*Figure 11.1* depicts the potential outcome model in which an individual (or unit) i has covariates and is assigned a binary treatment indicator , which takes the value one if the individual i receives the treatment and zero if i receives the control (indicating not treated or receiving a placebo). Let be the “potential outcome” if we give the treatment to i. Let be the “potential outcome” if we give the control to i.

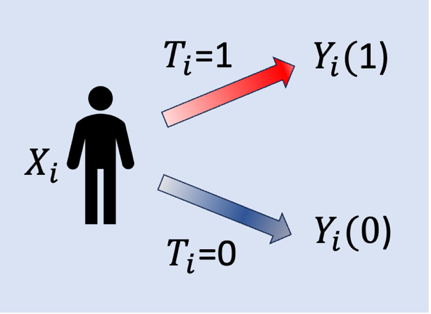


Figure 11.1—Potential outcome model

The **conditional average treatment effect (CATE**) is the treatment effect estimand focusing on subpopulations in which all units have the same value assigned to a covariant vector. The CATE is a function defined as the expected difference between the two POs conditioned on the covariates ,

(11.1)

The CATE provides insight into the effect of treatment cases when the effect differs between subgroups, also known as the **heterogeneous treatment effec (HTE)** (Athey and Imbens 2016), (Kunzel, et al. 2019), (Nie and Wager 2021). The CATE estimation problem is to

The CATE estimation problem is to compute an estimate of the true of equation (11.1) using as input the observational data consisting of 3-tuples for the treatment , the covariates , the observed outcome , for every unit indexed from one to , where is the total sum of all units in both the treatment and control group,

(11.2)

The core issue in causal inference is that real data can possess the potential outcome for either the treatment or the control condition but not simultaneously for any given unit. The missing data problem is that for any unit i receiving the treatment (), we only observe the potential outcome and lack information on the potential outcome had the unit received the control . For any unit i administered the control (), we only know the potential outcome and lack knowledge of the potential outcome had the unit received the treatment, .

Conditional average treatment effect

The conditional average treatment effect (CATE) is the treatment effect estimand focusing on subpopulations in which all units have the same value assigned to a covariant vector. The CATE is a function defined as the expected difference between the two POs conditioned on the covariates ,

(11.2)

The CATE provides insight into the effect of treatment cases when the effect differs between subgroups, also known as the heterogeneous treatment effect (HTE) (Athey and Imbens 2016), (Kunzel, et al. 2019), (Nie and Wager 2021).

The CATE estimation problem is to compute an estimate of the conditional expectation of the difference between the potential outcomes under treatment and control conditioning on value instantiations of the covariates as expressed by equation (11.2). The CATE estimation takes as input the observational data consisting of 3-tuples for the treatment , the covariates , the observed outcome , for every unit indexed from one to , where is the total sum of all units in both the treatment and control group,

(11.3)

Estimating the CATE is identifiable under the three POF assumptions: the Exchangeability assumption, the Positivity assumption, and the Stable Unit Treatment Value Assumption (SUTVA). See the “Identifying Causal Effects" section in Chapter 10.

Estimating the CATE quantifies the causal effect of a treatment on specific subgroups of a population, providing more specific and actionable insights than what is obtained from estimating the overall average treatment effect (ATE). Application areas of CATE estimation include a) personalized medicine to customize therapy to the individual patient; b) targeted marketing and recommendation systems to identify customers most likely to respond to marketing offers and to product suggestions; c) social science research by identifying subgroups who will benefit most from targeted social initiativesl d) economic policy setting by assessing workforce development initiatives effect on various demographics employment and labor markets.

Machine learning algorithms are employed to perform the CATE estimation. Due to the missing data problem (the actual individual treatment effects are never directly observable), evaluating ML algorithms for estimating the CATE poses a challenge, even when the identifying assumptions of the Neyman-Rubin model are satisfied. To provide proof of concept for ML algorithms in CATE estimators, we may rely on synthetic or semi-synthetic datasets (with real covariates) that possess a generative data model, in which both potential outcomes are simulated and hence known, to showcase their properties.

Identifying Causal Effects

In this section, we present the conditions the potential outcomes model requires to ensure that the treatment effect is identifiable. Identifiability implies that the causal effect estimand can be estimated from the observational dataset. The observational dataset contains n-value tuples for the treatment T, the confounders X, and the outcome Y, respectively. The dataset represents an n-independent identically distributed (iid) sample from an underlying distribution .

This section is organized as follows. First, we describe the potential outcome framework (POF) assumptions to identify the ATE estimand. Next, we demonstrate that we can derive the ATE based on the probability distribution of the observational data under the conditions of the POF assumptions.

POF Assumptions to Identify the ATE

The POF must make three assumptions to ensure the causal effect is identifiable and consistently estimate the ATE from observational data. These assumptions are the Exchangeability assumption, the Positivity assumption, and the Stable Unit Treatment Value Assumption (SUTVA).

Exchangeability

The exchangeability assumption is alternatively named the unconfoundedness assumption, the conditional-independence assumption, or the ignorability assumption. It states that conditional on the covariates the treatment is independent of the potential outcomes (That is,

(10.13)

The exchangeability assumption requires that the potential outcomes for any two units within strata defined by covariates , the probability of their potential outcomes should be independent of the treatment assignment

(10.14)

The exchangeability assumption permits the counterfactual (unobservable) outcomes in an exposed group (T=1) to be proxied by the observed results in an unexposed group (T=0) (Igelström, et al. 2022).

Positivity

The positivity assumption, also called the overlap assumption, is as follows. Conditional on the covariates , every unit has some non-zero probability of receiving treatment () or control (). In other words, the treatment assignment should be stochastic, which means the propensity score must be a positive number and must be less than 1,

(10.15)

Unlike the case of the exchangeability assumption, which is not possible to verify from the data alone, the positivity assumption can sometimes be verified from the data (Hernán and Robins, Causal Inference: What If 2020, Lee and Lee 2022).

Stable Unit-Treatment Value Assumption (SUTVA)

The SUTVA is composed of two assumptions: Consistency and No interference. The consistency assumption says that If treatment is applied, the observed outcome is the potential outcome for treatment , that is:

(10.16)

The consistency assumption states that for each unit under every treatment scenario, , , the potential outcomes are well-defined and have a single value. This assumption will be broken if multiple treatment versions are in place, as might be the case with surgery, for example, if different surgeons perform the surgery and if these various treatment versions result in different possible outcomes (VanderWeele and Hernán 2013). The consistency assumption is also called the no-multiple-versions-of-treatment assumption. Cole et al. (Cole and Frangakis 2009) pointed out that consistency is a challenge in observational studies with treatments that are hard to imagine manipulating as in cases where the treatment is a biological feature like body weight; there are numerous competing methods to (potentially) give a person a body mass index of 25 kg/m2, and each method may have a distinct causal effect on the result. Pearl (J. Pearl 2010) has argued that the consistency assumption should be framed as a "theorem in the logic of counterfactuals."

The no interference assumption is that no interference or spillover effect creates interdependencies between the unit's potential outcomes. When there is no interference, then the potential outcomes for a unit are functions of the treatment assignments , to the unit regardless of the treatment assignments to all the other units That is,

(10.17)

Gerber and Green (Gerber and Green 2012) describe many situations where the no interference assumption may be violated. For instance, in the case of vaccination treatment, it is conceivable that the potential outcomes of contracting a disease in response to treatment of a unit are likely to be influenced by the treatment of other units.

Identification under the POF Assumptions

The following demonstrates that the average causal effect is identifiable under the POF assumptions. That means we can estimate the ATE using the observational data.