6.S897/HST.956 Machine Learning for Healthcare April 2, 2019

Lecture 14: Causal Inference, Part I

Instructors: David Sontag, Peter Szolovits *Scribes: Georgy Guryev*

1 Motivation to causal inference

Up until now we were talking about purely predictive questions. For the purely predictive questions one could think that the correlation is "good enough", i.e. we have some signs in our data that are predictive of some outcome of the interest and the causal directionality is somewhat irrelevant to this process. However, if one deals with a situation of the time-variant or non-stationary data, having a deeper understanding of the data might allow one to build a conditional robustness to the datashift.

Moreover, in health care we often want to answer the questions that are causal questions by nature rather then the predictive ones. For instance, in Lecture 4 & PS2 we applied machine learning methods to Marketscan dataset to build a risk stratification predictive model for early detection of type 2 diabetes onset within a 1-3 years time period. If we think for a moment how we want to deploy our classifier we will likely conclude that want to get patients to the clinic, to get them diagnosed. But the next set of questions is: what are we going to do with the patient based on that prediction; how should we intervene based on the diagnosis. At the end of the day the ultimate goal is to figure out how to prevent patients from developing complications of type 2 diabetes.

1.1 Examples of causal questions

1.1.1 Gastric bypass surgery

Recall, when we built the predictive model and introspected the most negative weights we might have noticed that the gastric bypass surgery had the largest negative weight. Does that imply that by giving the obese patient gastric bypass surgery we can prevent development of type-2 diabetes? That is an example of causal question raised by our predictive model. However, as we will see later, just looking at the weights along is not sufficient to infer the causal relationship.

1.1.2 Diagnosis of breast cancer

Before the spring break we had a number of lectures about diagnosis from imaging data. For this example we will consider a patient with breast cancer (Figure 1).

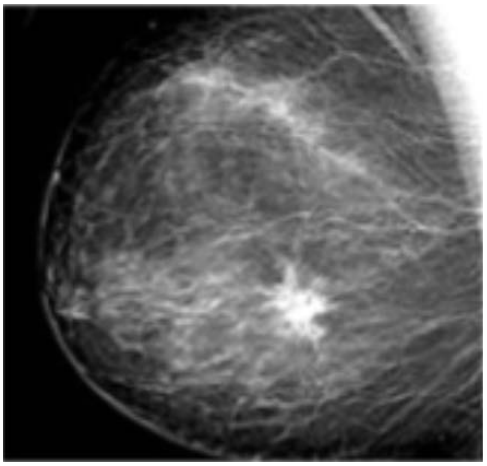


Figure 1: Patient with breast cancer

In this setting we want to answer the following question: what is the risk of this person dying in the next 5 years? We could train a deep learning model to predict the duration from diagnosis to death for a given patient. Based on this predictions you might take actions: if you predict that patient is not risky we might avoid treatment. However, that could be very dangerous: in the predictive models learned this way the outcome (death) is effected by what has happened in between.

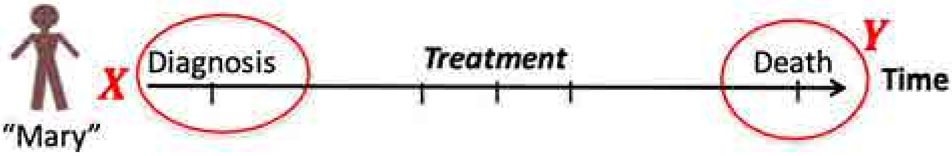


Figure 2: Patient with breast cancer

For instance, the patients from the data set might have being receiving treatment from the diagnosis till death that could have prolonged their life. I.e. if we ignore what happens in between and simply learn a mapping from image X to prediction Y we might come to a wrong conclusion for a new patient (e.g. we conclude that the patient does not require the treatment).

1.2 How do we guide treatment decisions?

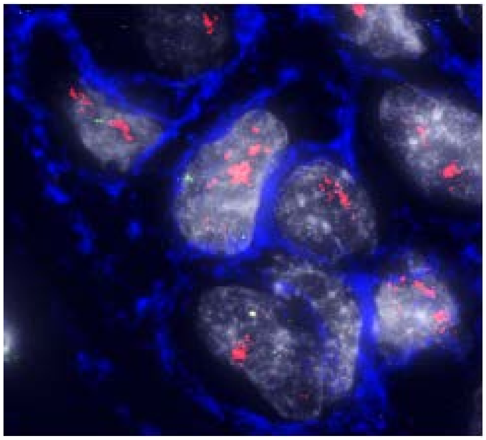


Figure 3: Pathology image

As data from pathology gets richer one might think to use computers to improve on human prediction accuracy of who is likely to benefit from a treatment. The challenge with using algorithms to do that is that humans respond differently to treatment and the data is biased to existing medical guidelines. The question that we might naturally ask is: what could go wrong if we trained to predict past treatment decisions? An example of past treatment history is given on Figure 4.

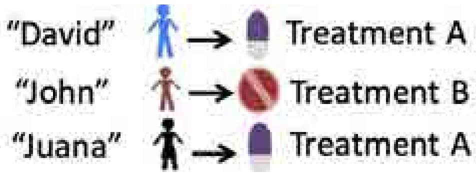


Figure 4: Example of treatment decisions

When a new patient comes in the best we can hope for is to do as good as existing clinical practice. However, if we want to go beyond (e.g. to recognize the heterogeneity response) then we have to modify the posed question.

1.3 Does smoking cause lung cancer?

The causal connection of smoking to lung cancer is the major question of societal importance. The reader might by familiar with the traditional way of addressing this question: the randomized control trial. Except, this is not exactly the type of setting where you can do the randomized control trial (it is not a feasible and unethical to ask a smoker to quit smoking, and/or ask a non-smoker to start smoking). In order to tackle such questions we have to start thinking on how we design ways to answer this questions using observational data. The challenge of this approach is that there is going to be a bias in data (i.e. who decided to smoke and who decides not to smoke). For instance, the most naive way to answer this question would be to look at the conditional likelihood of getting lung cancer among smokers, getting lung cancer among non-smokers. Unfortunately, the numbers in this approach might be quite misleading because there might be confounding factors that "cause" people to become smokers and cause people to develop cancer.

1.4 Conclusion to motivation section

Thus, when it comes to this field of machine learning in health care we need to think carefully about this type of questions, because the error in the way we formalize our problem could kill the people. In order to address this challenges properly we need to pose a questions as the causal ones. That is, rather than in traditional machine learning, where we have only inputs and outputs, we have to add a third quantity - interventions. Moreover, we have to start thinking about the causal relationship between these three. Thus one type of questions one might ask is to distinguish between different causal relationships expressed in terms of Bayesian directed graphical models.

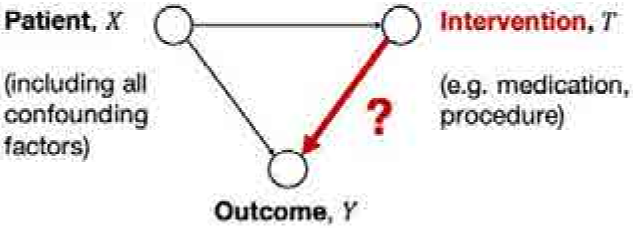

Patient, X
(including all confounding factors)
Intervention, T
(e.g. medication, procedure)
?
Outcome, Y

Figure 5: The simple causal graph

However, in this lecture we are going to talk about the simplest possible setting given by graph on Figure 5. The only unknowns here are the "strengths" of the edges. The treatment plan T here could be binary, continuous (dosage of a treatment) or a vector. For today's lecture we will assume the treatment plan to be binary. The nature of the directed graph tells us that the treatment plan T depends and the diagnoses for patient X, not the other way around, that becomes apparent if we compare graph with diagram from Figure 2. At the same time the outcome does depend on both diagnosis and the intervention plan T.

2 Potential outcomes framework

The questions we discussed in the previous sections have being studied for decades in political science, statistics, economics, bio-statistics, etc. However, the traditional approaches that statistics community used to work on no longer work on high-dimensional setting that is used in modern healthcare applications. In this lecture we will give an example of how we can bring the machine learning algorithms that are designed to work on high-dimensional data to answer this type of causal questions.

We shall start with introducing a language in order to formalizing these notions. We will work with the Rubin-Neyman causal model, where we will talk about the potential outcomes. The outcome $Y_0(x)$ then would correspond to the outcome when treatment zero (i.e. absence of treatment) was applied, while $Y_1(x)$ would correspond to the outcome when treatment 1 was administered. The preference of the type of treatment is formalized by the conditional average treatment effect for unit i that is given by formula:

$$CATE(x_i) = \mathbb{E}_{Y_1 \sim p(Y_1|x_i)}\left[Y_1|x_i\right] - \mathbb{E}_{Y_0 \sim p(Y_0|x_i)}\left[Y_0|x_i\right] \tag{1}$$

A simple example with a single feature (age) that explains the conditional average treatment effect is shown on Figure 6

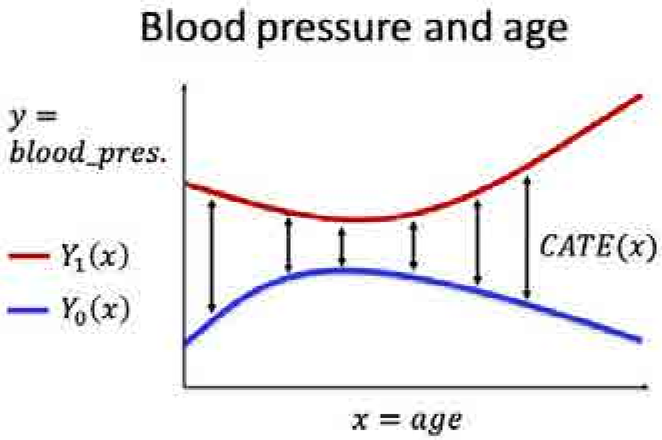


Figure 6: Example of CATE for blood pressure outcomes and feature age

Then average treatment effect over the distribution of people could be defined as

$$ATE = \mathbb{E}\left[Y_1 - Y_0\right] = \mathbb{E}_{x\sim p(x)}\left[CATE(x)\right] \tag{2}$$

A simple example with a single feature (age) that explains the average treatment effect is shown on Figure 7

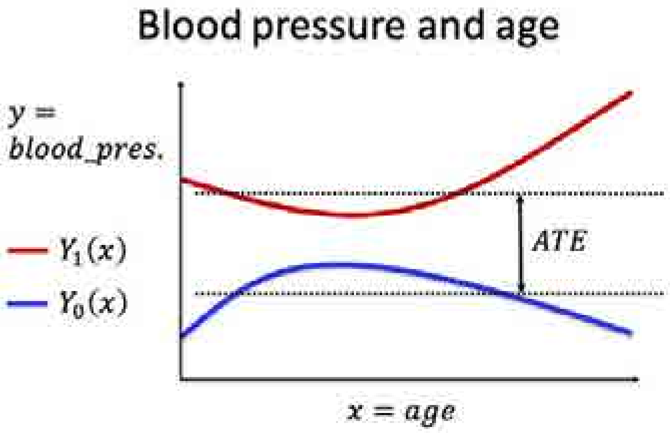


Figure 7: Example of ATE for blood pressure outcomes and feature age

The data you observe is the outcome for one of two interventions and reads:

$$y_i = t_1 Y_1(x_i) + (1 - t_1) Y_0(x_i) \tag{3}$$

where t_1 is a binary variable that indicates whether the i-th patient has received a treatment 1 or not. Thus, if the individual has received the treatment we would observe outcome $Y_1(x_i)$, if not - outcome $Y_0(x_i)$

On the other hand the unobserved counterfactual outcome, that is, the outcome for the opposite treatment, is given by:

$$y_i = (1 - t_1) Y_1(x_i) + t_1 Y_0(x_i) \tag{4}$$

From the definitions follows that the fundamental challenge of causal inference is that we observe only one of the two outcomes for the given patient.

The concepts of observed and unobserved counterfactual outcomes are summarized on Figure 8.

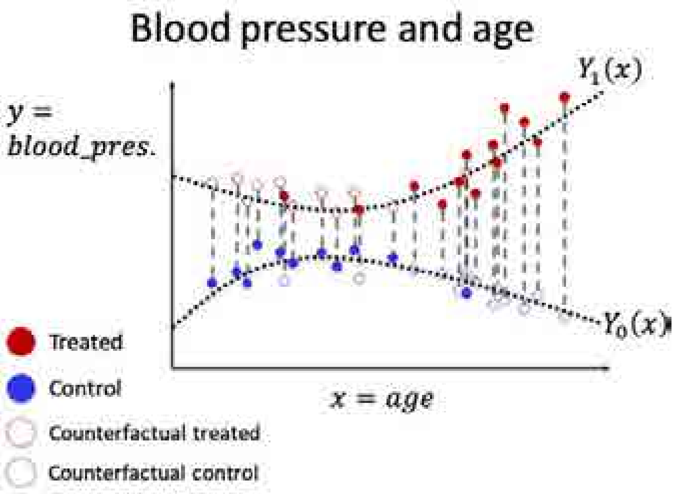
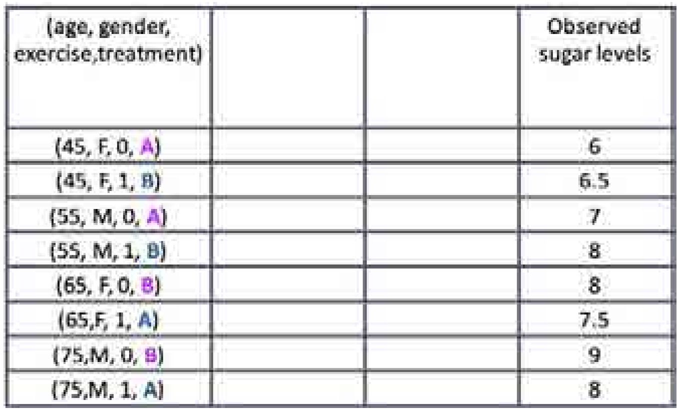


Figure 8: Counterfactual outcomes

Note that the dots are not on the curve because we want to allow some stochasticity, while the dotted lines represent the potential outcomes and the circles are the realisations.

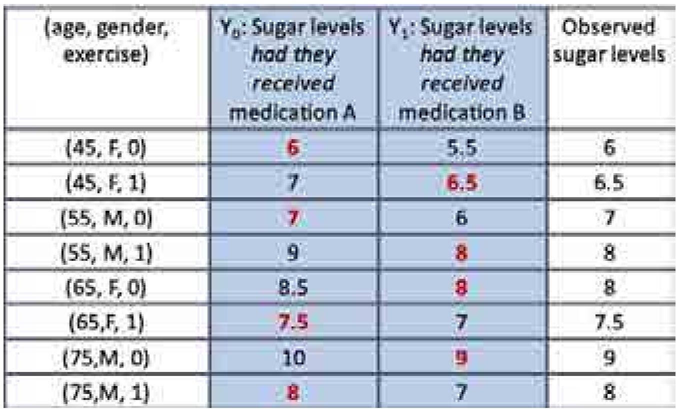
Let us look at the following example to get a better feeling of the introduced concepts. Consider the dataset shown in table on figure 9, where the first column consists of the patient-specific features such as age, gender, whether they exercise regularly and what treatment they got (A or B); the last column contains the observed sugar levels for these patients at the end of the year.



(age, gender, exercise,treatment)			Observed sugar levels
(45, F, 0, A)			6
(45, F, 1, B)			6.5
(55, M, 0, A)			7
(55, M, 1, B)			8
(65, F, 0, B)			8
(65,F, 1, A)			7.5
(75,M, 0, B)			9
(75,M, 1, A)			8

Figure 9: Example from Uri Shalit

From this table we want to infer the outcomes for all patients had they received treatment A or treatment B (see figure 10).



(age, gender, exercise)	Y_0: Sugar levels *had they received* medication A	Y_1: Sugar levels *had they received* medication B	Observed sugar levels
(45, F, 0)	6	5.5	6
(45, F, 1)	7	6.5	6.5
(55, M, 0)	7	6	7
(55, M, 1)	9	8	8
(65, F, 0)	8.5	8	8
(65,F, 1)	7.5	7	7.5
(75,M, 0)	10	9	9
(75,M, 1)	8	7	8

Figure 10: Caption

By comparing the two tables provided we note that we observe only the outcomes (observed sugar levels) for a single treatment at the time for the given patients (highlighted in red), while we do not know (do not observe) the counterfactual outcome (black numbers). In order to fill in an unobserved data we hypothesise: suppose someone told us that if the patient had the other treatment then we will see the values in black.

Now we are going to demonstrate the difference between the naive estimator of the averaging treatment effect and the true average treatment effect. We start by computing the average sugar level of the individuals with treatment B (we take red numbers from the third column only). That is equivalent to taking the rows for patients that have received treatment B from the first table. The resulting average sugar level for these patients is 7.875. Now we want to compute a similar quantity for patients that have received treatment A (the average of red numbers from the second column). The resulting value for sugar level in this case is 7.125. Now we compute the difference between this two and end up with the value of 0.75.

Now, we will estimate an average sugar levels in a different way. This time we will average all outcomes have the patients received treatment A/B (now we use both red and black numbers from a given column). Then computing the difference in expected sugar levels between treatments B and A we end up with -0.75. Now let us interpret the results in terms of a potential health insurance company's policies that are based on one of the predictions, i.e. the insurance company is trying to figure out if it is going to reimburse for treatment B? If they have used the naive estimation (the first approach we considered) then they would be reluctant to do that, since the treatment B is worse than the treatment A. However, if they used a proper estimation model (model 2) they would likely accept the opposite policy. This is a simple example that illustrates the difference between the conditioning and actually computing the counterfactual.

3 Assumptions in Neyman-Rubin causal model

Now the question one might ask is: How can I do anything without observing the actual counterfactual outcomes. Intuitively we understand, that generally there is no way we can recover the unobserved counterfactuals and thus it is an impossible problem to solve. However we might make a number of assumptions that will allow as to approximate the unobserved outcomes. Generally, the Neyman-Rubin causal model relies on the following assumptions:

- Stable Unit Treatment Value Assumption (SUTVA).
- Ignorability
- Common support

In this lecture we will focus on the last two assumptions, while information on SUTVA can be found here [Rub86].

3.1 Ignorability assumption

The ignorability assumption means that there are no unmeasured confounders. Mathematically speaking this means, that the potential outcomes Y_1, Y_2 are conditionally independent of treatment assignment conditioned on covariates x:

$$(Y_0, Y_1) \perp\!\!\!\perp T|X \tag{5}$$

The ignorability condition is represented as a directed graph on Figure 11. Note, that there are no edges from treatment T to the potential outcomes Y_0 and Y_1

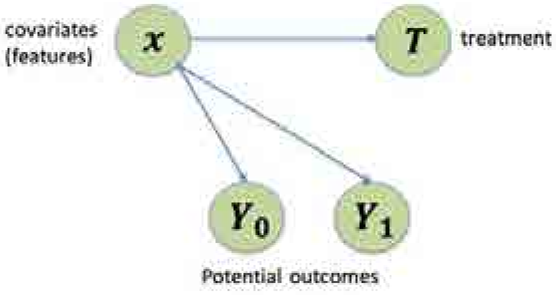


Figure 11: Illustration of the ignorability assumption

The intuition behind this graph is that we already know all the potential outcomes for given covariates so it does not make any sense to introduce edges from T to Y_0, Y_1. The example of violation of igorability condition implies that there exists some latent variable h such, that it has an impact on both the treatment decision and on the potential outcomes (see graph on Figure 12).

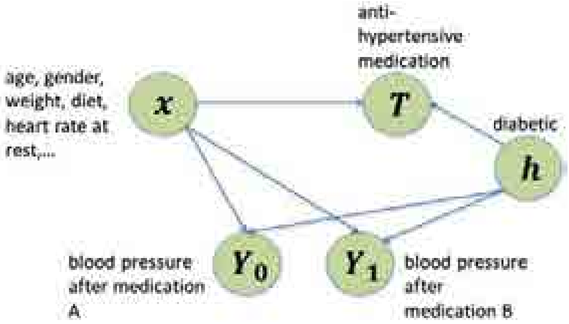


Figure 12: Illustration of the violation of the ignorability assumption

Thus, the assumption of no unmeasured confounders holds when we observe all possible factors that influence both treatment decisions and potential outcomes. That implies, that in order to be confident in the validity of the ignorability condition (and of the specified graph model) one has to consult with a specialists to make sure that all factors that influence doctor's decisions and potential outcomes are observed.

To learn how to assess the robustness to violations of these assumptions reader should refer to paper [] on sensitivity analysis.

3.2 Common support

The common support assumption says that there should always be some stochasticity in treatment decisions. That means that any group of patients/features should have nonzero probabilities for all considered treatments. The above statement could be formulated in the following way:

$$P(T = t|X = x) > 0, \quad \forall t, x \tag{6}$$

The expression above goes by the name of propensity score: the probability of receiving the treatment for each individual. Thus we assume, that this probability is bounded between 0 and 1, while any violation of this condition is going to completely validate the conclusions drawn from the data.

4 Approximate Average treatment effect with data

In this section we will assume that the ignorability and common support conditions hold for our model. Now the question we are interested in is: what methods should we use to estimate average treatment effect (ATE) and conditional average treatment effect (CATE).

4.1 Average treatment effect

Recall, that the ATE is defined as the expectation of the difference between two distinct treatment outcomes. The key tool that we are going to use is called the adjustment formula (also known as *g-formula* in statistical community). Due to linearity of the expectation operator we may rewrite the expectation of a difference in outcomes as the difference of expectations of treatment effects. The expectation of the potential treatment effect reads:

$$\mathbb{E}[Y_1] = \mathbb{E}_{x\sim p(x)}[\mathbb{E}_{Y_1\sim p(Y_1|x)}[Y_1|x]] \tag{7}$$

For the next step recall, that we assumed the conditional independence of the potential outcomes from the treatment decisions conditioned on covariates. Applying this condition to the above formula we get

$$\mathbb{E}[Y_1] = \mathbb{E}_{x\sim p(x)}[\mathbb{E}_{Y_1\sim p(Y_1|x)}[Y_1|x, T = 1]] = \mathbb{E}_{x\sim p(x)}[\mathbb{E}[Y_1|x, T = 1]] \tag{8}$$

The above steps are completely analogous for the expectation of the potential outcome Y_0. Then, the formula for the average treatment effect can be written as

$$ATE = \mathbb{E}[Y_1 - Y_0] = \mathbb{E}_{x\sim p(x)}[\mathbb{E}[Y_1|x, T = 1] - \mathbb{E}[Y_0|x, T = 0]] \tag{9}$$

Note that both $\mathbb{E}[Y_1|x, T = 1]$ and $\mathbb{E}[Y_0|x, T = 0]$ are the quantities that we can estimate from the data. However there are still some quantities that we can not directly estimate from data: these are:

- $\mathbb{E}[Y_0|x, T = 1]$ - the potential outcome for treatment 0 if patient x received treatment 1.
- $\mathbb{E}[Y_1|x, T = 0]$ - the potential outcome for treatment 1 if patient x received treatment 0.
- $\mathbb{E}[Y_0|x]$ - the potential outcome for treatment 0 given patient x
- $\mathbb{E}[Y_1|x]$ - the potential outcome for treatment 1 given patient x

Now let's return back to the formula (9). Empirically we have samples from $p(x|T = 1)$ for $\mathbb{E}[Y_1|x, T = 1]$ and $p(x|T = 0)$ for $\mathbb{E}[Y_0|x, T = 0]$ respectively. However, in order to compute an expectation over x we still have to extrapolate $\mathbb{E}[Y_1|x, T = 1]$ and $\mathbb{E}[Y_0|x, T = 0]$ for some individuals (extrapolate to p(x)).

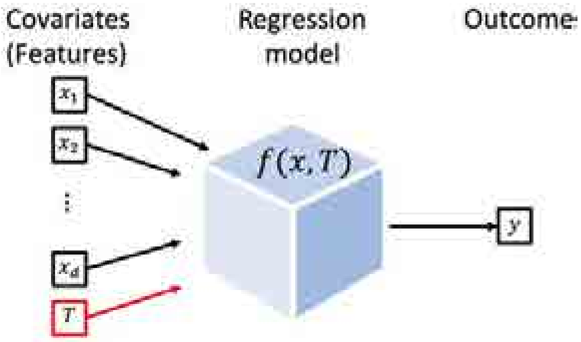


Figure 13: Covariate adjustment model

5 Types of causal inference methods

There exists a large number of types of causal inference training, that include covariate adjustment, propensity score, doubly robust estimators, matching, etc. In this lecture we will cover the first method.

5.1 Covariate adjustment (Response surface modelling)

Covariate adjustment is the natural way to perform the aforementioned extrapolation. In this method the key objective is to find function $f(x,T)$ which takes as an input x and T and its goal is to predict the potential outcome (see figure). We can think of $f(x,T)$ as of an approximation to the conditional probability $p(Y_t|x,T=t)$.

Thus, the covariate adjustment method is a regression model that explicitly models the relation between treatments, confounders and outcomes. The approximations of ATE and CATE for this method are given by

$$\widehat{ATE} = \frac{1}{n}\sum_{i=1}^{n}(f(x_i,1) - f(x_i,0)) \tag{10}$$

and

$$\widehat{CATE}(x_i) = f(x_i,1) - f(x_i,0) \tag{11}$$

The natural question to ask is: how do we know that the models have learned a meaningful relationship between covariates, the itervention T and the outcome (is the reduction to ML valid)? For instance it could happen that the treatment decision is among a huge number of factors that effect the outcome Y. Moreover, it could be that there are much more important factors than T, and we apply a regularizer to our data (here assume that data is high-dimensional). In this setup it is possible that the regression algorithm would simply ignore actual dependence on T (It might learn to ignore T).

This is the point where the causal inference and machine learning start to differ: we do not care about the accuracy of the predicted outcomes, we want to learn the causation instead(T is the parameter of interest). That is the gap of our understanding today. The challenge is how to change the machine learning paradigm to recognize that using machine learning in the causal inference you are actually interested in something different. This fundamental problem does not show up that often when one deals with a low-dimensional data. However when one moves to a higher-dimensional space the difference in goals becomes extremely important.

5.1.1 Examples of when covariate adjustment fails

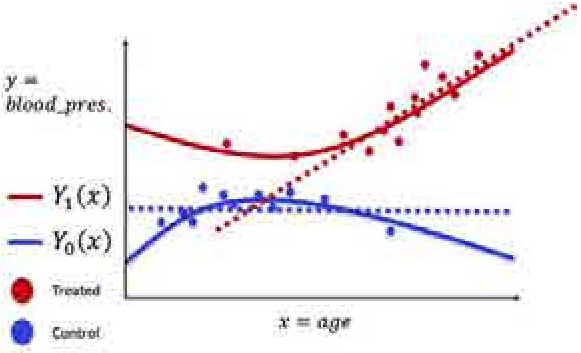


Figure 14: When covariate adjustment fails

Here we discuss what causes the covariate adjustment method to fail. The simple of such behaviour is when the data violates the assumption of common support/overlap. In this case the regression method does not have enough data to extrapolate fro the outcomes correctly.

Another example is somewhat more subtle: we made an assumption that we sample enough data to extrapolate correctly. However, that also implies that we have selected a family of functions that is powerful enough to extrapolate correctly. For instance,if the true outcome functions are quadratic functions and we fit them with a linear functions then we will get wrong estimates regardless of the number of samples we have.

References

[BAB$^+$18] Gabriel A Brat, Denis Agniel, Andrew Beam, Brian Yorkgitis, Mark Bicket, Mark Homer, Kathe P Fox, Daniel B Knecht, Cheryl N McMahill-Walraven, Nathan Palmer, and Isaac Kohane. Post-surgical prescriptions for opioid naive patients and association with overdose and misuse: retrospective cohort study. *BMJ*, 360, 2018.

[HM19] Robins JM Hernn MA. *Causal Inference*. Boca Raton: Chapman Hall/CRC, 10 February 2019.

[Rub86] Donald B. Rubin. Statistics and Causal Inference: Comment: Which Ifs Have Causal Answers. *Journal of the American Statistical Association*, 81(396):961–962, 1986.