note that in this case, we do NOT remove observations with missing entries–we still use their observed variables to construct our model. This is possible when we assume the missingness is missing at random (MAR; see Section 12.4) and use a proper parametric model.

However, MAR is just an assumption. It may be violated (which is often called missing not at random-MNAR). When the MAR is violated, it is often hard to obtain an ignorable missingness approach to deal with missing data. Note that sometimes we are still able to construct an ignorable missingness procedure using the selection model and inverse probability weighting estimator (see Section 12.5.1).

- **Imputation.** The imputation is another popular approach that practitioners used in solving missing data problem. The idea is very simple: we impute the missing entries with a proper value that leads to a complete dataset. Then we can treat the problem as if there is no missingness.

  Here is a caveat. If the imputation is done in a deterministic way, i.e., every time a missing entry is imputed, it always be imputed with a fixed number, the imputed data is often problematic because we do not take into account the intrinsic variation of that missing value. This would lead to bias in the later estimation procedure.

  A better approach is to use a stochastic imputation that we impute the missing entries by drawing from a distribution. Later we will show that if the distribution being drawn is the actual distribution that generates the data, the stochastic imputation leads to a dataset without any bias (Section 12.3.2).

  A challenge here is that in general, we do not know the actual distribution so how do we perform the stochastic imputation is a problem.

12.3.1 Simple cases

Consider a regression problem where we have a binary covariate $X \in \{0, 1\}$ and a continuous response $Y \in \mathbb{R}$. However, in our data, some response variables are missing and only the covariates are observed. So our data can be represented as

$$(X_1, Y_1), \cdots, (X_n, Y_n), (X_{n+1}, \star), \cdots, (X_{n+m}, \star).$$

The symbol $\star$ denotes a missing value. Namely, we have n observations that are fully observed while the other m observations that we only observe the covariate, not the response. Suppose that the parameter of interest is the marginal median of the response variable m_Y. How should we estimate the median?

We can introduce an additional variable R to denote the missingness such that $R = 0$ means that Y is not observed whereas $R = 1$ means that Y is observed. Note that R itself is another random variable.

Without any assumptions on the missing data, we are not able to accurately estimate the median consistently. There are two common assumptions people made about the missingness:

1. **MCAR:** missing completely at random. This means that the missingness is independent of any variables. Under the above notations, MCAR means that

$$R \perp X, Y.$$

2. **MAR:** missing at random. Under MAR, the missingness depends only on the observed pattern. In our case.

$$P(R = 0|X, Y) = P(R = 0|X)$$

   since Y is not observed when $R = 0$.

When the missingness is neither MCAR nor MAR, it is called MNAR–missing completely at random.

Under MCAR, we can completely ignore the data with missing values and just use the sample median as an estimate of m_Y. However, under MAR, we cannot do such thing because the missingness may depends on X and if the distribution of covariate is different under fully observed data ($R = 1$) and partially observed data ($R = 0$), we will obtain a biased estimate.

While there are other ways to estimate the median under MAR, we will focus on the method of imputation.

12.3.2 Imputation

The idea of imputation is to impute a value to the missing entry so that after imputing all missing entries, we obtain a data without any missingness. Then we can simply apply a regular estimator (in the above example, sample median) to estimate the parameter of interest.

However, we cannot impute any number to the missing entry because this would cause bias in the estimation. We need to impute the value in a smart way. Generally, we want to impute the value according to the conditional density

$$p(y|x, R = 0),$$

the conditional density of response variable Y given the covariate X and the missing pattern $R = 0$. Namely, for $n + i$-th observation where only X_{n+i} is observed, we want to draw a random number

$$\tilde{Y}_{n+i} \sim p(y|X_{n+i}, R = 0).$$

If indeed Y_{n+1} is from the above density function, one can show that the sample median

$$\text{median}\{Y_1, \cdots, Y_n, \tilde{Y}_{n+1}, \cdots, \tilde{Y}_{n+m}\}$$

is an unbiased estimator of m_Y.

This idea works regardless of what missing assumption is. However, the problem is that the density function $p(y|x, R = 0)$ cannot be estimated using our data because the only case we observed Y is when $R = 1$.

Under this case, MAR implies a powerful result:

$$p(y|x, R = 0) = p(y|x, R = 1). \tag{12.13}$$

Namely, the conditional density of Y given X is independent of the missing indicator R. To see how equation (12.13) is derived, note that MAR implies

$$P(R = 1|X, Y) = 1 - P(R = 0|X, Y) = 1 - P(R = 0|X) = P(R = 1|X).$$

Thus, the conditional density

$$\begin{aligned}
p(y|x,R=0) &= \frac{p(y,x,R=0)}{P(x,R=0)}\\
&= \frac{p(x,y)P(R=0|x,y)}{P(x,R=0)}\\
&= p(x,y)\frac{P(R=0|x)}{P(x,R=0)}\\
&= p(x,y)\frac{1}{p(x)}\\
&= p(x,y)\frac{P(R=1|x)}{P(x,R=1)}\\
&= \frac{p(x,y)P(R=1|x,y)}{P(x,R=1)}\\
&= \frac{p(y,x,R=1)}{P(x,R=1)}\\
&= p(y|x,R=1).
\end{aligned}$$

Thus, we obtain equation (12.13).

The power of equation (12.13) is that $p(y|x,R=1)$ can be estimated by a KDE:

$$\begin{aligned}
\widehat{p}(y|x,R=1) &= \frac{\frac{1}{nh}\sum_{i=1}^n K\left(\frac{Y_i-y}{h}\right)I(X_i=x)}{\frac{1}{n}\sum_{i=1}^n I(X_i=x)}\\
&= \frac{1}{n_x h}\sum_{i=1}^n K\left(\frac{Y_i-y}{h}\right)I(X_i=x),
\end{aligned}$$

where $n_x=\sum_{i=1}^n I(X_i=x)$ is the number of $X_i=x$ in the completely observed data and $x\in\{0,1\}$. Namely, $\widehat{p}(y|x,R=1)$ is the KDE applied to the completely observed data with the covariate $X=x$.

Given an observation $X_{n+i}=x$, how should we sample $\widehat{Y}_{n+i}$ from $\widehat{p}(y|x,R=1)$? It is very simple. We first sample the index I such that

$$P(I=i|\mathsf{data}) = \frac{1}{n_x}I(X_i=x).$$

Namely, I is chosen from those fully observed data with the covariate $X_i=x$ with equal probability. Given I we then sample Y_{n+i} from the density function

$$q(y)=\frac{1}{h}K\left(\frac{Y_I-y}{h}\right).$$

Although this may look scary, if the kernel function is Gaussian, $q(y)$ is the normal density with mean Y_I and variance h^2. Namely, when K is a Gaussian,

$$\widehat{Y}_{n+i}\sim N(Y_I,h^2).$$

Remark.

- The use of KDE is just one example. You can use any density estimator for $\widehat{p}(y|x,R=1)$ as long as you are able to sample from it.

- The equation (12.13) relies on the MAR assumption along with the fact that only one variable is subject to missing. When there are more than one variables that can be missing, we no longer have such a simple equivalence.
- The imputed data can be used for other estimators as well, not limited to estimating the median. You may notice that during our imputation process, we do not use any information about the estimator.
- There imputation methods that only imputes a fixed, non-random number for each missing entries. This is often called a deterministic imputation. For certain problem, a deterministic imputation works but in general, it may not work. So a rule thumb is to use a random imputation if possible.

12.3.3 Multiple imputation

After doing the imputation for all missing entries, we obtain a complete data

$$(X_1, Y_1), \cdots, (X_n, Y_n), (X_{n+1}, \widehat{Y}_{n+1}), \cdots, (X_{n+m}, \widehat{Y}_{n+1}).$$

The estimate of m_Y is just the sample median of this impute dataset. However, there will be Monte Carlo errors in this estimator because every time we do the imputation, we will not get the same number (due to sampling from $p(y|x, R = 0)$). If we just impute the data once (this is often called *single imputation*), we may suffer from the Monte Carlo errors a lot. Thus, a better approach is to perform a *multiple imputation*.

Multiple Imputation.[1] After obtaining a complete data, we do the same imputation procedure again, which gives us another new complete data. Then we keep repeating the above process, leading to several complete data, which can be represented as

$$\begin{aligned}
&(X_1, Y_1), \cdots, (X_n, Y_n), (X_{n+1}, \widehat{Y}^{(1)}_{n+1}), \cdots, (X_{n+m}, \widehat{Y}^{(1)}_{n+1})\\
&(X_1, Y_1), \cdots, (X_n, Y_n), (X_{n+1}, \widehat{Y}^{(2)}_{n+1}), \cdots, (X_{n+m}, \widehat{Y}^{(2)}_{n+1})\\
&\cdots\\
&(X_1, Y_1), \cdots, (X_n, Y_n), (X_{n+1}, \widehat{Y}^{(N)}_{n+1}), \cdots, (X_{n+m}, \widehat{Y}^{(N)}_{n+1}).
\end{aligned}$$

We then combine all these datasets to a huge dataset and compute the estimator of the parameter of interest (in our case, median of the response variable). This estimator has a smaller Monte Carlo error.

12.4 Missing data: general problems and missing at random

When there are more than one variable that are subject to missing, the problem gets a lot more complex. Consider the case where each individual has d variables $X_1, \cdots, X_5$ and all of them may be missing and we may even have many of them missing at the same time. There are two categories of the missing patterns:

1. **Monotone missingness.** In this case, if X_t is missing, then X_s is also missing for any $s > t$. This occurs a lot in medical research due to *dropout* of the individuals. For instance, let X_t denote the BMI of an individual at year t. If this individual left the study at time point τ, then we only observe $X_1, \cdots, X_\tau$ from this individual. Any information beyond year τ is missing.
2. **Non-monotone missingness.** When the missing pattern is not monotone, it is called non-monotone missingness. The non-monotone missing data is a lot more challenging than monotone missing data

[1] For more introduction on this topic, see https://stats.idre.ucla.edu/stata/seminars/mi_in_stata_pt1_new/

because there are many possible missing pattern that can occur in the data. If there are d variables, them monotone missing data has d different missing patterns but the non-monotone case may have up to 2^d different missing patterns!

Let $R \in \{0,1\}^5$ be a multi-index set that denotes the observed pattern and we use the notation $X_R = (X_i : R_i = 1)$. For instance, $R = 11001$ means that we observe variable X_1, X_2, and X_5 and $X_{11001} = (X_1, X_2, X_5)$. Under this notation, the MAR assumption can be written as

$$P(R = r|X) = P(R = r|X_r),$$

namely, the probability of seeing a pattern $R = r$ only depends on the observed variable.

MAR is a very popular assumption that people often assumed in practice (although it may not be reasonable in some cases). However, under the non-monotone case, MAR tells us little about the missingness and it is actually not very to work with. Why is the MAR still so popular in practice?

There are two reasons for why MAR is so popular. The first reason is that in both monotone and non-monotone case, MAR makes the likelihood inference a lot easier. The second reason is that under monotone missing data problem, MAR provides an elegant way to identify the entire distribution function.

12.4.1 Likelihood inference with MAR

The MAR has a nice property called the *ignorability*, which holds in both monotone and non-monotone missingness. Consider the joint density function $p(x, r)$ of both variable of interest X and the missing pattern R. Recall that $X_R = (X_i : R_i = 1)$ are the observed variables under pattern R. We also denote $X_{\bar{R}} = (X_i : R_i = 0)$ as the missing variables.

We can then factorize it into

$$p(x, r) = P(R = r|X = x)p(x).$$

Suppose we use parametric models separately for both $P(R = r|X = x)$ and $p(x)$, leading to

$$p(x, r; \phi, \theta) = P(R = r|X = x; \phi)p(x; \theta) \overset{(MAR)}{=} P(R = r|X_r = x_r; \phi)p(x; \theta),$$

where θ is the parameter for modeling $p(x)$ and ϕ is the parameter for modeling the missing probability $P(R = r|X_r = x_r)$ (this separability of parameter together with MAR is often called *ignorability*). In our data, what we observe are (x_r, r) so we should integrate over the missing variables $x_{\bar{r}}$:

$$p(x_r, r; \phi, \theta) = \int p(x, r; \phi, \theta)dx_{\bar{r}} = P(R = r|X_r = x_r; \phi) \int p(x; \theta)dx_{\bar{r}}.$$

Thus, the log-likelihood function is

$$\begin{aligned}
\ell(\theta, \phi|x_r, r) &= \log P(R = r|X_r = x_r; \phi) + \log \int p(x; \theta)dx_{\bar{r}} \\
&= \ell(\phi|x_r, r) + \ell(\theta|x_r), \\
\ell(\phi|x_r, r) &= \log P(R = r|X_r = x_r; \phi) \\
\ell(\theta|x_r, r) &= \log \int p(x; \theta)dx_{\bar{r}}.
\end{aligned}$$

The above factorization is very powerful–it decouple the problem of estimating θ and the problem of estimating ϕ!

Namely, if we are only interested in the distribution of X, we do not even need to deal with ϕ. We just need to maximize $\ell(\theta|x_r)$. So finding the MLE of θ can be done without estimating the parameter ϕ, leading to a simple procedure.

EM algorithm. Estimating θ via maximizing $\ell(\theta|x_r)$ is often done via the EM algorithm. The EM algorithm is an iterative algorithm that finds a stationary point. It consists of two steps, an expectation step (E-step) and a maximization step (M). Given an initial guess of the parameter $\theta^{(0)}$, the EM algorithm iterates the following two steps until convergence ($t = 0, 1, 2, 3, \cdots$):

1. **E-steps.** Compute
$$Q(\theta; \theta^{(t)}|X_r) = \mathbb{E}(\ell(\theta|X); X_r, \theta^{(t-1)}) = \int \ell(\theta|, x_{\bar{r}}, X_r) p(x_{\bar{r}}|X_r; \theta^{(t)}) dx_{\bar{r}}.$$

2. **M-steps.** Update
$$\theta^{(t+1)} = \mathsf{argmax}_\theta Q(\theta; \theta^{(t)}|X_r).$$

Note that in practice, we have n observations so the Q function will be

$$Q_n(\theta; \theta^{(t)}) = \frac{1}{n} \sum_{i=1}^{n} Q(\theta; \theta^{(t)}|X_{i,R_i})$$

and the M-step will be

$$\theta^{(t+1)} = \mathsf{argmax}_\theta Q_n(\theta; \theta^{(t)}).$$

Under good conditions, the EM algorithm has the ascending property, i.e.,

$$\ell(\theta^{(t+1)}|X_r) \geq \ell(\theta^{(t)}|X_r),$$

and will converge to a stationary point. However, the problem is that the stationary point is not guarantee to be the global maximum (MLE). It could be a local mode or even a saddle point.

A good introduction on the EM algorithm and missing data is Section 8 of the following textbook:

> Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). John Wiley & Sons.

12.4.2 MAR under monotone case

Under the monotone missing problem, let T denotes the index of the last observed variable. Namely, the individual dropouts after time point T. We use the notation $X_{\leq t} = (X_1, \cdots, X_t)$. Then the MAR can be written as

$$P(T = t|X) = P(T = t|X_{\leq t}).$$

The above equation gives us a very powerful result–we can estimate the missing probability $P(T = t|X)$ for every $t = 1, \cdots, d$!

To see this, consider the case $t = 1$ so MAR implies

$$P(T = 1|X) = P(T = 1|X_1).$$

Note that $P(T > 1|X) = 1 - P(T = 1|X) = P(T \neq 1|X_1) = P(T > 1|X_1)$. Thus, we can estimate $P(T = 1|X_1)$ by comparing pattern $T = 1$ against $T > 1$ given the variable X_1, which is always observed. Thus, $P(T = 1|X)$ is estimatible. For $t = 2$, the MAR implies

$$P(T = 2|X) = P(T = 2|X_1, X_2).$$

Thus,

$$P(T > 2|X) = 1 - P(T = 2|X) - P(T = 1|X) = 1 - P(T = 2|X_1, X_2) - P(T = 1|X_1) = P(T > 2|X_1, X_2).$$

Again, we can compare the pattern $T = 2$ against $T > 2$ and estimate the probability $P(T = 2|X)$. We can keep doing this procedure, and eventually all missing probability $P(T = t|X)$ can be estimated.

For instance, if we are interested in estimating the parameter of interest $\rho = \mathbb{E}(\omega(X_1, \cdots, X_d))$, we can then use the IPW estimator[2] as in the causal inference problem:

$$\widehat{\rho} = \frac{1}{n\widehat{P}(T = d|X)} \sum_{i=1}^{n} \omega(X_{i,1}, \cdots, X_{i,p}) I(T_i = d),$$

where $\widehat{P}(T = d|X)$ is an estimate of $P(T = d|X)$. Similar to the case of causal inference, $P(T = t|X)$ is also called the propensity score.

Actually, MAR under monotone missingness is equivalent to the *available case missing value* (ACMV) assumption:

$$p(x_{t+1}|x_{\leq t}, T = t) = p(x_{t+1}|x_{\leq t}, T > t)$$

for every t. The right-hand side can be estimated by conditional KDE so the density function[3]

$$p(x_{>t}|x_{\leq t}, T = t) = \prod_{s=t}^{p-1} p(x_{s+1}|x_{\leq s}, T = s)$$

can be estimated under ACMV assumption. Why is the above density being estimatible so useful? This is because the joint density function has the following pattern mixture model formulation:

$$p(x) = \sum_{t=1}^{p} p(x, t) = \sum_{t=1}^{p} p(x_{>t} \mid x_{\leq t}, T = t) p(x_{\leq t} \mid T = t) p(T = t),$$

where both $p(x_{\leq t} \mid T = t)$ and $P(T = t)$ can be directly estimated using our data so what remains unknown is the density function $p(x_{>t} \mid x_{\leq t}, T = t)$. ACMV implies an estimator of this density function, so the entire joint density function can be estimated. The equivalence between MAR and ACMV is shown in

> Molenberghs, G., Michiels, B., Kenward, M. G., & Diggle, P. J. (1998). *Monotone missing data and pattern-mixture models.* Statistica Neerlandica, 52(2), 153-161.

12.5 Missing data: strategies for missing not at random

In MNAR, the missing data problem becomes a lot more complicated. There are two common strategies for handling MNAR–the selection models and the pattern mixture models approaches.

[2] See https://en.wikipedia.org/wiki/Inverse_probability_weighting for more details.

[3] Also called the *extrapolation density*.

To simplify the problem, we consider monotone missing data problem. Even in this scenario, we will see several identifiability issues so we have to be very careful about our choice of model.

Recall that X denotes the study variable and T is the dropout time. We are interesting in the *full-data density* $p(x,t)$; note that $p(x,t)$ implies the joint PDF of the study variable $p(x)$.

A useful reference: https://content.sph.harvard.edu/fitzmaur/lda/C6587_C018.pdf.

12.5.1 Selection models

Selection models decompose the full-data density using

$$p(x,t) = P(T=t|x)p(x),$$

where $P(T=t|x)$ is called the missing probability or missing data mechanism.

A common strategy in selection model is to identify $P(T=d|x)$, where d is the end time of the study. There are two reasons for identifying $P(T=d|x)$. First, identifying this quantity is enough for constructing a consistent *inverse probability weighting (IPW)* estimator, similar to the one we saw in the causal inference. The other reason is that we can easily estimate the PDF $p(x,T=d)$ by using the observations without missing entries. If $P(T=d|x)$ is known, then we can identify $p(x)$ using $p(x) = \frac{p(x,T=d)}{P(T=d|x)}$.

The MAR and MCAR conditions are often expressed in a selection model framework. Formally, the MCAR is

$$P(T=t|X) = P(T=t).$$

Namely, the probability of any dropout time is totally independent of the study variable X. The MAR is

$$P(T=t|X) = P(T=t|X_{\leq t}).$$

In other words, the conditional probability of the dropout time only depend on the observed variables.

As we have mentioned, the selection model allows a simple way to construct a consistent estimator of a parameter of interest via the IPW procedure. Here is a simple example. Suppose that the parameter of interest is a linear statistical functional $\theta = \theta(F) = \int \omega(x)dF(x)$, then it can be further written as

$$\theta = \int \omega(x)p(x)dx = \int \omega(x)\frac{p(x,T=d)}{P(T=d|x)}dx = \int \omega(x)\frac{dF(dx,T=d)}{P(T=d|x)}.$$

With an estimator of the selection probability $\widehat{P}(T=d|x)$ (and we only need to estimate the probability of fully-observed case), a simple IPW estimator of θ is

$$\widehat{\theta}_0 = \int \omega(x)\frac{d\widehat{F}(dx,T=d)}{\widehat{P}(T=d|x)} = \frac{1}{n}\sum_{i=1}^n \frac{\omega(X_i)I(T_i=d)}{\widehat{P}(T=d|X_i)}. \tag{12.14}$$

You can show that $\widehat{\theta}_0$ is a consistent estimator (and it has asymptotical normality as well due to the Slutsky theorem). Moreover, the influence function (recall from the bootstrap lecture note) of $\widehat{\theta}_0$ can be easily derived so the variance of $\widehat{\theta}_0$ can be estimated via a plug-in estimate.

Although $\widehat{\theta}_0$ is elegant, it may not be the best estimator in the sense that after estimating the propensity score $P(T=t|x)$, we only rely on the completely observed data (the ones with $T_i=d$) to form the final estimator. Other observations are discarded entirely. Intuitively, this leads to an *inefficient* estimator.

To construct an efficient estimator, consider augmenting $\widehat{\theta}_0$ with an additional term

$$\widehat{\theta}_1 = \widehat{\theta}_0 + \frac{1}{n}\sum_{i=1}^n (I(T_i = \tau) - \widehat{P}(T_i = \tau | X_{i,\leq\tau}))g_\tau(X_{i,\leq\tau})I(T_i = \tau),$$

where $\tau < d$ is any time point and g_τ is a function of variable $x_{\leq\tau}$. The augmented term has an asymptotic mean 0 so $\widehat{\theta}_1$ is still a consistent estimator. The insight here is that the function g_τ is something we can choose– namely, we can choose it to minimize the variance of $\widehat{\theta}_1$ and this may leads to a reduction in the total variance compared to the estimator $\widehat{\theta}_0$. The same idea can be applied to every time point $\tau = 1, \cdots, d-1$, leading to an *augmented inverse probability weighting (AIPW)* estimator

$$\widehat{\theta}_{\mathsf{AIPW}} = \widehat{\theta}_0 + \frac{1}{n}\sum_{i=1}^n \sum_{\tau=1}^{d-1} (I(T_i = \tau) - \widehat{P}(T_i = \tau | X_{i,\leq\tau}))g_\tau(X_{i,\leq\tau})I(T_i = \tau).$$

With a proper choice of $g_\tau : \tau = 1, \cdots, d-1$, we can construct an estimator with the least variance. This leads to an efficient estimator. How to construct the functions $g_\tau : \tau = 1, \cdots, d-1$ is a central topic of *semi-parametric inference.*

Note that sometimes the AIPW (and IPW) estimators are constructed from solving an estimating equation. This occurs when the parameter of interest $\theta_0 = \theta(F)$ is defined through solving the equation

$$0 = \mathbb{E}(S(X;\theta_0)) = \int S(x;\theta_0)dF(x) = \int S(x;\theta)\frac{dF(dx, T = d)}{P(T = d|x)}.$$

In this case, the IPW estimator will be the solution to

$$0 = \int S(x;\widehat{\theta}_0)\frac{d\widehat{F}(dx, T = d)}{\widehat{P}(T = d|x)} = \frac{1}{n}\sum_{i=1}^n \frac{S(X_i;\widehat{\theta}_0)I(T_i = d)}{\widehat{P}(T = d|X_i)}$$

and we can augment it with a set of mean 0 terms to improve the efficiency.

If you are interested in the construction of AIPW, I would recommend the following textbook:

Tsiatis, A. (2007). *Semiparametric theory and missing data.* Springer Science & Business Media.

Note: although we introduce AIPW estimators in the MNAR framework, they are often used in the MAR scenario because the identification of propensity score/selection probability $P(T = t|X)$ is challenging in MNAR. The MAR is a simple case where we can identify the propensity score entirely so AIPW estimators can be constructed easily. Essentially, as long as you can identify the selection probability, you can construct an IPW estimator and attempt to augment it to obtain AIPW estimator to improve the efficiency. So the direction of research is often on how to identify the selection probability.

12.5.2 Pattern mixture models

Pattern-mixture models (PMMs) use another factorization of the full-data density:

$$p(x, t) = p(x_{>t}|x_{\leq t}, t)p(x_{\leq t}|t)P(T = t),$$

where the first term $p(x_{>t}|x_{\leq t}, t)$ is called the *extrapolation density* and the later two terms $p(x_{\leq t}|t)P(T = t)$ are called *observed-data density*. The extrapolation density is unobservable and unidentifiable–it describes the distribution of the missing entries. The observed-data density is identifiable since at each dropout time $T = t$, we do observe variables $x_1, \cdots, x_t$.

The PMMs provide a clean separation about what is identifiable and what is not identifiable. So the strategy for identifying $p(x,t)$ is to make the extrapolation density be identifiable.

In monotone missing problems, the extrapolation density has the following product form:

$$p(x_{>t} \mid x_{\leq t}, t) = \prod_{s=t+1}^{d} p(x_s \mid x_{<s}, T = t).$$

Thus, it suffices to identify each term in the product form to identify the extrapolation density. Several identifying restrictions have been proposed in the literature to identify the extrapolation density. For instance, the complete case missing value (CCMV) restriction equates that

$$p(x_s \mid x_{<s}, T = t) \overset{CC}{=} p(x_s \mid x_{<s}, T = d),$$

the available case missing value (ACMV) restriction assumes that

$$p(x_s \mid x_{<s}, T = t) \overset{AC}{=} p(x_s \mid x_{<s}, T \geq s),$$

and the nearest case missing value (NCMV) restriction requires that

$$p(x_s \mid x_{<s}, T = t) \overset{NC}{=} p(x_s \mid x_{<s}, T = s).$$

In general, one can specify any subset of patterns $\mathcal{A}_{ts} \subset \{s, s+1, \cdots, d\}$ and construct a corresponding identifying restriction

$$p(x_s \mid x_{<s}, T = t) \overset{\mathcal{A}_{ts}}{=} p(x_s \mid x_{<s}, T \in \mathcal{A}_{ts});$$

this is called the donor-baed identifying restriction in the following paper:

> Chen, Y. C., & Sadinle, M. (2019). Nonparametric Pattern-Mixture Models for Inference with Missing Data. arXiv preprint arXiv:1904.11085.

If you make any of these assumptions, the extrapolation density can be identified from the data so you can then estimate the full-data density $p(x,t)$.

Here is a nice review on PMMs for MNAR:

> Linero, A. R., & Daniels, M. J. (2018). Bayesian approaches for missing not at random outcome data: The role of identifying restrictions. *Statistical Science*, 33(2), 198-213.

12.5.3 Imputation and pattern mixture models

In the previous section, we introduce the idea of imputation when there is only one variable missing. But it can be applied to cases where there are multiple missing entries. Suppose that we have an imputation procedure such that if we observe $X_{\leq T} = (X_1, \cdots, X_T)$ and the dropout time T, the procedure generates random numbers $X_{>T} = (X_{T+1}, \cdots, X_d)$ from a distribution Q.

Then you can always view this imputation procedure as a PMM such that the PDF corresponds to the imputation distribution Q is the underlying model on the extrapolation density. So any imputation method can be viewed as implicitly handling the problem with a PMM.

From this point of view, you may notice that if we always impute the same number when observing $(X_{\leq T}, T)$, then this imputation procedure is problematic since the corresponding imputation distribution is not a good estimator of the underlying extrapolation distribution unless we are interesting in some very special parameter of interest. The commonly-used mean imputation or median imputation are thus bad ideas to apply in practice.

12.5.4 Nonparametric Saturation

In MNAR, we need to make identifying restrictions so that the full-data distribution $F(x,t)$ (or $p(x,t)$) is identifiable. However, there is one property that an identifying restriction should have: the implied joint distribution should be compatible/consistent with what we observe. This property is called nonparametric saturation/nonparametric identification/just identification.

The idea is simple: because we can identify $F(x,t)$, we can pretend the implied joint distribution is the true generating distribution and generates a new missing data from it. The generated missing data should be similar to the original data we have.

MAR and any pattern mixture models satisfies this property (when we attempt to estimate the joint distribution via a nonparametric estimator). However, some identifying restrictions, such as the MCAR, does not satisfy this. Whenever you proposed a new MNAR restriction, you should always think about if the implied full-data distribution satisfies this property or not.

12.5.5 Sensitivity analysis

Sensitivity analysis is a common procedure in handling the missing data problem. In short, sensitivity analysis is to perturb the missing data assumption a bit and see how the conclusion changes. This is often required in handling missing data because as we have shown previously, there is no way to check if a missing data assumption is correct (unless we have additional information) so our conclusion relies heavily on our assumption of missingness. By perturbing the assumption on missingness, we are able to examine if our conclusion is robust to the missing data assumption.

In MAR, one common approach for sensitivity analysis is to introduce the model (called the exponential tilting strategy)

$$\log \frac{P(T=t|X)}{P(T=t|X_{\leq t})} = \gamma^T X,$$

where $\gamma \in \mathbb{R}^d$ is a sensitivity parameter such that if $\gamma = 0$, we have $\frac{P(T=t|X)}{P(T=t|X_{\leq t})} = 1$, which is the MAR condition. We vary γ and examine how the estimator changes as a function of γ and use this as a way to how sensitivity the estimator depends on the MAR assumption.

12.5.6 Nonmonotone missing data problem

When the missingness is non-monotone (which occurs very often in a survey sample), the problem becomes a lot more complicated. Even we are willing to assume MAR, the full-data distribution $p(x)$ may not be unique. The following paper proposed a pattern mixture model to obtain a full-data distribution that satisfies MAR:

> Robins, J. M., & Gill, R. D. (1997). Non-response models for the analysis of non?monotone ignorable missing data. *Statistics in medicine*, 16(1), 39-56.

However, it only identifies one full-data distribution satisfying MAR, not all possible distributions.

The problem is even more challenging under MNAR case. In general, nonmonotone MNAR problem is still a very open problems. There are some attempts to deal with it but we have very limited options. Here are some recent work related to nonmonotone MNAR:

1. Sadinle, M., & Reiter, J. P. (2017). Itemwise conditionally independent nonresponse modelling for incomplete multivariate data. Biometrika, 104(1), 207-220.

2. Tchetgen, E. J. T., Wang, L., & Sun, B. (2018). Discrete choice models for nonmonotone nonignorable missing data: Identification and inference. *Statistica Sinica*, 28(4), 2069-2088
3. Malinsky, D., Shpitser, I., & Tchetgen, E. J. T. (2019). Semiparametric Inference for Non-monotone Missing-Not-at-Random Data: the No Self-Censoring Model. arXiv preprint arXiv:1909.01848.
4. Chen, Y. C., & Sadinle, M. (2019). Nonparametric Pattern-Mixture Models for Inference with Missing Data. arXiv preprint arXiv:1904.11085.

In particular, the first and the third model consider the following interesting assumptions:

$$X_j \perp R_j | X_{-j}, R_{-j},$$

where $R_j \in \{0, 1\}$ is the response indicator that $R_j = 1$ if variable X_j is observed. This assumption is known as ICIN (Itemwise conditionally independent nonresponse) and NSC (no self-censoring) assumption. It has a beautiful graphical representation induced by the conditional independence.