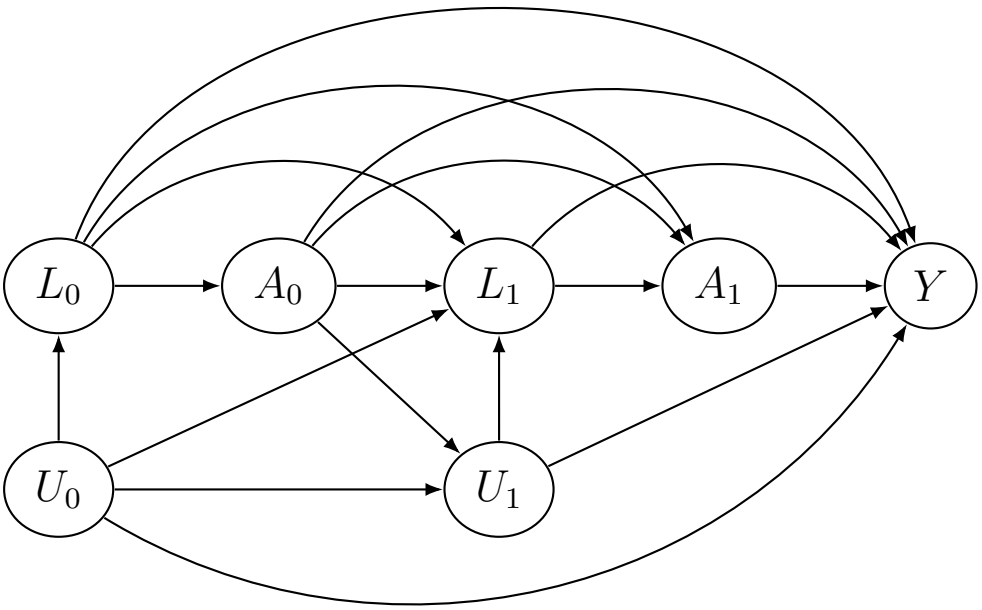
CAUSAL INFERENCE WITH A VIEW TOWARDS LONGITUDINAL DATA ANALYSIS



5.213a
Aalborg University
Mathematics

AALBORG UNIVERSITY
STUDENT REPORT

Department of Mathematical Sciences
Mathematics
Skjernvej 4A
9220 Aalborg Øst
http://math.aau.dk

Title:

Causal Inference with a view towards Longitudinal Data Analysis

Project:

Master's thesis

Project Period:

February 2022 – June 2022

Project Group:

5.213a

Participants:

Jacob Nissen
Tóra Oluffa Stenberg Olsen

Supervisor:

Torben Tvedebrink

Pages: 66

Finished: 2nd of June 2022

Abstract:

This master's thesis is a study of causal inference with a view towards longitudinal data analysis. The aim is to study the theory of causal inference and various methods to estimate causal effects. We present the TMLE, CV-TMLE and L-TMLE methods and properties of these methods. Moreover, we apply these methods to a subset of the Framingham Heart Study, which is a longitudinal study, to determine if smoking has a causal effect on stroke in both a non longitudinal study and a longitudinal study.

In particular, we first define an (average) causal effect and present the identifiability conditions which are required in order to identify (average) causal effects for observational studies. Then we present the inverse probability weighting and standardisation methods. Furthermore, we describe how causal directed acyclic graphs can be used to illustrate relations between the covariates, treatment and outcome.

Then we present structural causal models in order to specify the target parameter. Furthermore, we present the TMLE and CV-TMLE methods and show asymptotic linearity of the CV-TML estimator. Afterwards, we present a longitudinal extension of (average) causal effects, the identifiability conditions and the TMLE method.

Finally, we apply the TMLE and CV-TMLE methods as well as the L-TMLE method in order to examine if smoking has a causal effect on stroke within 24 years in a non longitudinal and longitudinal study, respectively.

Preface

This master's thesis has been written by group 5.213a throughout the course of the 4th semester of the Master's degree programme at the Department of Mathematical Sciences at Aalborg University. To read this master's thesis basic knowledge of graph theory, survival analysis and machine learning is required.

We would like to express our gratitude to Torben Tvedebrink for supervision during the course of the project period.

Signatures

Jacob Nissen

Tóra Oluffa Stenberg Olsen

Contents

1 Introduction

In this master's thesis, we study causal inference with a view towards longitudinal data analysis. Causal inference is the process of determining whether a treatment has an effect on an outcome of interest. If this is the case, we refer to such an effect as a causal effect, that is, the treatment has a causal effect on the outcome. The focus of this master's thesis is to study the theory of causal inference and various methods used to estimate causal effects in practice. In particular, for the practical aspect of this master's thesis, we examine a subset of a data set called the Framingham Heart Study. The Framingham Heart Study is a longitudinal study which examines cardiovascular diseases among a population in Framingham (Massachusetts, USA). When examining this study, we analyse whether smoking has a causal effect on stroke within 24 years in both a non longitudinal and a longitudinal setting. In order to estimate a possible causal effect, we first present the theory of causal inference for both non longitudinal and longitudinal studies and outline various methods which can be used to estimate causal effects in practice.

Hence, we first outline fundamental theory of causal inference in Chapter 2 for a non longitudinal study. In this chapter, we define a causal effect for an individual and for a population more precisely and clarify the difference between causation and association. Notice that we restrict this thesis to binary treatments and outcomes. Also we present study designs for which we can identify possible causal effects. Moreover, we present the identifiability conditions which are conditions required in order to identify causal effects in observational studies. Furthermore, we use directed acyclic graphs to illustrate possible biases which we need to adjust for in order to identify possible causal effects. In addition, we present methods which can be used to adjust for such biases.

In Chapter 3, we present a method for estimating causal effects in a non longitudinal study. First, we present structural models which we use to define the target parameter, that is, the measure for the causal effect of interest. Having defined the parameter of interest, we present the method consider in this chapter namely targeted maximum likelihood estimation (TMLE). This method consists of two steps where we in the first step obtain an initial estimator of the parameter of interest by using the super learner method. Then in step two, we target the initial estimator towards the parameter of interest in order to obtain an optimal bias-variance trade off for the parameter of interest. However, it has been observed in practice that when using too data adaptive methods in the super learner method, the TMLE method suffers which motivates the use of a more robust method.

Thus, we in Chapter 4, consider an extension of the TMLE method, that is, cross validated targeted maximum likelihood estimation (CV-TMLE). This method adds an additional layer of cross validation to the TMLE method in order to obtain a more robust method. Hence, the CV-TMLE method can also be used to estimate causal effects in practice. Having presented the

method, we then show that the estimator obtained from the CV-TMLE method is asymptotically linear. This yields a method which is asymptotically unbiased and also provides a method of obtaining confidence intervals for the estimator. In order to show the asymptotic linearity of this estimator, we first present influence functions.

In Chapter 5, we then extend the theory of causal inference to longitudinal studies. Specifically, for longitudinal studies, we consider time-varying treatments and hence, we extend the definition of a causal effect in order to incorporate the changes in the treatment over time. Moreover, we also present study designs where causal effects of time-varying treatments on an outcome can be identified. Furthermore, we extend the identifiability conditions to longitudinal settings in order to identify causal effect in observational longitudinal studies. Having extended the theory of causal inference to longitudinal studies, we then extend the TMLE method to longitudinal studies which is called the longitudinal targeted maximum likelihood estimation (L-TMLE) method.

In Chapter 6, we estimate causal effects in practice by considering the beforehand mentioned Framingham Heart Study. For the non longitudinal setting, we apply the TMLE method as well as the CV-TMLE method in order to analyse the following question: *"Does smoking at the time of the first examination have a causal effect on stroke within a period of* 24 *years"*. Hence, in this case we consider smoking at the time of the first examination as the treatment and stroke within 24 years as the outcome. For the longitudinal setting, that is, for a time-varying treatment, we analyse the following question: *"Does smoking at the time of each of the examinations have a causal effect on stroke within a period of* 24 *years compared to not smoking at any of the examinations"*. Thus, the treatment in this setting is smoking at the time of each of the examinations of the Framingham Heart Study. Moreover, based on the asymptotic linearity, we also provide 95% confidence intervals for the estimators obtained from the TMLE and the CV-TMLE methods for the non longitudinal setting and for the L-TMLE method for the longitudinal setting.

2 Causal Effects

In this chapter, we present the fundamental theory of causal inference. Specifically, we define a causal effect and an average causal effect as well as study designs where such average causal effects can be identified. Moreover, we present the conditions required to identify these effects in observational studies. Furthermore, we use graph theory to illustrate the relations between the variables and later present methods to identify an average causal effect.

Let Y and A be random variables denoting an outcome and a treatment, respectively. In this master's thesis, we restrict Y and A to binary variables which attain values zero and one since this aligns with the data set examined in Chapter 6. Furthermore, in this chapter, we assume that there is no random variation. Moreover, in this master's thesis, we assume that all variables are perfectly measured, that is, there is no *measurement bias*. Measurement bias is out of scope for this master's thesis and we refer to [Miguel A. Hernán, 2020, Chapter 9].

2.1 Causation and Association

This section is based on [Miguel A. Hernán, 2020, pp. 3–5, 11–12]. In this section, the aim is to determine when a treatment has a causal effect on an outcome. Moreover, we in this section clarify the distinction between causation and association. In order to do so, we first define the *potential outcomes* for an individual.

Definition 2.1.1. Potential Outcomes
Consider the outcome Y_i for individual i as a function of treatment A. Then the potential outcomes for individual i are defined as $Y_i(a)$ for all a which are realisations of A. ◀

Notice that Y_i is a deterministic function. Hence, the potential outcomes consist of the outcomes under every value of treatment for an individual. Having defined the potential outcomes, we now define a causal effect for an individual.

Definition 2.1.2. Causal Effect for an Individual
The treatment A has a non-zero causal effect on the outcome for individual i if $Y_i(1) \neq Y_i(0)$. ◀

Thus, if for an individual, the potential outcomes differ for different values of treatment then the treatment has a causal effect on the outcome for this individual. However, the fundamental issue of causal inference is that, for each individual, we only observe one of the potential outcomes since every individual only receives one treatment. Hence, we are unable to determine whether

or not the potential outcomes differ and thus identify causal effects for an individual. Therefore, we instead focus on the *average causal effect* in a population.

Definition 2.1.3. Average Causal Effect in a Population
The treatment A has a non-zero average causal effect on the outcome Y if $\mathbb{E}[Y(1)] \neq \mathbb{E}[Y(0)]$. ◀

When considering causal effects, it is important to distinguish between causation and association. We say that A and Y are associated if $\mathbb{E}[Y \mid A = 1] \neq \mathbb{E}[Y \mid A = 0]$. Thus, association is determined based on the actual treatment of the individuals in the population, that is, we have two disjoint populations based on the values of A while causation is determined based on the same population under the two different values of A. Hence, in general, association does not imply causation, that is,

$$\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] \neq \mathbb{E}[Y \mid A = 1] - \mathbb{E}[Y \mid A = 0]. \tag{2.1}$$

In the following, we present an example where association is not causation.

Example 2.1.4. Association is not Causation
Consider the case where all potential outcomes of the individuals in a population are known. These potential outcomes are shown in Table 2.1 which also includes two possible assignments of the treatment denoted as A and A'.

Table 2.1: Potential outcomes for each individual in the population and two possible assignments of treatment.

Individual	Y(0)	Y(1)	A	A'
1	0	1	1	0
2	1	0	1	0
3	0	0	1	1
4	0	1	1	1
5	1	0	1	0
6	1	1	0	1
7	0	1	0	0
8	1	1	0	1
9	1	0	0	0
10	0	0	0	1

Hence, since the potential outcomes are known, we can calculate the average causal effect as

$$\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = 0.5 - 0.5 = 0. \tag{2.2}$$

Thus, the average causal effect is zero which implies that there is no average causal effect.

Now assume that given a treatment, the actual outcome coincides with the corresponding potential outcome. Hence, if we assume that the treatment assignment of the individuals is A of Table 2.1, then we can compute the association as

$$\mathbb{E}[Y \mid A = 1] - \mathbb{E}[Y \mid A = 0] = \frac{2}{5} - \frac{3}{5} \neq 0. \tag{2.3}$$

Therefore, there is an association between treatment and outcome while there is no causation. Hence, in this case, association is not causation. ◀

In the above example, we illustrated a case for which association is not causation. However, if we had assigned treatment A' instead of treatment A in Table 2.1, causation would have been association since

$$\mathbb{E}[Y \mid A' = 1] - \mathbb{E}[Y \mid A' = 0] = \frac{3}{5} - \frac{3}{5} = 0. \tag{2.4}$$

Hence, the assignment of the treatment affects the relation between association and causation. Therefore, we want to determine for which cases association is causation since the association can be determined based on the observed treatment. Thus, we in the following section present two study designs where association is causation.

2.2 (Conditionally) Randomised Experiments

This section is based on [Miguel A. Hernán, 2020, pp. 4, 13–18, 31–35]. In a *randomised experiment*, that is, an experiment where the treatment is assigned randomly, association is causation by design which we clarify in the following. If treatment is randomised then, for any realisation a of A, $\mathbb{P}(Y = 1 \mid A = a)$ would be independent of which group receives which value of the treatment. Thus, the groups are *exchangeable* which implies that

$$\mathbb{P}(Y(1) = 1 \mid A = 1) = \mathbb{P}(Y(1) = 1 \mid A = 0) = \mathbb{P}(Y(1) = 1). \tag{2.5}$$

Analogously, we can derive $\mathbb{P}(Y(0) = 1)$. Therefore, exchangeability implies that $Y(a) \perp\!\!\!\perp A$ for all a which are realisations of A. Hence, in randomised experiments, association is causation since for binary A and Y, we get

$$\begin{aligned} \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] &= \mathbb{P}(Y(1) = 1) - \mathbb{P}(Y(0) = 1) \\ &= \mathbb{P}(Y(1) = 1 \mid A = 1) - \mathbb{P}(Y(0) \mid A = 0) \\ &= \mathbb{E}[Y(1) = 1 \mid A = 1] - \mathbb{E}[Y(0) \mid A = 0] \\ &= \mathbb{E}[Y \mid A = 1] - \mathbb{E}[Y \mid A = 0], \end{aligned} \tag{2.6}$$

where exchangeability is used in the second equality and the last equality holds under the assumption of *consistency*. Consistency is the case where for each individual

$$Y_i(a) = Y_i(A) = Y_i \quad \text{if } A = a. \tag{2.7}$$

That is, the consistency assumption ensures that the actual outcome Y coincides with the potential outcome $Y(A)$ under the corresponding treatment. Hence, under consistency $Y = Y(A)$. Notice that, consistency is not always fulfilled. Consistency depends on a precise definition of the treatment levels in order to obtain well-defined potential outcomes. This is the case since, if the treatment levels are not specified precisely, then we can have multiple versions of the treatment. We can illustrate this by considering a study where we examine the causal effect of the treatment heart transplant on the outcome death within five years. Then if the treatment heart transplant is not specified in further details, then the heart transplants given in the study may differ based on various pre-operative procedures, surgical technique etc. That is, we have multiple versions of the treatment heart transplant. If the multiple versions of the treatment have different causal effects on the outcome, then the potential outcomes are not well-defined since for each individual $Y_i(a)$ can attain multiple values based on the specific version of the treatment. In the particular example, consider the case where two different surgical techniques were used and one of them had a causal effect on the outcome and the other did not. Then for an individual in this study, its potential outcome under the treatment value heart transplant is not well-defined since it depends on the particular surgical technique. Hence, a precise specification of the *causal question*, that is, the formulated question of whether treatment has an average causal effect on the outcome or not, is needed in order to ensure consistency.

Moreover, notice that, do to the fundamental issue of causal inference then, in practice, both potential outcomes of an individual are not available. Therefore, we are generally unable to determine if exchangeability holds for a specific study. Furthermore, in some studies, it might be impossible or unethical to randomise treatment and thus we are interested in how to identify the causal effect in such cases. If the treatment is not randomised, then there exist factors which affect the assignment of treatment. An example could be when considering the causal effect of heart transplant on death within five years, there can be factors such as the individuals condition which affect whether the individual receives a heart transplant or not. In such a case, if we were to assign treatment with a relatively large probability to those in critical condition and with a lower probability to those not in critical condition, then the experiment would be a *conditionally randomised experiment*. This is the case since we used multiple randomisation probabilities conditional on the individuals condition.

Conditionally randomised experiments, generally, do not produce exchangeability by design. However, conditionally randomised experiments can be viewed as a combination of multiple randomised experiments. For example, in the previous example of heart transplant where the probability of treatment was dependent on the condition of the individual, then we could divide the population into subsets based on whether or not the individual was in critical condition. Thus, in each of these subsets, we assign the treatment based on one randomisation probability and hence the experiment within a subset can be seen as a randomised experiment. Therefore, if we denote the individuals condition by W then, for all w which are realisations of W and for all a which are realisations of A, it holds that

$$\mathbb{P}(Y(a) = 1 \mid A = 1, W = w) = \mathbb{P}(Y(a) = 1 \mid A = 0, W = w). \tag{2.8}$$

Hence, it holds that $Y(a) \perp\!\!\!\perp A \mid W = w$. Thus, conditionally randomised experiments produce *conditional exchangeability*.

Moreover, assuming that consistency holds, then conditionally randomised experiments are also cases of study designs where association is causation since

$$\begin{aligned}
&\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] \\
&\quad = \mathbb{E}\left[\mathbb{E}[Y(1) \mid W = w]\right] - \mathbb{E}\left[\mathbb{E}[Y(0) \mid W = w]\right] \\
&\quad = \mathbb{E}\left[\mathbb{P}(Y(1) = 1 \mid W = w)\right] - \mathbb{E}\left[\mathbb{P}(Y(0) = 1 \mid W = w)\right] \\
&\quad = \mathbb{E}\left[\mathbb{P}(Y(1) = 1 \mid A = 1, W = w)\right] - \mathbb{E}\left[\mathbb{P}(Y(0) = 1 \mid A = 0, W = w)\right] \\
&\quad = \mathbb{E}\left[\mathbb{E}[Y(1) \mid A = 1, W = w]\right] - \mathbb{E}\left[\mathbb{E}[Y(0) \mid A = 0, W = w]\right] \\
&\quad = \mathbb{E}\left[\mathbb{E}[Y \mid A = 1, W = w]\right] - \mathbb{E}\left[\mathbb{E}[Y \mid A = 0, W = w]\right]
\end{aligned} \tag{2.9}$$

where we apply the Law of Total Expectation in the first equality, conditional exchangeability in the third equality and consistency in the last equality.

In the following section, we present which conditions are required in order to identify average causal effects in observational studies.

2.3 Identifying Average Causal Effects in Observational Studies

This section is based on [Miguel A. Hernán, 2020, pp. 25–31] and [Neal, 2020, pp. 11–13]. In practice, most studies are not randomised experiments or conditionally randomised experiments but observational studies. Thus, the treatment is not necessarily randomly assigned in observational studies and hence we need a method for handling such studies for causal inference. In these cases, we analyse an observational study as if the treatment was randomised conditional on a set of covariates W. Notice that the result of the analysis relies on this assumption to be true and thus causal inference from observational studies is less convincing than causal inference from (conditionally) randomised experiments. Hence, in order to analyse an observational study, we consider the study as a conditionally randomised experiment. In order to do so, we require three conditions to hold which we refer to as the *identifiability conditions*. These identifiability conditions are:

(*i*) consistency see Equation (2.7),

(*ii*) conditional exchangeability see Equation (2.8),

(*iii*) *positivity*.

Positivity is the condition that for each realisation w of W with $\mathbb{P}(W = w) > 0$ then

$$\mathbb{P}(A = 1 \mid W = w) \in (0, 1). \tag{2.10}$$

The identifiability conditions ensure that we can identify average causal effects from an observational study since

$$\begin{aligned}
\mathbb{E}[Y(1) - Y(0)] &= \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] \\
&= \mathbb{E}\left[\mathbb{E}[Y(1) \mid W = w] - \mathbb{E}[Y(0) \mid W = w]\right] \\
&= \mathbb{E}\left[\mathbb{E}[Y(1) \mid A = 1, W = w] - \mathbb{E}[Y(0) \mid A = 0, W = w]\right] \\
&= \mathbb{E}\left[\mathbb{E}[Y \mid A = 1, W = w] - \mathbb{E}[Y \mid A = 0, W = w]\right]
\end{aligned} \tag{2.11}$$

where the third equality follows from conditional exchangeability and the fourth equality follows from consistency and positivity. Specifically, when assuming that W is discrete and finite, positivity ensures that the expression is well-defined since

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}[Y \mid A = 1, W = w] - \mathbb{E}[Y \mid A = 0, W = w]\right] \\
&= \sum_w \mathbb{P}(W = w)\big(\mathbb{P}(Y = 1 \mid A = 1, W = w) - \mathbb{P}(Y = 1 \mid A = 0, W = w)\big) \\
&= \sum_w \mathbb{P}(W = w)\left(\frac{\mathbb{P}(Y = 1, A = 1, W = w)}{\mathbb{P}(A = 1 \mid W = w)\,\mathbb{P}(W = w)} - \frac{\mathbb{P}(Y = 1, A = 0, W = w)}{\mathbb{P}(A = 0 \mid W = w)\,\mathbb{P}(W = w)}\right)
\end{aligned} \tag{2.12}$$

where we applied the definition of conditional probability in the last equality. Thus, positivity guarantees that the denominator of each of the terms in the last equality in Equation (2.12) is non-zero.

However, in practice, we cannot test whether conditional exchangeability holds. This is due to the fact that there might be additional unmeasured covariates which we need to condition on in order for the treatment and the potential outcomes to be conditionally independent. Thus, we cannot verify that the identifiability conditions hold in observational studies.

In order to be more certain that conditional exchangeability holds, then, in general, the more covariates we condition on, the more likely we are to obtain conditional exchangeability. However, the more covariates we condition on, the less likely it is that positivity holds. This is the case since positivity requires that, conditioned on each covariate value or a combination of covariate values, the probability of each treatment value occurring needs to be positive. However, dividing the population into smaller and smaller subsets increases the probability of receiving one subset where the probability of receiving one of the treatment values is zero. Hence, this is due to the curse of dimensionality. Therefore, there is a trade-off between the fulfilment of the identifiability conditions conditional exchangeability and positivity which further emphasizes the issue of fulfilling the identifiability conditions in observational studies.

In the next section, we consider directed acyclic graphs for visualising the relations between the variables as well as determining whether conditional independence holds.

2.4 Causal Directed Acyclic Graphs

This section is based on [Højsgaard et al., 2012, pp. 7–8, 13], [Miguel A. Hernán, 2020, pp. 70, 83–86, 93–94, 98–101] and [Neal, 2020, pp. 37–39]. In this section, we present a method to determine whether (conditional) exchangeability holds using graphs. Moreover, we consider various biases which arise in the cases where conditional exchangeability is not fulfilled.

Consider a graph $G = (V, E)$ where V denotes the set of vertices and E denotes the set of edges. Here the vertices represent both the observed and unobserved variables while the edges represent the relations between the variables. A *directed acyclic graph* (DAG) is a graph $G = (V, E)$ where the edges are directed and there are no directed cycles within the graph. Specifically, we focus on *causal DAGs*. In this case, the directed edges represent causal effects in the direction of the edge, that is, for variables $V_1, V_2 \in V$ where V_1 has a causal effect on V_2 then there is an edge from V_1 to V_2. Notice that, we in this case refer to V_1 as a *cause* of V_2 and we refer to V_2 as an *effect* of V_1. Furthermore, we note that the acyclic property ensures that a variable cannot cause itself.

An advantage of applying DAGs is that marginal as well as conditional independence can be inferred based on the DAGs. The sets $B, C \subset V$ are marginally independent if they are separated in the graph, that is, there is no path between the sets. In order to determine conditional independence, we first define the *ancestral graph of a set* $B \subset V$ which is the subgraph of G induced by the union of B and its ancestors. Furthermore, we define *moralisation* which is to add an undirected edge between each pair of parents and replace all directed edges in the DAG with undirected edges. Thus, this forms an undirected graph. We can now define *d-separation*.

Definition 2.4.1. d-separation
Let $B, C, D \subset V$. Then B and C are *d-separated* by D if and only if B and C are separated by D in the moralised ancestral graph of $B \cup C \cup D$. ◀

The sets $B, C \subset V$ are conditionally independent given $D \subset V$ if B and C are d-separated by D since d-separation yields conditional independence by the directed global Markov property [Lauritzen, 1996, pp. 32–47].

In the following, we present an example of applying d-separation for determining conditional independence.

Example 2.4.2. d-separation and Conditional Independence
Consider the causal DAG in Figure 2.1.

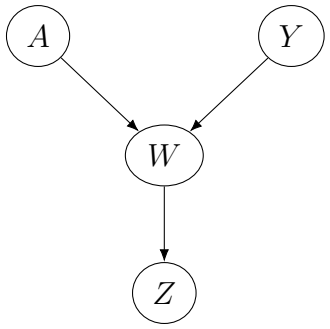


Figure 2.1: Causal DAG where A is the treatment, Y is the outcome, W is a common effect of A and Y and Z is a descendent of the common effect W.

The aim is to determine whether $A \perp\!\!\!\perp Y \mid W = w$ based on the causal DAG in Figure 2.1. This is done by applying d-separation from Definition 2.4.1. The procedure is shown in the following figures.

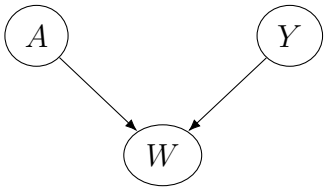


Figure 2.2: The ancestral graph of $A \cup Y \cup W$.

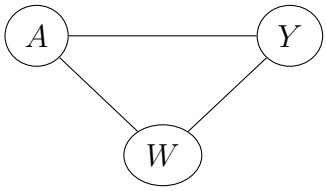


Figure 2.3: Moralised ancestral graph of $A \cup Y \cup W$.

Figure 2.4: Removal of W and its edges.

First, we determine the ancestral graph of $A \cup Y \cup W$ which is shown in Figure 2.2. Then we moralise this ancestral graph of $A \cup Y \cup W$ in Figure 2.3. At last we remove W and its corresponding edges in Figure 2.4 in order to determine if A and Y are separated in the resulting graph. Since A and Y are not separated in Figure 2.4, then A and Y are not d-separated by W and hence they are not conditionally independent given W. Therefore, in general, conditioning on a common effect of two variables $V_1, V_2 \in V$ does not yield conditional independence of V_1 and V_2 given the common effect. Furthermore, in general, conditioning on any descendent of a common effect of the treatment and the outcome such as Z in Figure 2.1 also does not yield conditional independence. ◀

We want to use causal DAGs to determine whether exchangeability or conditional exchangeability holds. However, we note that the causal DAGs do not include the potential outcomes explicitly. Thus, in order to determine (conditional) exchangeability, we introduce a *single-world intervention graph* (SWIG) which is a causal DAG which include the potential outcomes explicitly. Specifically, SWIGs consider the case where all individuals are assigned to a particular treatment value, that is, we intervene on the treatment. We describe interventions in further

details in Section 3.1. In the following, we present an example of applying a SWIG in order to determine whether conditional exhangeability holds in a particular case.

Example 2.4.3. SWIGs and Conditional Exchangeability
Consider the causal DAG and corresponding SWIG in Figures 2.5 and 2.6.

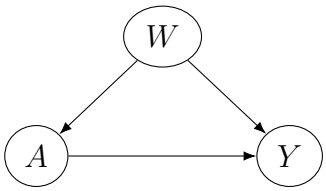


Figure 2.5: Causal DAG where W is a common cause of the treatment A and the outcome Y.

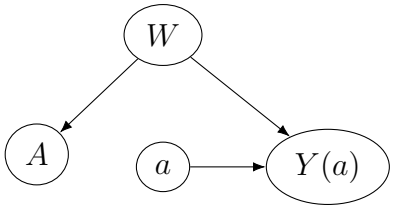


Figure 2.6: The SWIG corresponding to the causal DAG in Figure 2.5 under treatment assignment $A = a$.

On Figure 2.6, a denotes the assignment of all individuals to treatment $A = a$ and W is a common cause of the treatment A and the potential outcome $Y(a)$. Using d-separation, we can conclude that, $Y(a) \perp\!\!\!\perp A \mid W = w$, that is, conditional exhangeability holds. Hence, conditioning on the common cause of the treatment and the potential outcome blocks the association caused by this non-causal path between the treatment and the potential outcome. Thus, if we assume that positivity and consistency hold, then since conditional exchangeability holds, association is causation and we can identify the average causal effect. ◀

Having illustrated how conditional exchangeability can be determined based on a SWIG, we in the remainder of this master's thesis apply causal DAGs where the relation to the potential outcomes is as illustrated in Example 2.4.3.

We note that when (conditional) exchangeability does not hold, then association is not causation and we cannot identify the average causal effect. In these cases biases such as *selection bias* and *confounding* are introduced. Selection bias is the case where we condition on a common effect or a descendant of a common effect as in Example 2.4.2. As we saw in the example, conditioning on a common effect W or a descendant of the common effect Z does not yield conditional independence of A and Y and therefore conditional exchangeability is not fulfilled. Moreover, selection bias can also occur in more general cases. Specifically, selection bias of the treatment and the outcome can be defined as the bias introduced by conditioning on (a descendent of) a common effect of two variables. Here one of the variables is either the treatment or associated with the treatment and the other variable is either the outcome or associated with the outcome. Associated with for example the treatment in this context refers to the cases:

i) an ancestor of the treatment,

ii) a descendent of an ancestor of the treatment.

In the figures below, we illustrate examples of more general cases of selection bias of the treatment, A, and the outcome, Y, when conditioning on a variable C.

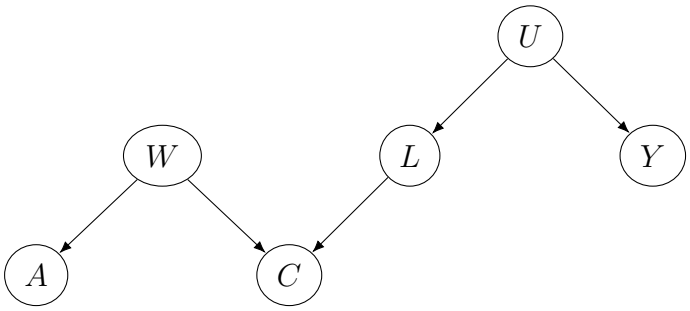


Figure 2.7: Selection bias of A and Y where C is a common effect of W which is a cause of A and L which is an effect of a cause U of Y.

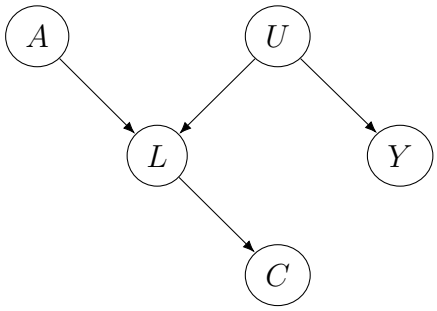


Figure 2.8: Selection bias of A and Y where C is an effect of L which is a common effect of A and U which is a cause of Y.

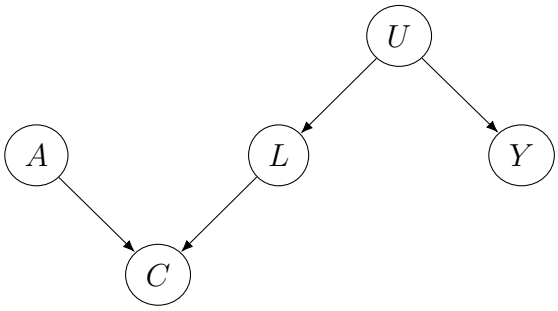


Figure 2.9: Selection bias of A and Y where C is a common effect of A and L which is an effect of U which is a cause of Y.

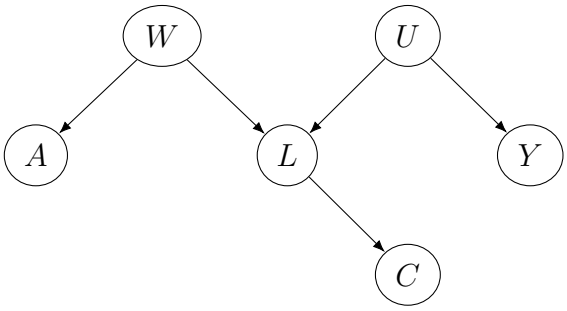


Figure 2.10: Selection bias of A and Y where C is an effect of the common effect L of W which is a cause of A and U which is a cause of the Y.

Figures 2.7–2.10 illustrate cases of selection bias of A and Y when conditioning on C. Applying d-separation, we conclude that A and Y are not d-separated by C in any of the Figures 2.7–2.10 and thus A and Y are not conditionally independent given C. Therefore, conditional exchangeability does not hold. Hence, when conditioning on C, a selection bias is introduced which implies that association is not causation.

Confounding is the bias caused by a common cause of two variables $V_1, V_2 \in V$. When confounding of the treatment and the outcome is present, exchangeability does not hold since the treatment is not randomised. However, by conditioning on the common cause, we get conditional exchangeability as shown in Example 2.4.3.

Since we require conditional exchangeability to be fulfilled in observational studies, we are interested in determining whether there exists a set of covariates W for which conditional exchangeability holds. That is, we want to block the paths of association corresponding to the confounding variables. These path are called *back-door paths*. A back-door path is a non-causal path between the treatment and the outcome where treatment is an effect of a variable on this path. If all back-

door paths between A and Y are blocked by conditioning on W where W does not contain any descendants of A, then we say that W satisfies the *back-door criterion*. In such a case, we refer to W as a *sufficient set of confounding adjustments* since it is sufficient to adjust for the variables in W in order to block all confounding. Hence, when considering an observational study, the aim is to determine a sufficient set of confounding adjustments in order to obtain conditional exchangeability such that the average causal effect can be identified. In this case, the average causal effect can be identified using the *back-door adjustment*.

Proposition 2.4.4. Back-door Adjustment
Let W be a sufficient set of confounding adjustment. Then for all a which are realisations of A and for all y which are realisations of Y, it holds that

$$\mathbb{P}(Y(a) = y) = \sum_{w} \mathbb{P}(Y = y \mid A = a, W = w)\, \mathbb{P}(W = w). \tag{2.13}$$

◀

The proof of this proposition is omitted here since the back-door adjustment coincides with the standardisation method derived in Section 2.5.

When we have unmeasured confounding, we cannot use the methods presented in Section 2.5 to adjust for confounding. In this case, we rely on other methods depending on the causal DAG. Consider for example the causal DAG in the following figure where M denotes measured variables and U denotes unmeasured variables.

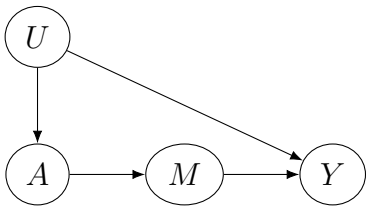


Figure 2.11: Unmeasured confounding in a causal DAG.

In Figure 2.11, we have unmeasured confounding of the treatment and the outcome while also having measured effects of treatment which are causes of the outcome. In such a case, we can use the *front-door criterion* in order to identify the average causal effect.

Proposition 2.4.5. Front-Door Criterion
Consider the causal DAG in Figure 2.11 and assume that positivity and consistency hold for M and Y. Then, for all a which are realisations of A, the front-door criterion yields

$$\mathbb{P}(Y(a) = 1) = \sum_{m} \mathbb{P}(M = m \mid A = a) \sum_{a'} \mathbb{P}(Y = 1 \mid M = m, A = a')\, \mathbb{P}(A = a'). \tag{2.14}$$

◀

Proof. By the Law of Total Probability, it holds that

$$\mathbb{P}(Y(a) = 1) = \sum_m \mathbb{P}(M(a) = m)\, \mathbb{P}(Y(a) = 1 \mid M(a) = m). \tag{2.15}$$

Since there is no confounding of A and M then $A \perp\!\!\!\perp M(a)$. Thus, we can write the first factor on the right hand side in Equation (2.15) as

$$\mathbb{P}(M(a) = m) = \mathbb{P}(M(a) = m \mid A = a) = \mathbb{P}(M = m \mid A = a) \tag{2.16}$$

where the last equality follows from consistency of M.

Now, $Y(a) = Y(m)$ for $M(a) = m$ since A only affects Y through M. Thus, the second factor on the right hand side of Equation (2.15) can be written as

$$\begin{aligned} \mathbb{P}(Y(a) = 1 \mid M(a) = m) &= \mathbb{P}(Y(m) = 1 \mid M(a) = m) \\ &= \mathbb{P}(Y(m) = 1) \end{aligned} \tag{2.17}$$

where the last equality follows from $Y(m) \perp\!\!\!\perp M(a)$ which follows from considering the SWIG corresponding to the causal DAG in Figure 2.11 under assignments $A = a$ and $M = m$. Furthermore, we note that there exists a back-door path between M and Y where A is a sufficient set for confounding adjustments. Hence, $Y(m) \perp\!\!\!\perp M \mid A$. Thus, we can apply the Back-door Adjustment 2.4.4 since A is a sufficient set for confounding adjustments which yields

$$\mathbb{P}(Y(m) = 1) = \sum_{a'} \mathbb{P}(Y = 1 \mid M = m, A = a')\, \mathbb{P}(A = a'). \tag{2.18}$$

Thus, combining Equations (2.16)–(2.18) with Equation (2.15) yields the desired result. ■

In the following section, we present methods which can be used to adjust for measured confounding and identify the average causal effect.

2.5 The Standardisation and Inverse Probability Weighting Methods

This section is based on [Pearl, 2010, pp. 3–4, 18–19], [Miguel A. Hernán, 2020, pp. 19–24] and [Neal, 2020, pp. 68–70]. In this section, we present two methods for identifying a possible average causal effect in a conditionally randomised experiment or an observational study where the identifiability conditions hold.

In order to measure a causal effect of a treatment on an outcome, we first consider the *causal risk difference* in stratum $W = w$ which is given by

$$\mathbb{P}(Y(1) = 1 \mid W = w) - \mathbb{P}(Y(0) = 1 \mid W = w). \tag{2.19}$$

Then applying the identifiability conditions yields

$$\begin{aligned}
&\mathbb{P}(Y(1)=1 \mid W=w) - \mathbb{P}(Y(0)=1 \mid W=w) \\
&\quad = \mathbb{P}(Y(1)=1 \mid A=1, W=w) - \mathbb{P}(Y(0)=1 \mid A=0, W=w) \\
&\quad = \mathbb{P}(Y=1 \mid A=1, W=w) - \mathbb{P}(Y=1 \mid A=0, W=w)
\end{aligned} \tag{2.20}$$

which we call the *risk difference* in stratum $W = w$. Notice that the risk difference in stratum $W = w$ does not depend on the potential outcomes.

Now we present the *standardisation* method which can be used to identity the average causal effect. By applying the Law of Total Probability and the identifiability conditions, then for all a which are realisations of A and for all y which are realisations of Y, it holds that

$$\begin{aligned}
\mathbb{P}(Y(a)=y) &= \sum_w \mathbb{P}(Y(a)=y \mid W=w)\,\mathbb{P}(W=w) \\
&= \sum_w \mathbb{P}(Y(a)=y \mid A=a, W=w)\,\mathbb{P}(W=w) \\
&= \sum_w \mathbb{P}(Y=y \mid A=a, W=w)\,\mathbb{P}(W=w).
\end{aligned} \tag{2.21}$$

Thus, the average causal effect can be identified as

$$\begin{aligned}
&\mathbb{P}(Y(1)=1) - \mathbb{P}(Y(0)=1) \\
&\quad = \sum_w \big(\mathbb{P}(Y=1 \mid A=1, W=w) - \mathbb{P}(Y=1 \mid A=0, W=w)\big)\,\mathbb{P}(W=w).
\end{aligned} \tag{2.22}$$

Thus, when using the standardisation method, we can identify the average causal effect as a weighted average of the risk differences in each stratum $W = w$.

Another method for identifying the average causal effect is the *inverse probability weighting* (IP-Weighting) method. Expanding on Equation (2.21), we obtain

$$\begin{aligned}
\mathbb{P}(Y(a)=y) &= \sum_w \mathbb{P}(Y=y \mid A=a, W=w)\,\mathbb{P}(W=w) \\
&= \sum_w \frac{\mathbb{P}(Y=y, A=a, W=w)}{\mathbb{P}(A=a, W=w)} \frac{\mathbb{P}(A=a, W=w)}{\mathbb{P}(A=a \mid W=w)} \\
&= \sum_w \frac{\mathbb{P}(Y=y, A=a, W=w)}{\mathbb{P}(A=a \mid W=w)}
\end{aligned} \tag{2.23}$$

where the second equality follows from the definition of conditional probability. The fractions $\mathbb{P}(A=a \mid W=w)^{-1}$ are called the *inverse probability weights* and thereby the name inverse probability weighting. Thus, we can identify the average causal effect by

$$\begin{aligned}
&\mathbb{P}(Y(1)=1) - \mathbb{P}(Y(0)=1) \\
&\quad = \sum_w \left(\frac{\mathbb{P}(Y=1, A=1, W=w)}{\mathbb{P}(A=1 \mid W=w)} - \frac{\mathbb{P}(Y=1, A=0, W=w)}{\mathbb{P}(A=0 \mid W=w)} \right).
\end{aligned} \tag{2.24}$$

We note that Equation (2.23) shows that the IP-Weighting method and standardisation method are equivalent methods.

The intuition behind the IP-Weighting method is that we create a *pseudo-population* by weighting the original population in order to adjust for confounding. A pseudo-population is a population twice the size of the original population, where each individual has received both values of treatment. Thus, we are able to identify a possible average causal effect in the pseudo population.

We note that if conditional exchangeability holds in the original population, then exchangeability holds in the pseudo-population. We illustrate this by the causal DAGs in Figures 2.12 and 2.13.

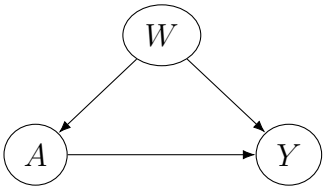


Figure 2.12: Causal DAG for the original population.

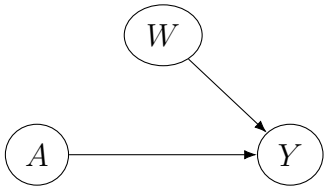


Figure 2.13: Causal DAG for the pseudo-population.

In Figure 2.12, conditional exchangeability holds by Example 2.4.3. If we as mentioned weigh the original population with the inverse probability weight, then $\mathbb{P}(A \mid W) = \mathbb{P}(A)$ such that A and W are independent in the pseudo-population, which implies that exchangeability holds, see Figure 2.13.

By using the standardisation and the IP-Weighting methods, we now show how to identify a possible average causal effect through two examples where we also clarify how the methods differ.

Example 2.5.1. The Standardisation Method
Assume that we have observed the data in Table 2.2.

Table 2.2: Covariate W, treatment A and outcome Y values for each individual.

Individual	W	A	Y
1	1	1	1
2	1	1	1
3	1	1	0
4	1	0	1
5	1	0	1
6	1	0	0
7	0	1	1
8	0	0	1
9	0	1	0
10	0	0	0

We now use the standardisation method to identify a possible average causal effect. First, from Table 2.2, we determine the probabilities

$$\mathbb{P}(W=0)=\frac{2}{5}, \tag{2.25}$$

$$\mathbb{P}(W=1)=\frac{3}{5} \tag{2.26}$$

and for all a which are realisation of A, we obtain

$$\mathbb{P}(Y=1 \mid A=a, W=1)=\frac{2}{3} \tag{2.27}$$

and

$$\mathbb{P}(Y=1 \mid A=a, W=0)=\frac{1}{2}. \tag{2.28}$$

Thus, we can identify the average causal effect, that is,

$$\begin{aligned}
&\mathbb{P}(Y(1)=1)-\mathbb{P}(Y(0)=1)\\
&\quad=\sum_{w}\big(\mathbb{P}(Y=1 \mid A=1, W=w)-\mathbb{P}(Y=1 \mid A=0, W=w)\big)\,\mathbb{P}(W=w)\\
&\quad=\left(\frac{2}{3}-\frac{2}{3}\right)\frac{3}{5}+\left(\frac{1}{2}-\frac{1}{2}\right)\frac{2}{5}\\
&\quad=0.
\end{aligned} \tag{2.29}$$

Hence, we can conclude that treatment A has no average causal effect on the outcome Y. ◀

Example 2.5.2. The IP-weighting Method
Consider again the population in Table 2.2. From this table, for all a which are realisations of A and w which are realisations of W, we determine the probabilities

$$\mathbb{P}(A=a \mid W=w)=\frac{1}{2} \tag{2.30}$$

and

$$\mathbb{P}(Y=1, A=a, W=1)=\frac{1}{5} \tag{2.31}$$

and

$$\mathbb{P}(Y=1, A=a, W=0)=\frac{1}{10}. \tag{2.32}$$

Hence, we obtain the average causal effect

$$\begin{aligned}
&\mathbb{P}(Y(1)=1)-\mathbb{P}(Y(0)=1)\\
&\quad=\sum_w \frac{\mathbb{P}(Y=1,A=1,W=w)}{\mathbb{P}(A=1\mid W=w)}-\sum_w \frac{\mathbb{P}(Y=1,A=0,W=w)}{\mathbb{P}(A=0\mid W=w)}\\
&\quad=2\left(\sum_w \mathbb{P}(Y=1,A=1,W=w)-\sum_w \mathbb{P}(Y=1,A=0,W=w)\right) \qquad (2.33)\\
&\quad=2\left(\frac{1}{5}+\frac{1}{10}-\frac{1}{5}-\frac{1}{10}\right)\\
&\quad=0
\end{aligned}$$

where the second equality follows from Equation (2.30). Hence, the treatment A has no average causal effect on the outcome Y. Moreover, we notice that we obtain the same result as when using the standardisation method in Example 2.5.1. ◀

3 | Targeted Maximum Likelihood Estimation

In this chapter, we present the *targeted maximum likelihood estimation* (TMLE) method. The TMLE method consists of two steps which aim to obtain an optimal bias-variance trade-off for a parameter of interest. Thus, we first present *structural causal models* in order to define this parameter of interest. Then we present the two steps of the TMLE method in order to estimate the average causal effect. The following is based on [van der Laan and Rose, 2011, pp. 7–9, 21–22].

Let $O = (Y, A, W) \sim P_0$ be a random vector with true distribution P_0. Furthermore, let $\mathcal{M}$ consist of the candidate distributions of O and let W denote a random vector consisting of covariate information. In addition, assume that for the random vector O, we have an i.i.d. sample $o_1, \ldots, o_n$. We want to define a function $\Psi : \mathcal{M} \to \mathbb{R}^d$ where $\Psi(P_0)$ is called the *target parameter*. Thus, in Section 3.1, we define this function and in Sections 3.2 and 3.3, we describe how to obtain an estimator of the target parameter. Notice that, the target parameter is the parameter of interest.

Statistical models are in practice often misspecified which introduces a bias. In this case, the target parameter is not defined as a parameter of any of the possible probability distributions since it is defined within the parametric statistical models assuming that the model is correctly specified. Hence, in the first step of the TMLE method, the initial estimator obtained in this step is estimated using only semi-parametric and non-parametric models. Moreover, notice that the Food and Drug Administration (FDA) does not allow parametric statistical models for causality assessments. Specifically, it is required that the statistical model reflects true knowledge and since parametric statistical models rely on assumptions of the underlying distribution of the observed data, the use of such models is troublesome.

In the following section, we describe how to use structural causal models to specify the causal question and define the target parameter.

3.1 Structural Causal Models

This section is based on [van der Laan and Rose, 2011, Chapter 2], [Pearl, 2009, Section 1.4.1] and [Neal, 2020, Section 4.5.2]. In this section, we present structural causal models which can be viewed as statistical models with possible additional non-testable assumptions. We refer to such a statistical model with these additional assumptions as a *model*. This model will be used to specify the causal question and define the target parameter $\Psi(P_0)$. Furthermore, we restrict this class of models in order to identify the causal parameter by assuming that the identifiability conditions hold.

Let (U, X) be the full data where $X = (X_1, \ldots, X_J)$ and $U = (U_{X_1}, \ldots, U_{X_J})$ are random vectors with entries X_j and U_{X_j} for $j = 1, \ldots, J$ which are called the *endogenous and exogenous variables*, respectively. An exogenous variable is unobserved and causes X_j, thus the notation U_{X_j}, and is not caused by any other variables. An endogenous variable, X_j, is a function of its causes, that is, its parents among the endogenous variables, which we denote as $\text{pa}(X_j)$, and U_{X_j}. Thus, we can write X_j as

$$X_j = f_{X_j}(\text{pa}(X_j), U_{X_j}) \tag{3.1}$$

where f_{X_j} is a deterministic function. We call such an equation a *structural equation*. Having defined a structural equation, we now define a structural causal model.

Definition 3.1.1. Structural Causal Model
Let (U, X) be the full data. Then a model is called a structural causal model (SCM) if the endogenous variables can be expressed as a distinct structural equations. ◀

We now present an example which illustrates the relation between causal DAGs and SCMs.

Example 3.1.2. Causal DAG and SCM
Consider the SCM for the full data $(U, X) = (U_W, U_A, U_Y, W, A, Y)$ with the following structural equations for the endogenous variables

$$W = f_W(U_W), \tag{3.2}$$
$$A = f_A(W, U_A), \tag{3.3}$$
$$Y = f_Y(A, W, U_Y). \tag{3.4}$$

For this SCM, we can illustrate the relations between the exogenous and endogenous variables by using a causal DAG. The corresponding causal DAG for this SCM is shown in Figure 3.1.

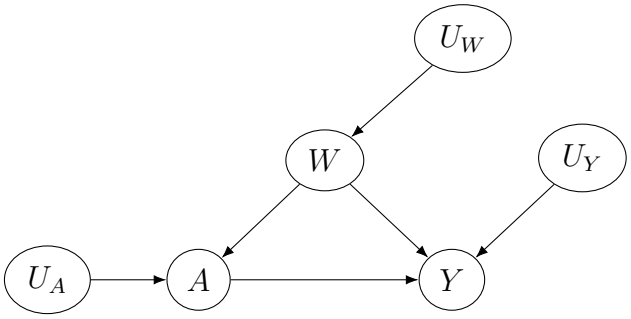


Figure 3.1: Causal DAG for the SCM with structural equations given by Equations (3.2)–(3.4).

◀

In the following, we assume that the exogenous variables have the joint distribution P_U and that the distribution $P_{U,X}$ of (U, X) implies the distribution P of O. Since we for the SCM have assumed that the distribution of (U, X) implies the distribution of O, we use the following notation $P := P(P_{U,X})$. The sets of possible distributions of (U, X) and O are denoted $\mathcal{M}^F$ and $\mathcal{M} = \{P(P_{U,X}) : P_{U,X}\}$, respectively. Specifically, $\mathcal{M}^F$ is restricted to the models for which the possible additional non-testable assumptions hold. Examples of such non-testable assumptions could be assumptions on P_U and that there is no unmeasured confounding of the treatment and the outcome. Note that since the true distribution of (U, X), that is, $P_{U,X,0}$, implies the true distribution of O, that is, P_0, it is possible to specify the SCM such that we obtain the true distribution of O.

In order to determine the distribution of O from the SCM, we need to specify the relation between O and X. In general, we assume that $O = \Phi(X)$ for some function Φ. By the Law of Total Probability, we obtain

$$P_X(X = x) = \sum_u P_f(X = x \mid U = u) P_U(U = u) \tag{3.5}$$

where $f = \{f_{X_j} : j = 1, \ldots, J\}$.

If we consider the special case $O = I(X)$, where I is the identity function, it follows that

$$P(X = o) = \sum_u P_f(X = o \mid U = u) P_U(U = u). \tag{3.6}$$

Notice that $O = X$ corresponds to the case where we for each observation of O observe all the endogenous variables.

As mentioned in Section 2.1, we cannot observe both potential outcomes for an individual. However, we note that we can use SCMs to determine all possible potential outcomes for an individual. In order to do so, we first introduce the notation $do(A = a)$ which indicates that we assign each individual to treatment a. Notice that using this notation then

$$\mathbb{P}\left(Y(a) = a\right) = \mathbb{P}\left(Y = a \mid do(A = a)\right). \tag{3.7}$$

An assignment, $do(A = a)$, is called an *intervention* and, in this case, A is called an *intervening variable*. We note that other observed endogenous variables than the treatment also can be intervening variables. In order to compute the potential outcomes, we assume the *modularity assumption for SCMs*. That is, we assume that intervening on a variable, for example $do(A = a)$, does not change the form of the structural equations for the remaining variables for the corresponding SCM denoted M_a. By assuming the modularity assumption, we have that $Y_i(a) = Y_{i,M_a}(a)$ which is called the *law of potential outcomes*. That is, the potential outcome in the original SCM equal the potential outcome for the SCM corresponding to the intervention $do(A = a)$. In this setup, we now use the notation $Y_a(i) = Y_{M_a}(i)$.

Interventions can be used to specify the target parameter since, as mentioned above, we can determine the potential outcomes based on the SCM. The potential outcomes for the SCM in Example 3.1.2 can for example be expressed as

$$Y_a(U) = f_Y(a, W, U_Y) \tag{3.8}$$

for all a which are realisations of A. In the following, we present interventions in a more general setup. Let $X = (A, L)$ where $A = (A_1, \dots, A_S)$ and $L = (L_1, \dots, L_{J-S})$ where A_s for $s = 1, \dots, S$ and L_r for $r = 1, \dots, J - S$ denote the intervening variables and non-intervening variables, respectively, and where J is the number of variables in X. An intervention is said to fulfil a *static intervention rule* if we assign treatment $A = a$ and a *dynamic intervention rule* if the intervention is determined by its parents. Thus, we let $d = \{d_s : s = 1, \dots, S\}$ denote the set of rules for each intervention.

Under one or more interventions, we denote the non-intervening endogenous variables by $L_{r,d}(U)$ for $r = 1, \dots, J - S$ which are called *post-intervention random variables*. An example of a post-intervention random variable is the potential outcome in Equation (3.8) where the static intervention rule is $do(A = a)$. By applying the Law of Total Probability, the distribution of $L_{r,d}(U)$ is given by

$$\begin{aligned} P(L_{r,d}(U) = l) &= \sum_u P_f(L_{r,d}(u) = l \mid U = u) P_U(U = u) \\ &= \sum_u \mathbb{1}[L_{r,d}(u) = l] P_U(U = u). \end{aligned} \tag{3.9}$$

Thus, we note that we can define the target parameter to be a function of $P(L_{r,d}(U) = l)$.

Now, let $\Psi^F \colon \mathcal{M}^F \to \mathbb{R}^d$ where $\Psi^F(P_{U,X,0})$ is the target parameter for $P_{U,X,0} \in \mathcal{M}^F$. By assuming that the identifiability conditions hold and applying that the distribution of (U, X) implies the distribution of O, we obtain

$$\Psi^F(P_{U,X}) = \Psi(P) \quad \forall P_{U,X} \in \mathcal{M}^{F*} \tag{3.10}$$

where $\mathcal{M}^{F*} \subseteq \mathcal{M}^F$ is a restricted set of models where the identifiability conditions hold and $\Psi \colon \mathcal{M} \to \mathbb{R}^d$ for $\mathcal{M} = \{P(P_{U,X}) : P_{U,X} \in \mathcal{M}^{F*}\}$. Thus, we can define the target parameter $\Psi(P_0)$ as $\Psi^F(P_{U,X,0}) = \Psi(P_0)$.

If we want to obtain an estimator of the average causal effect then we consider $\Psi \colon \mathcal{M} \to \mathbb{R}$ where the target parameter is given by

$$\Psi(P_0) = \mathbb{E}_{W,0}\left[\mathbb{E}_0[Y \mid A = 1, W = w] - \mathbb{E}_0[Y \mid A = 0, W = w]\right]. \tag{3.11}$$

We call the estimator of the target parameter the *target maximum likelihood estimator* which we refer to as the TML estimator. Notice that the notation $\mathbb{E}_0$ refers to the expected value with respect to P_0 and thus the target parameter depends on P_0 through these expected values.

In the following section, we present the first step of the TMLE method when the target parameter is given as in Equation (3.11).

3.2 Initial Estimator and the Super Learner Method

This section is based on [van der Laan and Rose, 2011, Chapter 3]. In this section, we describe the first step of the TMLE method. The first step of the TMLE method consists of determining an initial estimator $\bar{Q}_n^0$ of $\bar{Q}_0 \coloneqq \mathbb{E}_0[Y \mid A = a, W = w]$ where

$$\bar{Q}_0 \coloneqq \arg\min_{\bar{Q}} \left(\mathbb{E}_0 \left[L(O, \bar{Q}) \right] \right) \tag{3.12}$$

where L is a uniformly bounded loss function and $\bar{Q}$ are the candidates of $\bar{Q}_0$. We note that for $o = (y, a, w)$ being a realisation of O, then if Y is binary we can use either the negative log loss function given by

$$L(o, \bar{Q}) = -\log \left(\bar{Q}(a, w)^y \left(1 - \bar{Q}(a, w) \right)^{1-y} \right) \tag{3.13}$$

or the squared error loss function given by

$$L(o, \bar{Q}) = \left(y - \bar{Q}(a, w) \right)^2. \tag{3.14}$$

The initial estimator is obtained by applying the *super learner* method. In the following, we show how to obtain an initial estimate of $\bar{Q}_0$ using this method.

Consider m different regression methods to be applied in the super learner method. Notice that when considering methods containing tuning parameters then different tuning parameters result in different methods in this context. For instance consider the elastic nets method which is a convex sum of the Lasso and Ridge methods. Thus, the Lasso and Ridge methods are considered as different methods since they are special cases of the elastic nets method for different tuning parameter values.

Let $Z_p^{(1)}, \ldots, Z_p^{(k)}$ for $p = 1 \ldots, m$ be the predictions for the m methods by applying k-fold cross validation for each of the methods. The k-fold *cross validation risks* for the m methods are then calculated as follows

$$\bar{Z}_p = \frac{1}{k} \sum_{i=1}^{k} Z_p^{(i)} \tag{3.15}$$

for $p = 1, \ldots, m$. Then the super learner method determines the initial estimate by a weighted average of the cross validation risks, that is,

$$\sum_{p=1}^{m} \alpha_p \bar{Z}_p \tag{3.16}$$

where $\alpha_p \geqslant 0$ is an entry in $\alpha = (\alpha_1, \ldots, \alpha_m)$ with a sum to one constrain, $\sum_{p=1}^{m} \alpha_p = 1$. This is obtained by applying k-fold cross validation again which yields

$$\hat{\alpha} \coloneqq \arg\min_{\alpha} \left(\frac{1}{n} \sum_{i=1}^{n} L \left(o_i, \sum_{p=1}^{m} \alpha_p \bar{Z}_p \right) \right) \tag{3.17}$$

where $\hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_p)$.

At last, the m methods are fitted on the complete data set in order to obtain predictions $\bar{Z}_p^{\text{comp}}$ for $p = 1, \ldots m$. Thus, the super learner method obtains the initial estimate

$$\bar{Q}_n^0(a, w) = \sum_{p=1}^{m} \hat{\alpha}_p \bar{Z}_p^{\text{comp}}. \tag{3.18}$$

We note that the super learner method can be computationally heavy. By applying parallel programming when using k-fold cross validation for the m methods, we can separate the calculations for each of the m methods such that the super learner method becomes more computationally efficient.

Having obtained an initial estimator, we note that this initial estimator is not targeted towards the target parameter. The aim of the super learner method is to obtain an overall optimal fit of $\mathbb{E}_0[Y \mid A = a, W = w]$. Hence, this can result in a non-optimal bias-variance trade-off for the target parameter. Thus, in the following section, we describe how to update the initial estimator in order to obtain a more optimal bias-variance target parameter.

3.3 Targeting the Initial Estimator

This section is based on [van der Laan and Rose, 2011, pp. 37–39, Chapters 4–5, p. 459]. In this section, we present the second step of the TMLE method. Specifically, we in this step update the initial estimator obtained in the first step.

Recall that the target parameter for an average causal effect is given by

$$\Psi(P_0) = \mathbb{E}_{W,0} \left[\mathbb{E}_0[Y \mid A = 1, W = w] - \mathbb{E}_0[Y \mid A = 0, W = w] \right]. \tag{3.19}$$

The expected value over W can be estimated using the empirical mean which yields

$$\hat{\Psi}(Q_n) = \frac{1}{n} \sum_{i=1}^{n} \left(\bar{Q}_n(1, w_i) - \bar{Q}_n(0, w_i) \right) \tag{3.20}$$

where $Q_n = (\bar{Q}_n, Q_{W,n})$ where $\bar{Q}_n$ is an estimator of $\bar{Q}_0$ and $Q_{W,n}$ is the empirical distribution for the marginal distribution of W.

Now in the second step of the TMLE method, the aim is to obtain $\bar{Q}_n$ by updating the initial estimator of $\bar{Q}_0$ obtained in the first step of the TMLE method. Notice that, we only need to update the initial estimator $\bar{Q}_n^0$ since $Q_{W,n}^0$ is estimated as the sample mean which is an unbiased estimator and thus cannot produce bias for the target parameter.

Hence, the initial estimator $\bar{Q}_n^0$ is updated by targeting the estimator toward the parameter of interest. In the first step of the TMLE method, we estimated $\mathbb{E}_0[Y \mid A = a, W = w]$ as the overall optimal estimate. However, considering the causal question, we are only interested in the average causal effect of the treatment on the outcome. Thus, considering the treatment and the covariates W on equal terms does not exploit the fact that the treatment is the only variable of interest. Hence, we target the estimation towards treatment, and thus, we can obtain a less biased estimator for the parameter of interest while possibly obtaining more biased estimators for W.

In order to do so, we need to estimate $P_0(A = a \mid W = w)$ where we denote the estimator as $g_n(A \mid W = w)$. Specifically, we estimate $g_n(1 \mid W = w)$ and $g_n(0 \mid W = w)$, that is, we estimate the probability of treatment given the covariates $W = w$ under both values of treatment for all individuals. These estimators can be used to target the estimation towards the target parameter. This is done through the *clever covariate*. The expression of the clever covariate depends on the target parameter. For the target parameter for the average causal effect then the clever covariate is defined as

$$H_n^*(A, W) = \frac{\mathbb{1}[A = 1]}{g_n(1 \mid W = w)} - \frac{\mathbb{1}[A = 0]}{g_n(0 \mid W = w)}. \tag{3.21}$$

Note for each individual with $A = 1$, then

$$H_n^*(1, W) = \frac{1}{g_n(1 \mid W = w)} \tag{3.22}$$

and, analogously, for each individual with $A = 0$, then

$$H_n^*(0, W) = -\frac{1}{g_n(0 \mid W = w)}. \tag{3.23}$$

Hence, the clever covariate serves as a combined covariate of the (negative) reciprocal of the estimator of the probability of treatment given the additional covariates $W = w$. Furthermore, we note that we estimate g_n using the super learner method.

When updating the initial estimator $\bar{Q}_n^0$, we apply a parametric model of the outcome Y on the clever covariate where we use the initial estimator or a transformation of the initial estimator as the intercept. The parameter, ε, of this model is called the *fluctuation parameter* since it represents the fluctuation of the initial fit. For example, if the parametric model was a logistic regression, we get

$$\text{logit}\left(\bar{Q}_n^1(A, W)\right) = \text{logit}\left(\bar{Q}_n^0(A, W)\right) + \varepsilon_n^0 H_n^*(A, W) \tag{3.24}$$

where ε_n^0 is the coefficient of the clever covariate from the logistic regression and an estimator of the fluctuation parameter. Notice that since the outcome Y is assumed to be binary with values zero and one, then using the logistic regression ensures that the estimator is within the range of Y. Hence, we have updated the estimator of $\bar{Q}_0$ and this process is iterated until convergence, that is, $\varepsilon_n^k = 0$ where k denotes the k'th iteration.

If we were to update the estimator $\bar{Q}_n^1$ from Equation (3.24) then the update is

$$\text{logit}\left(\bar{Q}_n^2(A,W)\right) = \text{logit}\left(\bar{Q}_n^1(A,W)\right) + \varepsilon_n^1 H_n^*(A,W) \tag{3.25}$$

which yields $\varepsilon_n^1 = 0$ since we have already included the information of the clever covariate in the first update of $\bar{Q}_n^0$ and thus, there is no additional information in the clever covariate for next update $\bar{Q}_n^1$. Thus, we have convergence in one step. Hence, in order to estimate the target parameter, we now determine $\bar{Q}_n^1(1,W)$ and $\bar{Q}_n^1(0,W)$ by computing

$$\text{logit}\left(\bar{Q}_n^1(1,W)\right) = \text{logit}\left(\bar{Q}_n^0(1,W)\right) + \varepsilon_n^0 H_n^*(1,W) \tag{3.26}$$

and

$$\text{logit}\left(\bar{Q}_n^1(0,W)\right) = \text{logit}\left(\bar{Q}_n^0(0,W)\right) + \varepsilon_n^0 H_n^*(0,W). \tag{3.27}$$

Here the values of $H_n^*(1,W)$ and $H_n^*(0,W)$ were determined by setting $A = 1$ and $A = 0$ for all individuals. Thus, the final estimate of the target parameter in Equation (3.19) for the average causal effect is

$$\hat{\Psi}(Q_n^1) = \frac{1}{n}\sum_{i=1}^{n}\left(\bar{Q}_n^1(1,w_i) - \bar{Q}_n^1(0,w_i)\right). \tag{3.28}$$

Moreover, the TML estimator is an asymptotically linear estimator under regularity conditions. We omit the proof of this property since, in practice, it has been observed that the TMLE method suffers if the initial estimator is too adaptive. Thus, we now present an extension of the TMLE method where an additional layer of cross validation is introduced in order to overcome this issue. Furthermore, we show asymptotic linearity of the estimator obtained by this particular method.

4 Cross Validated Targeted Maximum Likelihood Estimation

In this chapter, we present the *cross validated targeted maximum likelihood estimation* (CV-TMLE) method. Specifically, we outline how the TMLE method presented in Chapter 3 is extended to the CV-TMLE method. Furthermore, we show asymptotic linearity of the estimator obtained from the CV-TMLE method. Moreover, in order to show the asymptotic linearity of this estimator, we also present influence functions.

4.1 Extending the Targeted Maximum Likelihood Estimation Method

This section is based on [Zheng and van der Laan, 2010]. In this section, we present the CV-TMLE method which is an extension of the TMLE method. Specifically, an additional layer of cross validation is added to the TMLE method in order to obtain a more robust estimator in the sense that we avoid the issue of a too adaptive initial estimator.

In the following, we outline how cross validation is used in the CV-TMLE method. Let the random vector $B_n \in \{0,1\}^n$ correspond a split of the n individuals, that is, the set $\{1,\dots,n\}$, into training and validation data sets. We let $\mathcal{T} = \{i : B_n(i) = 0\}$ and $\mathcal{V} = \{i : B_n(i) = 1\}$ denote the training and validation data sets, respectively. That is, the i'th individual belongs to the training data set if $B_n(i)$ attains the value zero or the validation data set if $B_n(i)$ attains the value one. Furthermore, we let P_n, P^0_{n,B_n} and P^1_{n,B_n} denote the empirical distributions for O, $\mathcal{T}$ and $\mathcal{V}$, respectively. Given a cross validation scheme and a parametric model, we define

$$\varepsilon^0_n := \arg\min_{\varepsilon} \left(\mathbb{E}_{B_n} \left[P^1_{n,B_n} L\left(\hat{Q}(P^0_{n,B_n})(\varepsilon) \right) \right] \right) \tag{4.1}$$

where $\mathbb{E}_{B_n}$ denotes the expected value of each split into training and validation data sets and where the initial estimators $\hat{Q}(P^0_{n,B_n})$ are obtained by using a super learner method on the training data set for each B_n. Moreover, notice that we use the notation that for a distribution P then $PS = \int S(o)\mathrm{d}P(o)$ for an integrable function S which for an empirical distribution reduces to the corresponding sum. Thus, in the next iteration, we obtain the estimators $\hat{Q}(P^0_{n,B_n})(\varepsilon^0_n)$ for each split B_n.

Hence, the additional layer of cross validation in the CV-TMLE method is applied such that we fit the initial estimator on the training data set and use the validation data set when targeting the initial estimator. Hence, if a highly adaptive method is used in the super learner method then

since the fluctuation parameter ε is determined based on the validation data set, we avoid the issue of too adaptive methods.

Hence, in general, the k'th update of ε_n^0 is determined by

$$\varepsilon_n^k := \arg\min_{\varepsilon} \left(\mathbb{E}_{B_n} \left[P_{n,B_n}^1 L \left(\hat{Q}(P_{n,B_n}^0)(\varepsilon_n^{k-1}) \right) \right] \right) \tag{4.2}$$

where, for all B_n, the k'th update of the initial estimator is $\hat{Q}(P_{n,B_n}^0)(\varepsilon_n^k)$. We denote the final update as $\hat{Q}^*(P_{n,B_n}^0)(\varepsilon_n^{k-1})$ where $\varepsilon_n^k = 0$. Thus, the ***CV-TML estimator*** of $\Psi(P_0)$ is then given by

$$\hat{\Psi}\left(P_n\right) = \mathbb{E}_{B_n} \left[\Psi \left(\hat{Q}^*(P_{n,B_n}^0)(\varepsilon_n^{k-1}) \right) \right]. \tag{4.3}$$

Notice that in the case where $\varepsilon_n^1 = 0$, we have *one step convergence* where the CV-TML estimator is given by

$$\hat{\Psi}\left(P_n\right) = \mathbb{E}_{B_n} \left[\Psi \left(\hat{Q}^*(P_{n,B_n}^0)(\varepsilon_n) \right) \right] \tag{4.4}$$

where $\varepsilon_n = \varepsilon_n^0$. We later refer to this estimator as *the one step CV-TML estimator*.

If the target parameter is the average causal effect, then the CV-TMLE method yields the following estimate

$$\hat{\Psi}(P_n) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i=1}^{n} \mathbb{1}[B_n(i,k) = 1] \left(\hat{Q}^*(P_{B_n(\cdot,k)}^0)(1, w_i) - \hat{Q}^*(P_{B_n(\cdot,k)}^0)(0, w_i) \right) \tag{4.5}$$

where K denotes the number of folds used in cross validation, $n_k = \sum_{i=1}^{n} B_n(i,k)$ and $P_{B_n(\cdot,k)}^0$ denotes the empirical distribution of the training data set for fold k.

In Section 4.3, we prove asymptotic linearity of the one step CV-TML estimator. In order to do so, we first present influence functions in the following section.

4.2 Influence Functions

This section is based on [Fisher and Kennedy, 2018, pp. 4–12, 20, 27–28] and [van der Laan and Rose, 2011, p. 89]. In this section, we define *influence functions* in a general setting. Furthermore, we define the *score-based influence function* and the *efficient influence function*.

Assume that for a random vector $Z \sim P$, we have observed an i.i.d. sample $z_1, \ldots, z_n$. We are interested in a target of the distribution P which we can write as a functional T of P. In this context, the term functional indicates that the input of T is a distribution and the output is in $\mathbb{R}^d$. An example of a functional or target could be the expected value of Z which we denote as $T_{\text{mean}}(P) := \mathbb{E}_P[Z]$. In order to estimate $T(P)$, we can estimate P, that is, the plug-in estimate is $T(\tilde{P})$ where $\tilde{P}$ is an estimate of P.

Now we are interested in how $T(\tilde{P})$ changes in response if $\tilde{P}$ is slightly improved. Here "slightly improved" refers to the accuracy of the estimator which we clarify in the following. Let p and $\tilde{p}$ be the densities of P and $\tilde{P}$, respectively. Furthermore, let P_ε be the distribution with the density

$$p_\varepsilon(z) = (1 - \varepsilon)p(z) + \varepsilon\tilde{p}(z) \tag{4.6}$$

where $\varepsilon \in [0, 1]$. We later write this distribution as $P_\varepsilon := P + \varepsilon(\tilde{P} - P)$. Hence, for $\varepsilon = 0$ the distribution P_ε reduces to the true distribution and for $\varepsilon = 1$ then P_ε reduces to the estimate $\tilde{P}$. Thus, we say that the accuracy of P_ε improves as ε approaches zero. That is, a slight improvement in $\tilde{P}$ can be viewed as letting ε approach zero for P_ε.

Now let $\mathcal{P} = \{P_\varepsilon\}_{\varepsilon \in [0,1]}$. Thus, $\mathcal{P}$ is a set of possible distributions which we refer to as the path that connects $\tilde{P}$ to P. An example of a path $\mathcal{P}$ is illustrated in Figure 4.1.

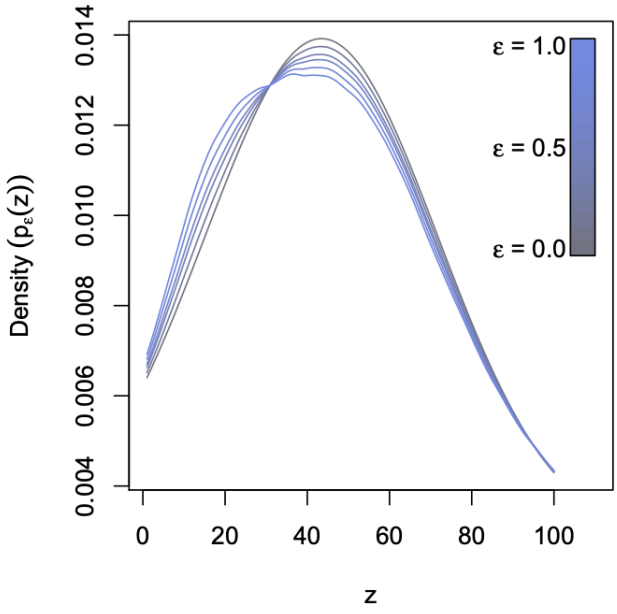


Figure 4.1: A path $\mathcal{P}$ between $\tilde{P}$ and P [Fisher and Kennedy, 2018, p. 8].

Notice that $T(P_\varepsilon)$ exists for each P_ε along this path but, in practice, $T(P_\varepsilon)$ can only be computed for $\varepsilon = 1$. We illustrate the values of T of the path $\mathcal{P}$ as the solid curve on Figure 4.2.

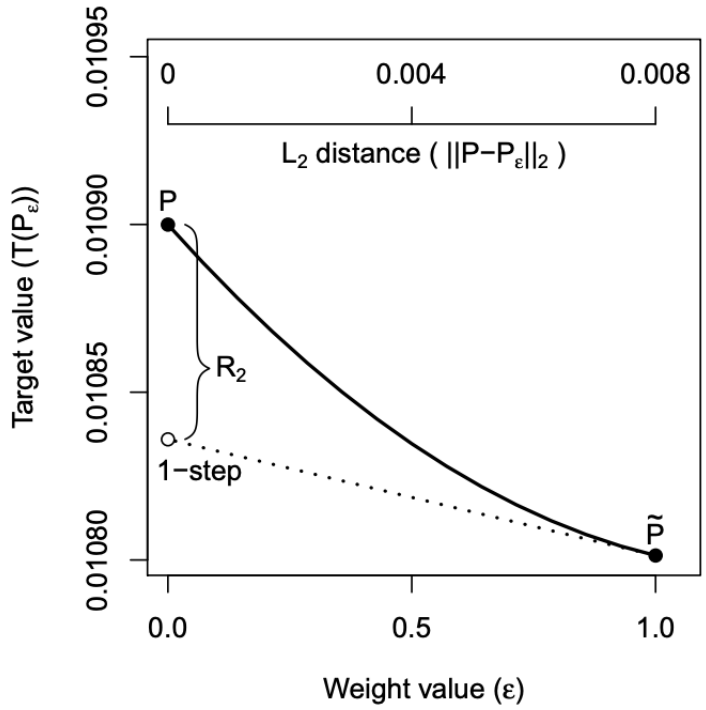


Figure 4.2: $T(P_\varepsilon)$ against ε where the solid curve represents the values of $T(P_\varepsilon)$ on the path $\mathcal{P}$, the dashed line represents the one step update of $\tilde{P}$ and R_2 is the residual [Fisher and Kennedy, 2018, p. 8].

Consider Figure 4.2. As mentioned above, the solid curve represents the value for T for the path $\mathcal{P}$. We refer to the corresponding function as v, that is,

$$v(\varepsilon) = T(P_\varepsilon). \tag{4.7}$$

Since we are interested in $T(P)$, this corresponds to determining the intercept with the y-axis for v since $v(0) = T(P_0) = T(P)$. However, v is unknown and can only be evaluated in $\varepsilon = 1$. Thus, we want to approximate v and evaluated this approximation in $\varepsilon = 0$. Specifically, by assuming that T is pathwise differentiable along $\mathcal{P}$, we can approximate the slope of v at $\varepsilon = 1$, that is,

$$v'(1) = \frac{\partial}{\partial \varepsilon} T(P_\varepsilon)\bigg|_{\varepsilon=1}. \tag{4.8}$$

Using this slope, we can approximate v as a linear function and thus, by evaluating at $\varepsilon = 0$, approximate $T(P)$. This is also illustrated as the dashed line in Figure 4.2. This one step update can be motivated by applying a Taylor expansion of v, that is,

$$\begin{aligned} T(P_0) = v(0) &= v(1) + v'(1)(0-1) - R_2 \\ &= T(P_1) + \frac{\partial}{\partial \varepsilon} T(P_\varepsilon)\bigg|_{\varepsilon=1} (0-1) - R_2 \end{aligned} \tag{4.9}$$

where R_2 is the remainder term which often can be shown to converge to zero. In Section 4.3, we show one example where the remainder term converges to zero by proving asymptotic linearity of the CV-TML estimator.

Now assume that Z is discrete with values in $\{z_1, \ldots, z_K\}$. We can approximate $v'(1)$ by applying the chain rule such that

$$\begin{aligned} \frac{\partial}{\partial \varepsilon} T(P_\varepsilon)\Big|_{\varepsilon=1} &= \sum_{k=1}^{K} \frac{\partial T(P_\varepsilon)}{\partial p_\varepsilon(z_k)} \frac{\partial p_\varepsilon(z_k)}{\partial \varepsilon}\Big|_{\varepsilon=1} \\ &= \sum_{k=1}^{K} \frac{\partial T(P_\varepsilon)}{\partial p_\varepsilon(z_k)}\Big|_{\varepsilon=1} \left(\tilde{p}(z_k) - p(z_k)\right). \end{aligned} \tag{4.10}$$

However, notice that in Equation (4.10), we slightly abuse notation of the partial derivative of $T(P_\varepsilon)$ with respect to $p_\varepsilon(z_k)$ since marginal increases to $p_\varepsilon(z_k)$ for some z_k must result in equal marginal decreases in $p_\varepsilon(z_{k'})$ for some $z_{k'}$ in order for p_ε to be a valid density function. Thus, if this is not fulfilled, then $\frac{\partial T(P_\varepsilon)}{\partial p_\varepsilon(z_k)}$ is ill-defined. In order to avoid this issue, we instead use the influence function defined in the following definition. Notice that, we consider Z as an arbitrary continuous random vector in the following definition.

Definition 4.2.1. Influence Function
Let $L_0^2(P)$ be the subspace of the Hilbert space $L^2(P)$ of mean zero function of Z and let T be a functional. Then the influence function for T is the function $IF \in L_0^2(P)$ which satisfies

$$\frac{\partial}{\partial \varepsilon} T\left(G + \varepsilon(Q - G)\right)\Big|_{\varepsilon=0} = \int IF(z, G)\left(q(z) - g(z)\right) \mathrm{d}z \tag{4.11}$$

for any two distributions G and Q with densities g and q, respectively. ◀

Notice that since $IF \in L_0^2(P)$ then, by definition, the influence function has mean zero, that is,

$$\int IF(z, G) g(z) \mathrm{d}z = 0, \tag{4.12}$$

and is square integrable which in combination implies that it has finite variance.

Moreover, notice that for any two distributions G and Q with densities g and q, then $G + \varepsilon(Q - G)$ in Equation (4.11) is a distribution with density $g(z) + \varepsilon\left(q(z) - g(z)\right)$.

Consider again the setup of Z being a discrete variable with values in $\{z_1, \ldots, z_K\}$. We can isolate the influence function in Equation (4.11) by letting Q be the point mass distribution at point z which is denoted as Δ_z with density δ_z. Notice that, if Z is discrete the integral in Equation (4.11) is a sum over the possible values of Z. Hence, Equation (4.11) reduces to

$$\frac{\partial}{\partial \varepsilon} T\left(G + \varepsilon(\Delta_z - G)\right)\Big|_{\varepsilon=0} = \sum_{k=1}^{K} IF(z_k, G)\delta_z(z_k) = IF(z, G). \tag{4.13}$$

where the first equality follows from Z being discrete and applying Equation (4.11) in combination with Equation (4.12).

Hence, the influence function measures the effect on T for infinitesimally small changes in G. Thus, the influence function can measure the influence of an additional observation (at point z) on the estimator of interest.

Returning to the approximating of $v'(1)$ and Z being a random vector where we have observed $z_1, \ldots, z_n$, we can use the influence function for this approximation which yields

$$\begin{aligned}
\frac{\partial}{\partial \varepsilon} T(P_\varepsilon)\Big|_{\varepsilon=1} &= \frac{\partial}{\partial \varepsilon} T\left(P + \varepsilon(\tilde{P} - P)\right)\Big|_{\varepsilon=1} \\
&= -\frac{\partial}{\partial \tilde{\varepsilon}} T\left(\tilde{P} + \tilde{\varepsilon}(P - \tilde{P})\right)\Big|_{\tilde{\varepsilon}=0} \\
&= -\int IF(z, \tilde{P})\left(p(z) - \tilde{p}(z)\right) \mathrm{d}z \\
&= -\int IF(z, \tilde{P}) p(z) \mathrm{d}z \\
&\approx -\frac{1}{n} \sum_{i=1}^{n} IF(z_i, \tilde{P})
\end{aligned} \tag{4.14}$$

where the second equality follows from rearranging P and $\tilde{P}$ in the expression for P_ε and the fourth equality follows from Equation (4.12). Thus, using this approximation in combination with Equation (4.9) yields

$$T(P) \approx T(\tilde{P}) + \frac{1}{n} \sum_{i=1}^{n} IF(z_i, \tilde{P}) - R_2. \tag{4.15}$$

Therefore, the one step estimator is

$$\hat{T}_1 := T(\tilde{P}) + \frac{1}{n} \sum_{i=1}^{n} IF(z_i, \tilde{P}). \tag{4.16}$$

Until now we have considered the case in which no prior knowledge or restrictions are assumed about P. However, we might consider semi parametric models when applying the (CV-)TMLE method and thus, we now expand the concept of the influence function. In particular, when some parameters of P are known, then some distributions along $\mathcal{P}$ may not fulfil these requirements. Hence, we encode these restrictions in a likelihood model $\mathcal{L}(z, e)$ for $e \in [0, 1]$ and with distribution W_e and density w_e. This implies that we no longer need to define the influence function for any G and Q but instead we the define the influence function in terms of the score of this likelihood shown in the following definition.

Definition 4.2.2. Score-based Influence Function
Let $L_0^2(P)$ be the subspace of the Hilbert space $L^2(P)$ of mean zero function of Z and let T be a functional. The influence function for T is the function $IF \in L_0^2(P)$ which satisfies

$$\frac{\partial}{\partial e} T(W_e)\bigg|_{e=0} = \mathbb{E}_{W_0}[IF(Z, W_0)s_0(Z)] \tag{4.17}$$

and

$$\mathbb{E}_{W_0}[IF(Z, W_0)] = 0 \tag{4.18}$$

for any W_e, where s_e is the score function given by

$$s_e(z) = \frac{\partial}{\partial e} \log\left(w_e(z)\right). \tag{4.19}$$

◀

If a function IF satisfies the conditions in Definition 4.2.2, it also satisfies the conditions in Definition 4.2.1. This holds since if we define $W_e := G + e(Q - G)$ for any two distributions G and Q then the score function for $e = 0$ is given by

$$\begin{aligned} s_0(z) &= \frac{\partial}{\partial e} \log\Big(g(z) + e\left(q(z) - g(z)\right)\Big)\bigg|_{e=0} \\ &= \frac{q(z) - g(z)}{g(z)}. \end{aligned} \tag{4.20}$$

Moreover, notice that $W_0 = G$. Hence, from Definition 4.2.2, we get

$$\begin{aligned} \frac{\partial}{\partial e} T(W_e)\bigg|_{e=0} &= \mathbb{E}_{W_0}[IF(Z, W_0)s_0(Z)] \\ &= \int IF(Z, G)s_0(Z)g(z)\mathrm{d}z \\ &= \int IF(Z, G)\left(\frac{q(z) - g(z)}{g(z)}\right)g(z)\mathrm{d}z \\ &= \int IF(Z, G)\left(q(z) - g(z)\right)\mathrm{d}z. \end{aligned} \tag{4.21}$$

Thus, by Equation (4.21), the influence function also satisfies Definition 4.2.1.

Notice that, the score-based influence function in Definition 4.2.2 does not need to be defined for any G and Q as opposed to the influence function in Definition 4.2.1. Thus, the conditions will be fulfilled by a set $\mathcal{S}$ of score-based influence functions. Therefore, we define the *efficient influence function* D^* as

$$D^*(z, P) := \underset{\widetilde{IF} \in \mathcal{S}}{\arg\min}\, \mathbb{V}\mathrm{ar}\left(\widetilde{IF}(Z, P)\right). \tag{4.22}$$

Thus, the efficient influence function is the score-based influence function among the possible score-based influence functions which fulfil the requirements in Definition 4.2.2 with the smallest variance. Therefore, we can estimate the derivative along allowed P_ε of $\mathcal{P}$ more efficiently.

Having presented influence function as well as score-based influence function and the efficient influence function, we in the following section show asymptotic linearity of the one step CV-TML estimator presented in Section 4.1. Notice that, we refer to both influence functions and score-based influence functions as influence functions in the following section.

4.3 Asymptotic Linearity and Efficiency

This section is based on [Zheng and van der Laan, 2010, pp. 2–11, 43], [van der Laan and Rose, 2011, p. 137, Chapter 27], [Herbrich, 2002, pp. 220–221] and [van der Vaart and Wellner, 1996, pp. 80–84]. In this section, we show that the one step CV-TML estimator is an asymptotic linear and efficient estimator. Moreover, based on the asymptotic distribution of the one step CV-TML estimator obtained in this section, we also determine 95% confidence intervals.

Specifically, we show that

$$\hat{\Psi}(P_n) - \psi_0 = (P_n - P_0)IF(P_0) + R_n \tag{4.23}$$

where ψ_0 denotes the target parameter, that is, $\psi_0 = \Psi(P_0)$ and R_n denotes the remainder term. Notice that we again use the notation introduced in Section 4.1, that is,

$$PIF(P) = \int IF(o, P)\mathrm{d}P(o). \tag{4.24}$$

for a distribution P. For the remaining of this section, we denote $IF(o, P) = IF(P)$ for all influence functions including the efficient influence function. Furthermore, notice that we now consider the random vector $O \sim P_0$ compared to $Z \sim P$ in Section 4.2 and thus also $L_0^2(P)$ now denotes the subspace of the Hilbert space $L^2(P)$ of mean zero function of O.

In order for the influence function to be well-defined, we assume that $\Psi \colon \mathcal{M} \to \mathbb{R}^d$ is a pathwise differentiable function at each $P \in \mathcal{M}$. Moreover, assume that $Q \colon \mathcal{M} \to \mathcal{Q}$ is chosen such that for some $\Psi^1 \colon \mathcal{Q} \to \mathbb{R}^d$ then

$$\Psi(P_0) = \Psi^1\left(Q(P_0)\right). \tag{4.25}$$

Notice that, we abuse notation and refer to both mappings Ψ and Ψ^1 as Ψ and thus write $\Psi\left(Q(P)\right)$ and $\Psi(P)$ interchangeably. Furthermore, for all $P \in \mathcal{M}$, let

$$D^*(P) = D^*\left(Q(P), g(P)\right), \tag{4.26}$$

that is, D^* depends on P through the relevant part Q of P and a nuisance parameter $g(P)$ for $g\colon \mathcal{M} \to \mathcal{G}$. Moreover, the efficient influence function is double robust, that is,

$$P_0 D^*(Q, g) = 0 \quad \text{if } Q = Q_0 \text{ or } g = g_0. \tag{4.27}$$

However, the proof of this property is out of scope for this master's thesis.

In addition, let $\mathcal{L}^\infty(K)$ be the class of function of O with bounded supremum norm over a set K such that $P_0(O \in K) = 1$. Then assume that there exists a loss function $L\colon \mathcal{Q} \to \mathcal{L}^\infty(K)$ which is uniformly bounded and for which it holds that

$$Q(P_0) \coloneqq \arg\min_{Q \in \mathcal{Q}} \left(P_0 L(Q)\right). \tag{4.28}$$

Moreover, assume for each $P \in \mathcal{M}$, for a parametric model $\{P(\varepsilon) : \varepsilon\} \subset \mathcal{M}$, that the corresponding $D^*(P)$ fulfils

$$\langle D^*(P) \rangle \subset \left\langle \frac{\partial}{\partial \varepsilon} L\left(Q\left(P(\varepsilon)\right)\right)\Big|_{\varepsilon=0} \right\rangle \tag{4.29}$$

where $\langle h \rangle$ denotes the set of linear combinations of the components of $h = (h_1, \ldots, h_d)$. Furthermore, given the initial estimators $\hat{Q}$ of $Q_0 = Q(P_0)$ and $\hat{g}$ of $g_0 = g(P_0)$ where we let $\{\hat{Q}(P_n)(\varepsilon) : \varepsilon\} \subset \mathcal{M}$ then we let $P_n \mapsto \hat{Q}(P_n)(\varepsilon)$ such that it fulfils

$$\langle D^*(\hat{Q}(P_n), \hat{g}(P_n)) \rangle \subset \left\langle \frac{\partial}{\partial \varepsilon} L\left(\hat{Q}(P_n)(\varepsilon)\right)\Big|_{\varepsilon=0} \right\rangle. \tag{4.30}$$

Now we present some definitions and lemmata which we use to show asymptotic linearity of the one step CV-TML estimator.

Definition 4.3.1. **ε-cover and Covering Number**
Let (X, d) be a normed space. Moreover, let $A \subseteq X$ and $\varepsilon > 0$. Then $B \subseteq X$ is an *ε-cover* of A if

$$\forall a \in A \; \exists b \in B : d(a, b) \leqslant \varepsilon. \tag{4.31}$$

Or equivalently,

$$A \subseteq \bigcup_{b \in B} \bar{B}_\varepsilon(b). \tag{4.32}$$

Let $\mathcal{B}$ denote the set of all ε-covers of A. The *covering number* $N(\varepsilon, A, d)$ is the minimal cardinality of an ε-cover of A, that is,

$$N(\varepsilon, A, d) = \min_{B \in \mathcal{B}} \{n : |B| = n\}. \tag{4.33}$$

◀

We note that the center of the balls do not need to belong to A and the balls do not need to be disjoint. In Figure 4.3, we illustrate an example on an ε-cover of a set A.

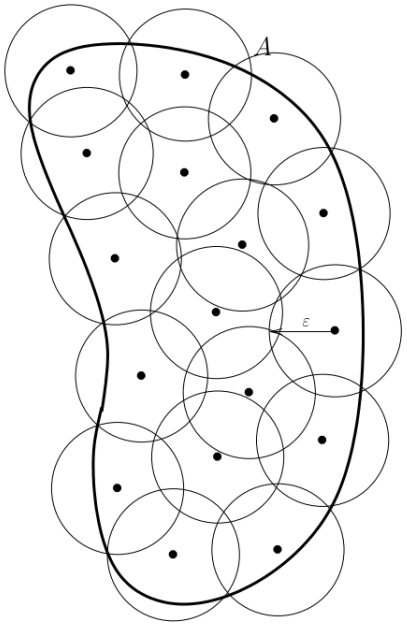


Figure 4.3: An ε-cover of a set A [Herbrich, 2002, p. 221] .

We now consider the covering number for a specific class of functions $\mathcal{F} := \{f \mid f \colon O \to \mathbb{R}\}$ where we use the $L_2(Q)$ norm given by

$$\|f\|_{Q,2} = \left(\int |f|^2 \mathrm{d}Q \right)^{\frac{1}{2}} \tag{4.34}$$

where Q is a probability measure. Moreover, we consider the *uniform number* given by

$$\log \left(\sup_Q \left(N \left(\varepsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q) \right) \right) \right) \tag{4.35}$$

where $\varepsilon > 0$ and $0 < QF^2 < \infty$ for an *envelope function* F of $\mathcal{F}$ which is a function that fulfils that $|f(o)| \leqslant F(o)$ for all $o \in O$ and for all $f \in \mathcal{F}$.

Definition 4.3.2. Entropy Integral
Let $\mathcal{F} = \{f \mid f \colon O \to \mathbb{R}\}$ and let F be an envelope function of $\mathcal{F}$. The *entropy integral* is defined as

$$\text{Entro}(\mathcal{F}) = \int_0^\infty \sqrt{\log\left(\sup_Q \left(N\left(\varepsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q)\right)\right)\right)} \mathrm{d}\varepsilon. \tag{4.36}$$

◀

We now present a lemma which we use to prove Lemma 4.3.4.

Lemma 4.3.3.
Let $\mathcal{F} = \{f \mid f \colon O \to \mathbb{R}\}$ and F be an envelope function of $\mathcal{F}$. Furthermore, let $G_n = \sqrt{n}(P_n - P_0)$. Then it holds that

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} |G_n f|\right] \leqslant \text{Entro}(\mathcal{F}) \sqrt{P_0 F^2}. \tag{4.37}$$

◀

The proof of Lemma 4.3.3 is beyond the scope of this master's thesis but can be found in [van der Vaart and Wellner, 1996, Section 2.14.1].

In order to prove asymptotic linearity of the one step CV-TML estimator, we use Lemma 4.3.4 which we now present.

Lemma 4.3.4.
Let $\varepsilon_0 \coloneqq \arg\min_\varepsilon \left(P_0 L\left(\hat{Q}(P_0)(\varepsilon)\right)\right)$ and let ε_n be defined as in Section 4.1. Furthermore, suppose that $||\varepsilon_n - \varepsilon_0|| \xrightarrow{\mathbb{P}} 0$. For each B_n, we condition on P^0_{n,B_n} and consider the class of measurable functions of O given by

$$\mathcal{F}(P^0_{n,B_n}) = \{f_\varepsilon(P^0_{n,B_n}) = f(\varepsilon, P^0_{n,B_n}) - f(\varepsilon_0, P_0) : \varepsilon\}. \tag{4.38}$$

For a deterministic sequence $\{\delta_n\}_{n \geqslant 1}$ where $\delta_n \to 0$ for $n \to \infty$, we define

$$\mathcal{F}_{\delta_n}(P^0_{n,B_n}) = \{f_\varepsilon \in \mathcal{F}(P^0_{n,B_n}) : ||\varepsilon - \varepsilon_0|| < \delta_n\}. \tag{4.39}$$

If it holds that

$$\mathbb{E}\left[\text{Entro}\left(\mathcal{F}_{\delta_n}(P^0_{n,B_n})\right) \sqrt{P_0 F_{\delta_n}(P^0_{n,B_n})^2}\right] \to 0 \text{ for } n \to \infty, \tag{4.40}$$

where $F_{\delta_n}(P^0_{n,B_n})$ is the envelope function of $\mathcal{F}_{\delta_n}(P^0_{n,B_n})$, then

$$\sqrt{n}(P^1_{n,B_n} - P_0)\left(f(\varepsilon_n, P^0_{n,B_n}) - f(\varepsilon_0, P_0)\right) = o_p(1). \tag{4.41}$$

◀

Proof. Let $G^1_{n,B_n} = \sqrt{n}(P^1_{n,B_n} - P_0)$. We want to show that $G^1_{n,B_n} f_{\varepsilon_n}(P^0_{n,B_n}) = o_p(1)$ for $f_{\varepsilon_n}(P^0_{n,B_n}) \in \mathcal{F}(P^0_{n,B_n})$ which holds if $G^1_{n,B_n} f_{\varepsilon_n}(P^0_{n,B_n}) \xrightarrow{\mathbb{P}} 0$ or equivalently

$$\mathbb{P}\left(|G^1_{n,B_n} f_{\varepsilon_n}(P^0_{n,B_n})| > \delta\right) \to 0 \text{ for } n \to \infty \tag{4.42}$$

for all $\delta > 0$.

We have that

$$\begin{aligned}
&\mathbb{P}\left(|G^1_{n,B_n} f_{\varepsilon_n}(P^0_{n,B_n})| > \delta\right) \\
&\quad = \mathbb{E}\left[\mathbb{P}\left(|G^1_{n,B_n} f_{\varepsilon_n}(P^0_{n,B_n})| > \delta \mid P^0_{n,B_n}\right)\right] \\
&\quad = \mathbb{E}\left[\mathbb{P}\left(|G^1_{n,B_n} f_{\varepsilon_n}(P^0_{n,B_n}) \mathbb{1}\left[||\varepsilon_n - \varepsilon_0|| < \delta_n\right]| > \delta \mid P^0_{n,B_n}\right)\right] \\
&\qquad + \mathbb{E}\left[\mathbb{P}\left(|G^1_{n,B_n} f_{\varepsilon_n}(P^0_{n,B_n}) \mathbb{1}[||\varepsilon_n - \varepsilon_0|| \geqslant \delta_n]| > \delta \mid P^0_{n,B_n}\right)\right] \\
&\quad \leqslant \mathbb{E}\left[\mathbb{P}\left(\sup_{f \in \mathcal{F}_{\delta_n}(P^0_{n,B_n})} \left(|G^1_{n,B_n} f|\right) > \delta \mid P^0_{n,B_n}\right)\right] \\
&\qquad + \mathbb{E}\left[\mathbb{P}\left(||\varepsilon_n - \varepsilon_0|| \geqslant \delta_n \mid P^0_{n,B_n}\right)\right] \\
&\quad = \mathbb{E}\left[\mathbb{P}\left(\sup_{f \in \mathcal{F}_{\delta_n}(P^0_{n,B_n})} \left(|G^1_{n,B_n} f|\right) > \delta \mid P^0_{n,B_n}\right)\right] + \mathbb{P}\left(||\varepsilon_n - \varepsilon_0|| \geqslant \delta_n\right)
\end{aligned} \tag{4.43}$$

where we in the first and last equality apply the Law of Total Expectation. The inequality follows since for $\|\varepsilon_n - \varepsilon_0\| < \delta_n$ then $f_{\varepsilon_n}(P^0_{n,B_n}) \in \mathcal{F}_{\delta_n}(P^0_{n,B_n})$ where the supremum of such functions, f, only increases the probability of $\left|G^1_{n,B_n} f\right|$ being greater than δ. In addition, we apply that for two events A and B where $A \subseteq B$ then $\mathbb{P}(A) \leqslant \mathbb{P}(B)$.

We have assumed that $||\varepsilon_n - \varepsilon_0|| \xrightarrow{\mathbb{P}} 0$ and thus, by definition of convergence in probability, we have that $\mathbb{P}(||\varepsilon_n - \varepsilon_0|| \geqslant \delta_n) \to 0$ for $n \to \infty$. Thus, we consider

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{P}\left(\sup_{f\in\mathcal{F}_{\delta_n}(P^0_{n,B_n})}\left(|G^1_{n,B_n}f|\right) > \delta \mid P^0_{n,B_n}\right)\right] \\
&\quad\leqslant \frac{1}{\delta}\,\mathbb{E}\left[\mathbb{E}\left[\sup_{f\in\mathcal{F}_{\delta_n}(P^0_{n,B_n})}\left(|G^1_{n,B_n}f|\right) \mid P^0_{n,B_n}\right]\right] \\
&\quad\leqslant \frac{1}{\delta}\,\mathbb{E}\left[\text{Entro}\left(\mathcal{F}_{\delta_n}(P^0_{n,B_n})\right)\sqrt{P_0F_{\delta_n}(P^0_{n,B_n})^2}\right]
\end{aligned} \tag{4.44}$$

where we in the first inequality apply Markov's inequality and in the second inequality apply Lemma 4.3.3. By the assumption in Equation (4.40), we obtain that

$$\frac{1}{\delta}\,\mathbb{E}\left[\text{Entro}\left(\mathcal{F}_{\delta_n}(P^0_{n,B_n})\right)\sqrt{P_0F_{\delta_n}(P^0_{n,B_n})^2}\right] \to 0 \text{ for } n \to \infty. \tag{4.45}$$

Thus, we have shown the convergence in probability in Equation (4.42). ■

In the following theorem, we show the asymptotic linearity of the one step CV-TML estimator.

Theorem 4.3.5. Asymptotic Linearity for the One Step CV-TML Estimator
Let $\hat{Q}(P_n)$ and $\hat{g}(P_n)$ be initial estimators of Q_0 and g_0, respectively, and let $\hat{Q}(P_0)$ and $\hat{g}(P_0)$ denote their respective limits. Moreover, suppose that B_n is uniformly distributed over a finite support. Furthermore, consider

$$\hat{\Psi}(P_n) = \mathbb{E}_{B_n}\left[\Psi\left(\hat{Q}(P^0_{n,B_n})(\varepsilon_n)\right)\right]. \tag{4.46}$$

If $P \mapsto \Psi\left(Q(P)\right)$ fulfils

A1:

$$\Psi\left(Q(P)\right) - \Psi(Q_0) = -P_0D^*\left(Q(P), g_0\right) + O_p\left(\left\|\Psi\left(Q(P)\right) - \Psi(Q_0)\right\|^2\right) \tag{4.47}$$

then

$$\begin{aligned}
&\hat{\Psi}(P_n) - \psi_0 \\
&\quad = \mathbb{E}_{B_n}\left[\left(P^1_{n,B_n} - P_0\right) D^*\left(\hat{Q}\left(P^0_{n,B_n}\right)(\varepsilon_n), \hat{g}\left(P^0_{n,B_n}\right)\right)\right] \\
&\qquad + \mathbb{E}_{B_n}\left[P_0\left(D^*\left(\hat{Q}\left(P^0_{n,B_n}\right)(\varepsilon_n), \hat{g}\left(P^0_{n,B_n}\right)\right) - D^*\left(\hat{Q}\left(P^0_{n,B_n}\right)(\varepsilon_n), g_0\right)\right)\right] \\
&\qquad - \mathbb{E}_{B_n}\left[P_0\left(D^*\left(Q_0, \hat{g}\left(P^0_{n,B_n}\right)\right) - D^*\left(Q_0, g_0\right)\right)\right] \\
&\qquad + O_p\left(\left\|\hat{\Psi}\left(P_n\right) - \psi_0\right\|^2\right).
\end{aligned} \tag{4.48}$$

Now consider $\varepsilon_0 \coloneqq \arg\min_{\varepsilon}\left(P_0 L\left(\hat{Q}(P_0)(\varepsilon)\right)\right)$ such that $\|\varepsilon_n - \varepsilon_0\| \xrightarrow{\mathbb{P}} 0$. Assume that

A2: For each B_n, condition on P^0_{n,B_n} and consider the class of functions

$$\mathcal{F}(P^0_{n,B_n}) = \left\{O \mapsto D^*\left(\hat{Q}(P^0_{n,B_n})(\varepsilon), \hat{g}(P^0_{n,B_n})\right) - D^*\left(\hat{Q}(P_0)(\varepsilon_0), \hat{g}(P_0)\right) : \varepsilon\right\} \tag{4.49}$$

where we choose the set for which ε varies such that this set contains ε_n with a probability approaching one. Furthermore, for a deterministic sequence $\{\delta_n\}_{n=1}^{\infty}$ where $\delta_n \to 0$ for $n \to \infty$, we define

$$\mathcal{F}_{\delta_n}(P^0_{n,B_n}) = \left\{f_\varepsilon \in \mathcal{F}(P^0_{n,B_n}) : \|\varepsilon - \varepsilon_0\| < \delta_n\right\}. \tag{4.50}$$

Moreover, assume that for $\{\delta_n\}_{n=1}^{\infty}$, it holds that

$$\mathbb{E}\left[\text{Entro}\left(\mathcal{F}_{\delta_n}(P^0_{n,B_n})\right)\sqrt{P_0 F_{\delta_n}(P^0_{n,B_n})^2}\right] \to 0 \text{ for } n \to \infty \tag{4.51}$$

where $F_{\delta_n}(P^0_{n,B_n})$ is the envelope function of $\mathcal{F}_{\delta_n}(P^0_{n,B_n})$. Then

$$\begin{aligned}
\hat{\Psi}(P_n) - \psi_0 = (P_n - P_0)D^*\left(\hat{Q}(P_0)(\varepsilon_0), \hat{g}(P_0)\right) + o_p\left(\frac{1}{\sqrt{n}}\right) \\
+ \mathbb{E}_{B_n}\left[P_0\left(D^*\left(\hat{Q}(P^0_{n,B_n})(\varepsilon), \hat{g}(P^0_{n,B_n})\right) - D^*\left(\hat{Q}(P^0_{n,B_n})(\varepsilon), g_0\right)\right)\right] \\
- \mathbb{E}_{B_n}\left[P_0\left(D^*\left(Q_0, \hat{g}(P^0_{n,B_n})\right) - D^*(Q_0, g_0)\right)\right] \\
+ O_p\left(\left\|\hat{\Psi}(P_n) - \psi_0\right\|^2\right).
\end{aligned} \tag{4.52}$$

In addition, suppose that $\hat{g}(P_n) = g_0$. Then

$$\hat{\Psi}(P_n) - \psi_0 = (P_n - P_0)D^*\left(\hat{Q}(P_0)(\varepsilon_0), g_0\right) + o_p\left(\frac{1}{\sqrt{n}}\right). \tag{4.53}$$

If also $\hat{Q}(P_0)(\varepsilon_0) = Q_0$ then $\hat{\Psi}(P_n)$ is asymptotically efficient, that is,

$$\hat{\Psi}(P_n) - \psi_0 = (P_n - P_0)D^*(Q_0, g_0) + o_p\left(\frac{1}{\sqrt{n}}\right). \tag{4.54}$$

More generally, assume that $\hat{g}(P_0) = g_0$. Let the limit of $\hat{Q}(P_n)(\varepsilon_n)$ be $\tilde{Q}$ and assume that

A3:

$$\begin{aligned}&\mathbb{E}_{B_n}\left[P_0\left(D^*\left(\hat{Q}(P^0_{n,B_n})(\varepsilon_n), \hat{g}(P^0_{n,B_n})\right) - D^*\left(\hat{Q}(P^0_{n,B_n})(\varepsilon_n), g_0\right)\right)\right] \\ &- \mathbb{E}_{B_n}\left[P_0\left(D^*\left(\tilde{Q}, \hat{g}(P^0_{n,B_n})\right) - D^*(\tilde{Q}, g_0)\right)\right] = o_p\left(\frac{1}{\sqrt{n}}\right).\end{aligned} \tag{4.55}$$

A4: For a function $\widetilde{IF}(P_0) \in L^2_0(P_0)$, it holds that

$$\begin{aligned}&\mathbb{E}_{B_n}\left[P_0\left(D^*\left(\tilde{Q}, \hat{g}(P^0_{n,B_n})\right) - D^*\left(\tilde{Q}, g_0\right)\right)\right] \\ &\quad - \mathbb{E}_{B_n}\left[P_0\left(D^*\left(Q_0, \hat{g}(P^0_{n,B_n})\right) - D^*(Q_0, g_0)\right)\right] \\ &= (P_n - P_0)\widetilde{IF}(P_0) + o_p\left(\frac{1}{\sqrt{n}}\right).\end{aligned} \tag{4.56}$$

Then $\hat{\Psi}\left(P_n\right)$ is asymptotically linear, that is,

$$\hat{\Psi}\left(P_n\right) - \psi_0 = (P_n - P_0)\left(D^*\left(\hat{Q}(P_0)(\varepsilon_0), g_0\right) + \widetilde{IF}(P_0)\right) + o_p\left(\frac{1}{\sqrt{n}}\right). \tag{4.57}$$

◀

Proof. Since D^* is double robust, see Equation (4.27), then

$$P_0 D^*(Q_0, g) = 0 \quad \forall g. \tag{4.58}$$

Moreover, by the definition of ε_n and the one-step convergence of $\hat{Q}(P^0_{n,B_n})(\varepsilon_n)$, we get

$$\mathbb{E}_{B_n}\left[P^1_{n,B_n} D^*\left(\hat{Q}(P^0_{n,B_n})(\varepsilon_n), \hat{g}(P^0_{n,B_n})\right)\right] = 0. \tag{4.59}$$

Applying these results in combination with **A1** yields

$$\begin{aligned}
&\hat{\Psi}(P_n) - \psi_0 \\
&\quad = \mathbb{E}_{B_n}\left[\Psi\left(\hat{Q}(P^0_{n,B_n})(\varepsilon_n)\right) - \psi_0\right] && (4.60)\\
&\quad = \mathbb{E}_{B_n}\left[-P_0 D^*\left(\hat{Q}(P^0_{n,B_n})(\varepsilon_n), g_0\right) + O_p\left(\left\|\Psi\left(\hat{Q}(P^0_{n,B_n})(\varepsilon_n)\right) - \psi_0\right\|^2\right)\right] && (4.61)\\
&\quad = \mathbb{E}_{B_n}\left[-P_0 D^*\left(\hat{Q}(P^0_{n,B_n})(\varepsilon_n), g_0\right)\right] - \mathbb{E}_{B_n}\left[P_0 D^*\left(Q_0, \hat{g}(P^0_{n,B_n})\right)\right] \\
&\qquad + \mathbb{E}_{B_n}\left[O_p\left(\left\|\Psi\left(\hat{Q}(P^0_{n,B_n})(\varepsilon_n)\right) - \psi_0\right\|^2\right)\right] && (4.62)\\
&\quad = \mathbb{E}_{B_n}\left[\left(P^1_{n,B_n} - P_0\right) D^*\left(\hat{Q}\left(P^0_{n,B_n}\right)(\varepsilon_n), \hat{g}\left(P^0_{n,B_n}\right)\right)\right] \\
&\qquad + \mathbb{E}_{B_n}\left[P_0\left(D^*\left(\hat{Q}\left(P^0_{n,B_n}\right)(\varepsilon_n), \hat{g}\left(P^0_{n,B_n}\right)\right) - D^*\left(\hat{Q}\left(P^0_{n,B_n}\right)(\varepsilon_n), g_0\right)\right)\right] \\
&\qquad - \mathbb{E}_{B_n}\left[P_0\left(D^*\left(Q_0, \hat{g}\left(P^0_{n,B_n}\right)\right) - D^*\left(Q_0, g_0\right)\right)\right] \\
&\qquad + O_p\left(\left\|\hat{\Psi}\left(P_n\right) - \psi_0\right\|^2\right) && (4.63)
\end{aligned}$$

where we use the definition of the estimator $\hat{\Psi}(P_n)$ in the first equality, the assumption in **A1** in the second equality, Equation (4.58) in the third equality and Equation (4.59) in the fourth equality. This verifies Equation (4.48) of **A1**.

The first term in Equation (4.63) can we rewritten as

$$\begin{aligned}
&\mathbb{E}_{B_n}\left[\left(P^1_{n,B_n}-P_0\right)D^*\left(\hat{Q}\left(P^0_{n,B_n}\right)(\varepsilon_n),\hat{g}\left(P^0_{n,B_n}\right)\right)\right] \\
&\quad=\mathbb{E}_{B_n}\left[\left(P^1_{n,B_n}-P_0\right)\left(D^*\left(\hat{Q}\left(P^0_{n,B_n}\right)(\varepsilon_n),\hat{g}\left(P^0_{n,B_n}\right)\right)-D^*\left(\hat{Q}\left(P_0\right)(\varepsilon_0),\hat{g}\left(P_0\right)\right)\right)\right] \\
&\qquad+\mathbb{E}_{B_n}\left[\left(P^1_{n,B_n}-P_0\right)D^*\left(\hat{Q}\left(P_0\right)(\varepsilon_0),\hat{g}\left(P_0\right)\right)\right].
\end{aligned} \tag{4.64}$$

Now notice that the assumptions in **A2** coincide with the assumptions of Lemma 4.3.4 for

$$f(\varepsilon,P^0_{n,B_n})=D^*\left(\hat{Q}(P^0_{n,B_n})(\varepsilon),\hat{g}(P^0_{n,B_n})\right). \tag{4.65}$$

Thus, Lemma 4.3.4 implies that for each B_n

$$\begin{aligned}
&\left(P^1_{n,B_n}-P_0\right)\left(D^*\left(\hat{Q}\left(P^0_{n,B_n}\right)(\varepsilon_n),\hat{g}\left(P^0_{n,B_n}\right)\right)-D^*\left(\hat{Q}\left(P_0\right)(\varepsilon_0),\hat{g}\left(P_0\right)\right)\right) \\
&\quad=o_p\left(\frac{1}{\sqrt{n}}\right).
\end{aligned} \tag{4.66}$$

Moreover, since B_n uniformly distributed over a finite support, then in fact

$$\begin{aligned}
&\mathbb{E}_{B_n}\left[\left(P^1_{n,B_n}-P_0\right)\left(D^*\left(\hat{Q}\left(P^0_{n,B_n}\right)(\varepsilon_n),\hat{g}\left(P^0_{n,B_n}\right)\right)-D^*\left(\hat{Q}\left(P_0\right)(\varepsilon_0),\hat{g}\left(P_0\right)\right)\right)\right] \\
&\quad=o_p\left(\frac{1}{\sqrt{n}}\right).
\end{aligned} \tag{4.67}$$

Using this result in combination with Equation (4.64) yields

$$\begin{aligned}
&\mathbb{E}_{B_n}\left[\left(P^1_{n,B_n}-P_0\right)D^*\left(\hat{Q}\left(P^0_{n,B_n}\right)(\varepsilon_n),\hat{g}\left(P^0_{n,B_n}\right)\right)\right] \\
&\quad=\mathbb{E}_{B_n}\left[\left(P^1_{n,B_n}-P_0\right)D^*\left(\hat{Q}\left(P_0\right)(\varepsilon_0),\hat{g}\left(P_0\right)\right)\right]+o_p\left(\frac{1}{\sqrt{n}}\right).
\end{aligned} \tag{4.68}$$

Thus, Equation (4.63) can be written as

$$\begin{aligned}
\hat{\Psi}(P_n) - \psi_0 = \mathbb{E}_{B_n}&\left[\left(P^1_{n,B_n} - P_0\right) D^*\left(\hat{Q}\left(P_0\right)\left(\varepsilon_0\right), \hat{g}\left(P_0\right)\right)\right] + o_p\left(\frac{1}{\sqrt{n}}\right) \\
&+ \mathbb{E}_{B_n}\left[P_0\left(D^*\left(\hat{Q}\left(P^0_{n,B_n}\right)\left(\varepsilon_n\right), \hat{g}\left(P^0_{n,B_n}\right)\right) - D^*\left(\hat{Q}\left(P^0_{n,B_n}\right)\left(\varepsilon_n\right), g_0\right)\right)\right] \\
&- \mathbb{E}_{B_n}\left[P_0\left(D^*\left(Q_0, \hat{g}\left(P^0_{n,B_n}\right)\right) - D^*\left(Q_0, g_0\right)\right)\right] \\
&+ O_p\left(\left\|\hat{\Psi}\left(P_n\right) - \psi_0\right\|^2\right) \qquad (4.69) \\
= \left(P_n - P_0\right) D^*&\left(\hat{Q}\left(P_0\right)\left(\varepsilon_0\right), \hat{g}\left(P_0\right)\right) + o_p\left(\frac{1}{\sqrt{n}}\right) \\
&+ \mathbb{E}_{B_n}\left[P_0\left(D^*\left(\hat{Q}\left(P^0_{n,B_n}\right)\left(\varepsilon_n\right), \hat{g}\left(P^0_{n,B_n}\right)\right) - D^*\left(\hat{Q}\left(P^0_{n,B_n}\right)\left(\varepsilon_n\right), g_0\right)\right)\right] \\
&- \mathbb{E}_{B_n}\left[P_0\left(D^*\left(Q_0, \hat{g}\left(P^0_{n,B_n}\right)\right) - D^*\left(Q_0, g_0\right)\right)\right] \\
&+ O_p\left(\left\|\hat{\Psi}\left(P_n\right) - \psi_0\right\|^2\right). \qquad (4.70)
\end{aligned}$$

This verifies Equation (4.52) of **A2**.

Moreover, if $\hat{g}(P_n) = g_0$, that is, the initial estimator of g_0 is exactly g_0, then Equation (4.70) reduces to

$$\begin{aligned}
\hat{\Psi}\left(P_n\right) - \psi_0 = \left(P_n - P_0\right) D^*&\left(\hat{Q}\left(P_0\right)\left(\varepsilon_0\right), g_0\right) + o_p\left(\frac{1}{\sqrt{n}}\right) \\
&+ O_p\left(\left\|\hat{\Psi}\left(P_n\right) - \psi_0\right\|^2\right)
\end{aligned} \qquad (4.71)$$

since if the initial estimator is g_0, that is, the initial estimator is correct, the limit of this estimator would also be g_0 which corresponds to $\hat{g}(P_0) = g_0$ which is used in the first term in Equation (4.70). Furthermore, the third and fourth terms in Equation (4.70) equal zero since for any Q then

$$\mathbb{E}_{B_n}\left[P_0 D^*\left(Q, \hat{g}(P^0_{n,B_n})\right)\right] = P_0 D^*(Q, g_0) \qquad (4.72)$$

when $\hat{g}(P_n) = g_0$. Notice that for the third term in Equation (4.70), $\hat{Q}$ also depend on P^0_{n,B_n} but since both the $\hat{Q}$'s in this term depend P^0_{n,B_n} in the same way, the terms cancel.

Furthermore, taking the norm on both sides of Equation (4.71) yields

$$\left\| \hat{\Psi}\left(P_n\right) - \psi_0 \right\| = o_p\left(\frac{1}{\sqrt{n}}\right) \tag{4.73}$$

where we refer to [Zheng and van der Laan, 2010, p. 10] for this result.

This yields asymptotic linearity of $\hat{\Psi}(P_n)$ since

$$\hat{\Psi}\left(P_n\right) - \psi_0 = \left(P_n - P_0\right) D^*\left(\hat{Q}\left(P_0\right)\left(\varepsilon_0\right), g_0\right) + o_p\left(\frac{1}{\sqrt{n}}\right). \tag{4.74}$$

Moreover, if $\hat{Q}(P_0)(\varepsilon_0) = Q_0$, then we obtain the efficient influence function $D^*(Q_0, g_0)$.

Now consider a more general case, that is, $\hat{g}(P_0) = g_0$. Furthermore, let $\tilde{Q}$ be the limit of $\hat{Q}(P_n)(\varepsilon_n)$. Then Equation (4.70) can be written as

$$\begin{aligned}
&\hat{\Psi}\left(P_n\right) - \psi_0 \\
&\quad = \left(P_n - P_0\right) D^*\left(\hat{Q}\left(P_0\right)\left(\varepsilon_0\right), g_0\right) + o_p\left(\frac{1}{\sqrt{n}}\right) \\
&\qquad + \mathbb{E}_{B_n}\left[P_0\left(D^*\left(\hat{Q}\left(P^0_{n,B_n}\right)\left(\varepsilon_n\right), \hat{g}\left(P^0_{n,B_n}\right)\right) - D^*\left(\hat{Q}\left(P^0_{n,B_n}\right)\left(\varepsilon_n\right), g_0\right)\right)\right] \\
&\qquad - \mathbb{E}_{B_n}\left[P_0\left(D^*\left(\tilde{Q}, \hat{g}\left(P^0_{n,B_n}\right)\right) - D^*\left(\tilde{Q}, g_0\right)\right)\right] \\
&\qquad + \mathbb{E}_{B_n}\left[P_0\left(D^*\left(\tilde{Q}, \hat{g}\left(P^0_{n,B_n}\right)\right) - D^*\left(\tilde{Q}, g_0\right)\right)\right] \\
&\qquad - \mathbb{E}_{B_n}\left[P_0\left(D^*\left(Q_0, \hat{g}\left(P^0_{n,B_n}\right)\right) - D^*\left(Q_0, g_0\right)\right)\right] \\
&\qquad + O_p\left(\left\| \hat{\Psi}\left(P_n\right) - \psi_0 \right\|^2\right).
\end{aligned} \tag{4.75}$$

Then consider the assumption in **A3**, that is,

$$\begin{aligned}
&\mathbb{E}_{B_n}\left[P_0\left(D^*\left(\hat{Q}\left(P^0_{n,B_n}\right)\left(\varepsilon_n\right), \hat{g}\left(P^0_{n,B_n}\right)\right) - D^*\left(\hat{Q}\left(P^0_{n,B_n}\right)\left(\varepsilon_n\right), g_0\right)\right)\right] \\
&- \mathbb{E}_{B_n}\left[P_0\left(D^*\left(\tilde{Q}, \hat{g}\left(P^0_{n,B_n}\right)\right) - D^*\left(\tilde{Q}, g_0\right)\right)\right] = o_p\left(\frac{1}{\sqrt{n}}\right)
\end{aligned} \tag{4.76}$$

and the assumption in **A4**, that is, for $\widetilde{IF}(P_0) \in L_0^2(P_0)$ then

$$\begin{aligned}
&\mathbb{E}_{B_n}\left[P_0\left(D^*\left(\tilde{Q},\hat{g}\left(P^0_{n,B_n}\right)\right)-D^*\left(\tilde{Q},g_0\right)\right)\right] \\
&\quad-\mathbb{E}_{B_n}\left[P_0\left(D^*\left(Q_0,\hat{g}\left(P^0_{n,B_n}\right)\right)-D^*\left(Q_0,g_0\right)\right)\right] \\
&\quad=\left(P_n-P_0\right)\widetilde{IF}\left(P_0\right)+o_p\left(\frac{1}{\sqrt{n}}\right).
\end{aligned} \tag{4.77}$$

Hence, applying Equations (4.76) and (4.77) to Equation (4.75) yields

$$\begin{aligned}
\hat{\Psi}\left(P_n\right)-\psi_0 =& \left(P_n-P_0\right)\left(D^*\left(\hat{Q}\left(P_0\right)\left(\varepsilon_0\right),g_0\right)+\widetilde{IF}\left(P_0\right)\right)+o_p\left(\frac{1}{\sqrt{n}}\right) \\
&+O_p\left(\left\|\hat{\Psi}\left(P_n\right)-\psi_0\right\|^2\right).
\end{aligned} \tag{4.78}$$

By taking the norm on both sides, we get

$$\left\|\hat{\Psi}\left(P_n\right)-\psi_0\right\| = o_p\left(\frac{1}{\sqrt{n}}\right). \tag{4.79}$$

Thus, we obtain the desired result, that is,

$$\hat{\Psi}\left(P_n\right)-\psi_0 = \left(P_n-P_0\right)\left(D^*\left(\hat{Q}\left(P_0\right)\left(\varepsilon_0\right),g_0\right)+\widetilde{IF}\left(P_0\right)\right)+o_p\left(\frac{1}{\sqrt{n}}\right). \tag{4.80}$$

Hence, we have shown asymptotic linearity of the one step CV-TML estimator. ■

Notice that Equation (4.57) can be rewritten as

$$\sqrt{n}\left(\hat{\Psi}\left(P_n\right)-\psi_0\right) = \sqrt{n}\left(P_n-P_0\right)\left(D^*\left(\hat{Q}\left(P_0\right)\left(\varepsilon_0\right),g_0\right)+\widetilde{IF}\left(P_0\right)\right)+o_p(1). \tag{4.81}$$

Analogously, we can derive similar expressions for Equations (4.53) and (4.54).

Furthermore, $o_P(1)$ converges to zero in probability as the sample size goes to infinity. Moreover, recall that the influence function and thus also the efficient influence function has mean zero and finite variance. Then since the mean and variance of $\hat{\Psi}\left(P_n\right)-\psi_0$ coincide with the mean and variance of the influence function, then by the Central Limit Theorem

$$\sqrt{n}\left(\hat{\Psi}(P_n)-\psi_0\right) \xrightarrow{d} N\left(0,\mathbb{V}\mathrm{ar}[\hat{\Psi}(P_n)]\right). \tag{4.82}$$

Thus, we can derive 95% confidence intervals by using the asymptotic normal distribution of $\hat{\Psi}(P_n)$ where we note that the variance of $\hat{\Psi}(P_n)$ is well approximated by the variance of the influence function.

5 Causal Inference for Longitudinal Studies

In this chapter, we extend theory of causal inference presented in Chapter 2 to longitudinal studies in order to identify (average) causal effects in such studies. In particular, we consider causal inference for a time-varying treatment and how (conditionally) randomised experiments and the identifiability conditions generalise to observational longitudinal studies. Furthermore, we outline how the TMLE method extends to longitudinal studies. Moreover, we assume that there is no random variation in Sections 5.1 and 5.2.

5.1 Treatment Strategies and Causal Effects

This section is based on [Miguel A. Hernán, 2020, pp. 235–237] and [Robins and Hernán, 2009, pp. 560–561]. In previous chapters, we have considered the case of identifying and estimating average causal effects of a treatment on an outcome where the treatment was only measured once. We refer to such a treatment as a *fixed treatment*. In many cases, including the data set considered in Chapter 6, the treatment varies over time which we refer to as a *time-varying treatment*. Thus, we need a more general definition of (average) causal effects that incorporate time explicitly.

Consider a binary time-varying treatment A_k for $k = 0, 1, \ldots, K$ where K denotes the number of times the treatment has been measured in a longitudinal study. Hence, we can denote *the treatment history* from time zero to time k for $k \in \{0, 1, \ldots, K-1\}$ as $\bar{A}_k = (A_0, A_1, \ldots, A_k)$ while we denote the entire treatment history as $\bar{A}$. Since we only have defined the (average) causal effect for a fixed treatment, we can based on Definitions 2.1.2 and 2.1.3 only identify the (average) causal effect at a single time $k \in \{0, 1, \ldots, K\}$ and thus not the (average) causal effect of the time-varying treatment over the entire period of the longitudinal study. Hence, we first consider various *treatment strategies* which are presented in the following.

Treatment strategies are rules to assign the treatment at each time $k \in \{0, 1, \ldots, K\}$. An example of such a treatment strategy could be assigning the treatment at every $k \in \{0, 1, \ldots, K\}$, that is, $\bar{a}_1 = (1, 1, \ldots, 1)$. Another treatment strategy could be $\bar{a}_0 = (0, 0, \ldots, 0)$, that is, an individual with this treatment strategy never receives the treatment. Thus, we now have two treatment strategies and hence, we can identify a non-zero causal effect of the time-varying treatment on the outcome for an individual i if $Y_i(\bar{A} = \bar{a}_1) \neq Y_i(\bar{A} = \bar{a}_0)$.

However, there are many other possible treatment strategies for a time-varying treatment. In particular, we have at least 2^{K+1} treatment strategies for a binary treatment A_k for $k = 0, 1, \ldots, K$. Moreover, further treatment strategies can be defined when considering a treatment which depends on covariates, that is, a dynamic treatment rule as defined in Section 3.1. Hence, we now define a causal effect of multiple treatment strategies for an individual in the following definition.

Definition 5.1.1. Causal Effect of Multiple Treatment Strategies for an Individual
Consider the treatment strategies $\bar{a}_1, \ldots, \bar{a}_m$ for $m \geqslant 2$. If for an individual i, $Y_i(\bar{a}_l) \neq Y_i(\bar{a}_j)$ for at least one pair (l, j) such that $l \neq j$ and $l, j \in \{1, \ldots, m\}$ then $\bar{A}$ has a non-zero causal effect on Y for individual i. ◀

However, as explained in Chapter 2, we consider average causal effects due to the fundamental issue of causal inference. Thus, in the following, we define an average causal effect of multiple treatment strategies of a population.

Definition 5.1.2. Average Causal Effect of Multiple Treatment Strategies of a Population
Consider the treatment strategies $\bar{a}_1, \ldots, \bar{a}_m$ for $m \geqslant 2$. If for at least one pair (l, j) such that $l \neq j$ and $l, j \in \{1, \ldots, m\}$ it holds that

$$\mathbb{E}[Y(\bar{a}_l)] - \mathbb{E}[Y(\bar{a}_j)] \neq 0 \tag{5.1}$$

then $\bar{A}$ has a non-zero average causal effect on Y. ◀

Notice that we only consider the treatment strategies which are relevant for examining the specific causal question.

Hence, when conducting causal inference with a time-varying treatment, we are considering the contrast between the expected values of the potential outcomes under two or more treatment strategies. Thus, the (average) causal effect is only well-defined if the treatment strategies are specified. Therefore, the definition of an (average) causal effect of a time-varying treatment on an outcome is dependent on the treatment strategies. Hence, the (average) causal effect of a time-varying treatment is not uniquely defined.

In the following section, we present cases where the average causal effect of a time-varying treatment can be identified.

5.2 Sequentially Randomised Experiments and the Identifiability Conditions for Observational Longitudinal Studies

This section is based on [Miguel A. Hernán, 2020, pp. 237–241] and [Robins and Hernán, 2009, p. 561]. In Section 2.2, we presented randomised experiments as well as conditionally randomised experiments which were study designs where the average causal effect of a fixed treatment could be identified. Now we generalise these study designs to longitudinal studies in order to identify average causal effects of time-varying treatments.

Consider a longitudinal study and let L_k denote the measured time-varying covariates and U_k denote the unmeasured time-varying covariates for $k = 0, 1, \ldots, K$ where K denotes the number of times the time-varying variables have been measured in the specific study. A *sequentially randomised experiment* is an experiment where the treatment is randomly assigned at each time

k for $k = 0, 1, \ldots, K$ to each individual in the study with known randomisation probabilities that may depend on $\bar{A}_{k-1}$ and $\bar{L}_k$ for $k > 0$ and on L_0 for $k = 0$. Notice that, $\bar{L}_k = (L_0, L_1, \ldots, L_k)$ which we refer to as the *history of the measured time-varying covariates*. On the following figures, we present examples of sequentially randomised experiments where $K = 1$ using causal DAGs.

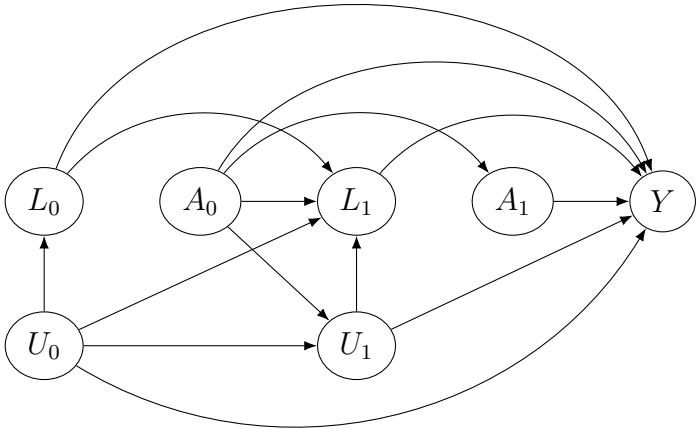


Figure 5.1: Causal DAG of a sequentially randomised experiment where the treatment is assigned at random at each time $k = 0, 1$.

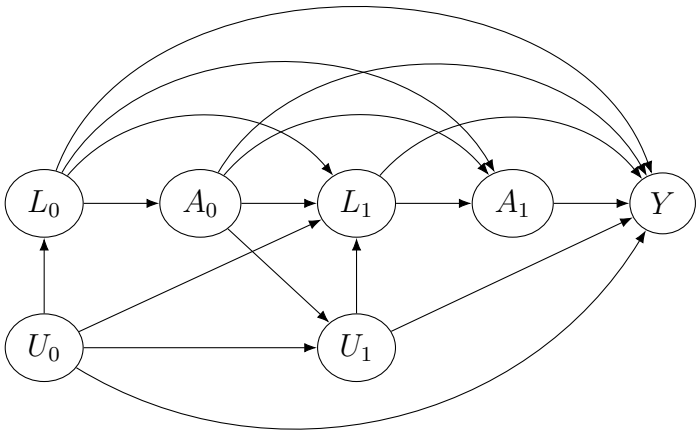


Figure 5.2: Causal DAG sequentially randomised experiment where the treatment is randomly assigned conditioned on the measured time-varying covariates L_k at each time $k = 0, 1$.

Figure 5.1 shows the causal DAG for a sequentially randomised experiment where the treatment is assigned at random at each time $k = 0, 1$ since the treatment is not affected by any covariates at any time. This example corresponds to the generalisation of randomised experiments for a fixed treatment. On the other hand, Figure 5.2 represents an example of a sequentially randomised experiment where the treatment is randomly assigned conditioned on the measured time-varying covariates L_k at each time $k = 0, 1$. Thus, Figure 5.2 corresponds to the generalisation of conditionally randomised experiments for a fixed treatment.

However, as for fixed treatments, we often want to identify average causal effects in observational longitudinal studies. In order to do so, we consider such a study as a sequentially randomised experiment by requiring sufficient identifiability conditions to hold. Thus, we generalise the identifiability conditions to observational longitudinal studies as follows:

(*i*) *Consistency*:

$$\begin{aligned} Y(\bar{a}) &= Y(\bar{a}^*) && \text{if } \bar{a}^* &&= \bar{a}, \\ Y(\bar{a}) &= Y && \text{if } \bar{A} &&= \bar{a}, \\ \bar{L}_k(\bar{a}) &= \bar{L}_k(\bar{a}^*) && \text{if } \bar{a}^*_{k-1} &&= \bar{a}_{k-1}, \\ \bar{L}_k(\bar{a}) &= \bar{L}_k && \text{if } \bar{A}_{k-1} &&= \bar{a}_{k-1}. \end{aligned} \tag{5.2}$$

(*ii*) *Sequential exchangeability*:

$$Y(\bar{a}) \perp\!\!\!\perp A_k \mid \bar{A}_{k-1} = \bar{a}(\bar{A}_{k-2}, \bar{L}_{k-1}), \bar{L}_k = \bar{l}_k \tag{5.3}$$

for all treatment strategies $\bar{a}$ which may depend on $\bar{A}_{k-2}$ and $\bar{L}_{k-1}$ for $k = 0, 1, \ldots, K$.

(*iii*) *Positivity*:
For $\mathbb{P}(\bar{A}_{k-1} = \bar{a}_{k-1}, \bar{L}_k = \bar{l}_k) > 0$ then

$$\mathbb{P}(A_k = a_k \mid \bar{A}_{k-1} = \bar{a}_{k-1}, \bar{L}_k = \bar{l}_k) > 0 \tag{5.4}$$

for all $\bar{a}_k$ and for all $\bar{l}_k$.

Hence, the consistency conditions are now required for both the outcome and the measured time-varying covariates. Specifically, considering the first equation in Equation (5.2), we assume that for two treatment strategies which coincide, then the potential outcomes of the outcome Y also coincide. Moreover, we assume in the second equation in Equation (5.2) that given a treatment strategy then the actual outcome equals the potential outcome corresponding to the given treatment strategy. Notice that this condition is the time-varying extension of the consistency condition in Equation (2.7). Furthermore, we assume consistency conditions for the measured time-varying covariates. In particular, in the third equation in Equation (5.2), we assume that given two treatment strategies which coincide until time $k-1$ then the potential outcomes of the history of the measured time-varying covariates coincide until time k. Moreover, in the fourth equation in Equation (5.2), the condition is that given a treatment strategy until time $k-1$ then the actual history of the measured time-varying covariates until time k equals the potential outcome corresponding to the given treatment strategy.

For sequential exchangeability first notice that $\bar{a}(\bar{A}_{k-2}, \bar{L}_{k-1})$ denotes a treatment strategy until time k where the k'th value of the treatment depends on $\bar{A}_{k-2}$ and $\bar{L}_{k-1}$. Thus, this notation includes dynamic treatment strategies explicitly since the assignment of treatment depends on the values of the measured time-varying covariates until time $k-1$. Sequential exchangeability hence states that for any treatment strategy $\bar{a}$, the treated and the untreated at each time k are exchangeable conditioned on the history of the measured time-varying covariates $\bar{L}_k$ and any treatment strategy $\bar{a}(\bar{A}_{k-2}, \bar{L}_{k-1})$.

The positivity condition requires that if there is a positive probability of $\bar{A}_{k-1}$ and $\bar{L}_k$ to occur simultaneously, then we assume that conditional on these two events occurring, there is a positive probability of a realisation a_k of A_k occurring for all a_k such that $(a_k, \bar{a}_{k-1})$ is a valid treatment strategy.

Thus, for two treatment strategies $\bar{a}$ and $\bar{a}^*$, we can identify the average causal effect based on similar derivations as for the fixed treatment derived in Equation (2.11).

Having extended the theory of causal inference for fixed treatments to time-varying treatments and outlined the conditions to identify average causal effects in observational longitudinal studies, we now present how the TMLE method extends to longitudinal studies.

5.3 Longitudinal Targeted Maximum Likelihood Estimation

This section is based on [Lendle et al., 2017] and [Schomaker et al., 2019]. In this section, we present the longitudinal TMLE (L-TMLE) method which is an extension of the TMLE method for longitudinal studies.

Consider a longitudinal study. Let L_k, A_k and Y_k denote the measured time-varying covariates, the binary time-varying treatment and the binary time-varying outcome at time k for $k = 0, 1, \ldots, K$, respectively, and let $O = (L_0, A_0, Y_0, \ldots, L_K, A_K, Y_K) \sim P_0$. Moreover, if right censoring is present in the longitudinal study, then let C_k denote right censoring at time k. Notice that the ordering of A_k and C_k depends on the particular longitudinal study which we describe in further detail in Section 6.4. Moreover, let $S_{k,i}$ be a binary variable of whether individual i is alive at time k or not where $S_{k,i} = 1$ refers to the individual being alive. Then we let S_k denotes the vector with entries $S_{k,i}$ for all i. This binary variable is not needed in all cases and depends on the censoring mechanism.

First, we note that the structural equations for longitudinal studies are given as in Equation (3.1) and the SCM is defined as in Definition 3.1.1. Moreover, the target parameter for the average causal effect of two treatment strategies is given as for non longitudinal studies.

We now present the L-TMLE method. In the L-TMLE method we apply the *sequential g-formula* given by

$$\begin{aligned}&\mathbb{E}\left[Y_K(\bar{a})\right]\\&\quad=\mathbb{E}\Bigg[\mathbb{E}\Bigg[\cdots\mathbb{E}\Big[\mathbb{E}\big[Y_K\,|\,\bar{A}_{K-1}=\bar{a}_{K-1},\bar{L}_K\big]\,|\,\bar{A}_{K-2}=\bar{a}_{K-2},\bar{L}_{K-1}\Big]\cdots\,|\,A_0=a_0,\bar{L}_1\Bigg]\,|\,L_0\Bigg]\end{aligned} \tag{5.5}$$

Consider Equation (5.5). The L-TMLE method consists of four steps where the first three steps are iterated for $k = K, \ldots, 1$. The reason why we iterate for $k = K, \ldots, 1$ is to sequentially estimate and update each of the expected values in Equation (5.5) starting from the innermost expected value. Thus, in the first step, we estimate $\mathbb{E}[Y_k \mid \bar{A}_{k-1} = \bar{a}_{k-1}, \bar{L}_k]$ for all individuals which are uncensored and alive until time $k-1$ applying the super learner method. Notice that in the first iteration, that is, $k = K$, then the outcome refers to the measured outcome while for later iterations then the outcome refers to the prediction from step three. In step two, set $\bar{A}_{k-1} = \bar{a}_{k-1}$ based on the rule $\bar{a}_k$ and predict the outcome at k, which we denote as $\tilde{Y}_k^{\bar{a}_k}$, based on step one. In step three, the initial estimator is targeted as for the TMLE method described in Section 3.3. However, the intercept is now the predicted outcome from step two and the clever covariate in this case is given by

$$\begin{aligned}H(\bar{A},\bar{C},\bar{L})_{k-1} = \prod_{s=0}^{k-1} &\frac{\mathbb{1}[\bar{A}_s=\bar{a}_s]}{\mathbb{P}(A_s=a_s \mid \bar{L}_s=\bar{l}_s,\bar{S}_s=1,\bar{C}_{s-1}=1,\bar{A}_{s-1}=\bar{a}_{s-1})}\\ &\cdot\frac{\mathbb{1}[\bar{C}_s=1]}{\mathbb{P}(C_s=1 \mid \bar{A}_s=\bar{a}_s,\bar{L}_s=\bar{l}_s,\bar{S}_s=1,\bar{C}_{s-1}=1)}\\ &\cdot\frac{\mathbb{1}[\bar{S}_s=1]}{\mathbb{P}(S_s=1 \mid \bar{C}_s=1,\bar{A}_s=\bar{a}_s,\bar{L}_s=\bar{l}_s,\bar{S}_{s-1}=1)}.\end{aligned} \tag{5.6}$$

Notice that we assume that all individuals are uncensored and alive prior to the beginning of the study, that is, $C_{-1} = S_{-1} = 1$. Also for $s = 0$, that is at time zero, the denominator of the first factor in Equation (5.6) reduces to $\mathbb{P}(A_0 = a_0 \mid L_0 = l_0, \bar{S}_0 = 1, C_{-1} = 1)$. Moreover to estimate the probabilities in the clever covariate in Equation (5.6), we again apply the super learner method.

Thus, step three yields a targeted update of $\tilde{Y}_k^{\bar{a}_k}$. In step four, we estimate the expression in Equation (5.5) by the empirical mean of the prediction from step three for $k = 1$.

6 | Applying the TMLE Methods in Practice

In this chapter, we analyse a subset of the Framingham Heart Study[1]. This study examines cardiovascular diseases among a population in Framingham (Massachusetts, USA). The particular data set considered in this chapter consists of $4,434$ individuals and 39 variables. A list of the variables can be found in Table 6.1 where also a short description of the variables is given. The study consists of three examinations periods approximately six years apart in the period 1956 to 1968. For each of these examinations, the *Time-varying variables* in Table 6.1 are measured. At the first examination attended also the *Fixed variables* are measured except for `Hdlc` and `Ldlc` which only are measured at the third examination. Moreover, each individual is observed for a total of 24 years for the outcome of the *Variables measured for 24 years* in Table 6.1. Furthermore, the data set is subject to right censoring and thus, when considering the entire data set, we need to adjust for censoring.

The aim of this chapter is to determine if smoking has a causal effect on stroke within 24 years. That is, smoking is viewed as the treatment and stroke within 24 years as the outcome. We note that smoking is defined as whether or not an individual is currently smoking at an examination. In particular, we want to analyse this causal question for both a fixed treatment and for a time-varying treatment. For the fixed treatment, we do not consider the measurements of the time-varying variables at the second and third examinations for each individual and examine the causal question *"does smoking at the time of the first examination have a causal effect on stroke within a period of* 24 *years"*. For the time-varying treatment, we consider all measurements of the variables of the data set. In this particular case, the causal question is *"does smoking at the time of each of the examinations have a causal effect on stroke within a period of* 24 *years compared to not smoking at any of the examinations"*. In order to analyse these causal questions, we assume the identifiability conditions hold for both the observational non-longitudinal study considered for the fixed treatment and for the observational longitudinal study considered for the time-varying treatment. We apply the TMLE and the CV-TMLE methods when examining the causal question of the fixed treatment and the L-TMLE method when examining the causal question of the time-varying treatment. Furthermore, all analysis of this chapter is done in `R` where we used the packages `tidyverse`, `missForest`, `sl3`, `SuperLearner`, `tmle3` and `ltmle`. Specific functionalities mentioned in this chapter are implemented in these packages. All packages are available through The Comprehensive R Archive Network (CRAN) except the `tmle3` and `sl3` packages which are available through GitHub see [Coyle, 2021] and [Coyle et al., 2021], respectively. Furthermore, all our code is available at `https://github.com/AalborgGit/TMLEandDataprep` [Last assessed 02-06-2022].

[1]The data set can be found at: `https://biolincc.nhlbi.nih.gov/studies/framcohort/` [Last assessed 02-06-2022].

Table 6.1: List and description of all variables included in the data set [*Framingham Heart Study Longitudinal Data Documentation*, 2021].

Variable Name	Description
Fixed variables	
`Randid`	ID for the individual
`Sex`	The sex of the individual (0=Men, 1=Women)
`Educ`	Education (1=0–11 years, 2=High School Diploma, GED, 3=Some College, Vocational School and 4=College (BS, BA) degree or more)
`Hdlc`	High Density Lipoprotein Cholesterol, mg/dL
`Ldlc`	Low Density Lipoprotein Cholesterol, mg/dL
Time-varying variables	
`Period`	Examination periods (1=First examination, 2=Second examination, 3=Third examination)
`Time`	Number of days since the first examination
`Cursmoke`	Smoking status (0=Not current smoker, 1=Current smoker)
`Cigpday`	Number of cigarettes smoked per day
`Age`	The age of the individual
`Sysbp`	Systolic Blood Pressure, mmHg
`Diabp`	Diastolic Blood Pressure, mmHg
`Bpmeds`	The use of Anti-hypertensive medication (0=Not currently used, 1=Current Use)
`Totchol`	Serum Total Cholesterol, mg/dL
`BMI`	Body Mass Index (BMI)
`Glucose`	Casual serum glucose, mg/dL
`Diabetes`	Diabetes status (0=Not diabetic, 1=Diabetic)
`Heartrte`	Heart rate, beats/min
`Prevap`	Prevalent Angina Pectoris (0=No disease, 1=Prevalent disease)
`Prevchd`	Prevalent Coronary Heart Disease (0=No disease, 1=Prevalent disease)
`Prevmi`	Prevalent Myocardial Infarction (0=No disease, 1=Prevalent disease)
`Prevstrk`	Prevalent Stroke (0=No disease, 1=Prevalent disease)
`Prevhyp`	Prevalent Hypertensive (0=No disease, 1=Prevalent disease)

Variables measured for 24 *years*

`Angina`	Angina Pectoris (0=The event did not occur during follow-up, 1=The event did occur during follow-up)
`Hospmi`	Hospitalized Myocardial Infarction (0=The event did not occur during follow-up, 1=The event did occur during follow-up)
`Mi_fchd`	Hospitalized Myocardial Infarction or Fatal Coronary Heart Disease (0=The event did not occur during follow-up, 1=The event did occur during follow-up)
`Anychd`	Angina Pectoris, Myocardial infarction (Hospitalized and silent or unrecognized), Coronary Insufficiency (Unstable Angina), or Fatal Coronary Heart Disease (0=The event did not occur during follow-up, 1=The event did occur during follow-up)
`Stroke`	Atherothrombotic infarction, Cerebral Embolism, Intracerebral Hemorrhage, or Subarachnoid Hemorrhage or Fatal Cerebrovascular Disease (0=The event did not occur during follow-up, 1=The event did occur during follow-up)
`Cvd`	Myocardial infarction (Hospitalized and silent or unrecognized), Fatal Coronary Heart Disease, Atherothrombotic infarction, Cerebral Embolism, Intracerebral Hemorrhage, or Subarachnoid Hemorrhage or Fatal Cerebrovascular Disease (0=The event did not occur during follow-up, 1=The event did occur during follow-up)
`Hyperten`	Hypertensive (0=The event did not occur during follow-up, 1=The event did occur during follow-up)
`Death`	Death from any cause (0=The event did not occur during follow-up, 1=The event did occur during follow-up)
`Timeap`	Number of days from first examination to first `Angina` during the follow-up or number of days from first examination to censor date
`Timemi`	Defined as above for the first `Hospmi` event during follow-up
`Timemifc`	Defined as above for the first `Mi_fchd` event during follow-up
`Timechd`	Defined as above for the first `Anychd` event during follow-up
`Timestrk`	Defined as above for the first `Stroke` event during follow-up
`Timecvd`	Defined as above for the first `Cvd` event during follow-up
`Timehyp`	Defined as above for the first `Hyperten` event during follow-up
`Timedth`	Defined as above for `Death` during follow-up

In order to analyse the beforehand mentioned causal questions, we clean the data set in Section 6.2 in order to perform these analyses. Since there are missing values in the data set, we first describe the method used to handle these missing values in the following section.

6.1 Handling Missing Values

This section is based on [Stekhoven and Buhlmann, 2011, pp. 112–114]. Since FDA does not allow parametric statistical models for causal inference, as we described in the beginning of Chapter 3, we choose to use a non parametric imputation method to handle the missing values in the data set. Hence, in this section, we present the *miss forest* imputation method.

First, we examine which variables contain missing values. In Table 6.2, we present which variables in the data set we encountered missing values for as well as the number of missing values for these variables.

Table 6.2: The variables with missing values in the data with their associated number of missing values.

Variable Name	Number of Missing Values
`Heartrte`	6
`Bmi`	52
`Cigpday`	79
`Educ`	295
`Totchol`	409
`Bpmeds`	593
`Glucose`	1440
`Hdlc`	8600
`Ldlc`	8601

Notice that the data set is in long format, that is, each individual is represented by multiple rows in the data set and thus the number of missing values can exceed the number of individuals. Moreover, recall that `Hdlc` and `Ldlc` only are measured at the third examination and thus are missing by design for first and second examinations. We now introduce the notation used in order to describe the miss forest imputation method.

Let $X = (X_1, X_2, \ldots, X_p)$ where p is the number of variables and let n be the number of observations. Furthermore, for a variable X_s where $s \in \{1, \ldots, p\}$ containing missing values, let $i^{(s)}_{\text{mis}} \subset \{1, \ldots, n\}$ and $i^{(s)}_{\text{obs}} = \{1, \ldots, n\} \backslash i^{(s)}_{\text{mis}}$ be the entries containing missing values and observed values for the variable X_s, respectively. In the miss forest imputation method, for each variable X_s, categorical or continuous, we make the following split. Let $y^{(s)}_{\text{obs}}$ and $y^{(s)}_{\text{mis}}$ denote the observed values and missing values of X_s, respectively. Furthermore, let $x^{(s)}_{\text{obs}}$ and $x^{(s)}_{\text{mis}}$ denote the remaining variables at entries $i^{(s)}_{\text{obs}}$ and $i^{(s)}_{\text{mis}}$, respectively.

We now present the miss forest imputation method in Algorithm 1 where the stop criterion γ of the algorithm is described after the description of the method.

Algorithm 1 Miss Forest

Require: An $n \times p$ matrix X and stopping criterion γ.
  $X_{\text{new}}^{\text{imp}} \leftarrow$ Mean imputation for missing values in X;
  $k \leftarrow$ Sorted indices of variables in X with missing values in increasing order;
  while not γ **do**
    $X_{\text{old}}^{\text{imp}} \leftarrow X_{\text{new}}^{\text{imp}}$;
    for s in k **do**
      Fit a random forest model with $y_{\text{obs}}^{(s)}$ as the response variable and $x_{\text{obs}}^{(s)}$ as the predictors;
      Predict $y_{\text{mis}}^{(s)}$ with the data $x_{\text{mis}}^{(s)}$;
      $X_{\text{new}}^{\text{imp}} \leftarrow X_{\text{old}}^{\text{imp}}$ with new prediction of $y_{\text{mis}}^{(s)}$;
    end for
    Update γ;
  end while
  return $X_{\text{new}}^{\text{imp}}$

Thus, Algorithm 1 first replaces all the missing values by using a mean imputation. Afterwards, the variables which contained missing values in X are updated using a random forest model. Specifically, the variables are updated in increasing order based on the number of missing values of the variables. This process in iterated until the stop criterion γ is fulfilled which we present in the following.

Let N denote the index set of the continuous variables in X. Then the difference between the continuous variables in successive iterations of the miss forest imputation method is given by

$$\Delta_N = \frac{\sum_{j \in N} \left\| X_{\text{new},j}^{\text{imp}} - X_{\text{old},j}^{\text{imp}} \right\|^2}{\sum_{j \in N} \left\| X_{\text{new},j}^{\text{imp}} \right\|^2} \tag{6.1}$$

where $X_{\text{new},j}^{\text{imp}}$ and $X_{\text{old},j}^{\text{imp}}$ denotes the j'th column of $X_{\text{new}}^{\text{imp}}$ and $X_{\text{old}}^{\text{imp}}$, respectively. Furthermore, let F denote the index set of the categorical variables in X. The difference between these variables in successive iterations of the miss forest imputation method is given by

$$\Delta_F = \frac{\sum_{j \in F} \sum_{i=1}^{n} \mathbb{1}[(X_{\text{new},j}^{\text{imp}})_i \neq (X_{\text{old},j}^{\text{imp}})_i]}{\sum_{j \in F} |i_{\text{mis},F}^{(j)}|} \tag{6.2}$$

where $(X_{\text{new},j}^{\text{imp}})_i$ and $(X_{\text{old},j}^{\text{imp}})_i$ denotes the i'th entry in $X_{\text{new},j}^{\text{imp}}$ and $X_{\text{old},j}^{\text{imp}}$, respectively. Furthermore, $|i_{\text{mis},F}^{(j)}|$ denotes the number of missing values for j'th categorical variable. Then if there is an increase in Δ_N and Δ_F, for X consisting of both continuous and categorical variables, the stopping criterion, γ is fulfilled.

Having described the method used for handling the missing values of the data set considered in this chapter, we now present the cleaning of this data set in the following section.

6.2 Data Cleaning

In this section, we present the process of cleaning the data set introduced in the beginning of this chapter. The aim of this section is to prepare the data set for applying the L-TMLE method as well as the TMLE and the CV-TMLE methods in order to analyse the corresponding causal questions introduced previously in this chapter. Notice that for the TMLE and the CV-TMLE methods, we only use a subset of the data set used for the L-TMLE method. Thus, in particular, we present how the data set is prepared for the L-TMLE analysis.

The subset of the Framingham Heart Study considered in this chapter is as mentioned provided in long format, that is, for each individual, we have one row representing one examination attended by the individual. Hence, we have a maximum of three rows for each individual. For each of these rows, besides the time-varying variables in Table 6.1 which are measured at each examination, the variables measured for 24 years and the fixed variables are also provided. Notice that these particular variables are considered as fixed variables. In order to apply the L-TMLE method, we need to provide the data set in wide format, that is, one individual is represented by only one row. Moreover, we need to handle the missing values where we use the method presented in Section 6.1. In the following, we outline the process of converting the data set from long to wide format and how the missing values are imputed as well as how we treat unattended examinations which are not due to right censoring.

First, we notice that the variables `Hdlc` and `Ldlc` are only measured at the third examination. Thus, all missing values of these variables for the first two examinations are missing by design. Since we do not have any observations for these examinations, we decide to consider these variables as fixed. However, for individuals right censored before the third examination, we have no observation of these variables. Moreover, for some of the individuals who attended the third examination, these variables are also missing. Hence, we need to handle these specific missing values separately which we describe in the following.

Specifically, we first identify, the last examination attended for each individual. This is done by grouping the individuals based on `Randid` using the `group_by` function and then using the `filter` function in order to extract the the row with the maximal value of `Period`. This results in a data set consisting of only one row for each individual corresponding to the last attended examination. Having done so, we use `missForest` with parallel programming in order to obtain a more efficient imputation method to impute for the missing values of this data set. Then we use the `select` function to extract the columns `Randid`, `Period`, `Hdlc` and `Ldlc` in the imputed data set for the maximal period of each individual. This is done since we are only interested in the variables `Hdlc` and `Ldlc` where we use `Randid` and `Period` in order to combine these with the original data set.

The original data set is then left joined, using the `left_join` function, with this data set where we join by `Randid` and `Period`. This adds two new columns to the original data set containing the `Hdlc` and `Ldlc` columns from the imputed data set and hence, we remove the original `Hdlc` and `Ldlc` variables. Therefore, we now have one observation of `Hdlc` and `Ldlc` for each individual where this observation is present in the last attended examination.

Before handling the remaining missing values, we need to add rows for the individuals who did not attend an examination prior to their last attended examination. This could for example be an individual attending only examinations one and three. This is required since the `ltmle` function which we use for applying the L-TMLE method can only handle missing values due to right censoring. Hence, we use the `complete` function on `Randid` and `Period` in order to fill in rows for all individuals who did not attend all examinations, that is, one row for each individual for each value of `Period`. We note that the values of `Randid` and `Period` will also be filled in when applying the complete function. However, when doing so, we also add rows for the individuals who are lost to follow up due to right censoring. Since right censoring can be handled by the `ltmle` function, we do not want to impute for these individuals. Thus, we want to remove the rows which correspond to a right censoring which we explain in the following.

First, we identify which individuals did not attend all examinations. Hence, we begin with applying the `rle` function on the `Randid` variable in the original data set which returns a count of each value in `Randid`. That is, we have a list of how many examinations each individual attended. Then we extract from the data set the rows of the individuals who did not attend all three examinations. However, this data set then also contains the individuals who did not attend an examination prior to their last attended examination. Thus, we start by filtering the individuals which maximal value of `Period` is not three. This is done since for the individuals who attended the third examinations, we want to impute for the unattended examination. Thus, we now have a data set consisting of individuals who are right censored in the sense that they did not attend the third examination. For each of these individuals, we want to remove the row corresponding to the third examination. This is done by using `anti_join` on the complete data set and the data set containing the individuals censored in the third examination.

However, some of these individuals were already right censored for the second examination and thus, we want to identify these individuals in order to also remove the row corresponding to the second examination. This is done by extracting the rows corresponding to the individuals who were censored in the third examination from the data set constructed by anti joining the complete data set and the data set containing the censored individuals in the third examination. Then we `filter` based on the second `Period` and test which rows have missing values for `Sex`. Notice that the variable `Sex` had no missing values in the original data set and thus a missing value of this variable is introduced from adding a row representing an examination not attended. At last, we anti join this data set with the data set from the beforehand mentioned anti join. Thus, we have now removed the rows corresponding to right censoring.

Having added rows for the individuals who did not attend an examination prior to their last attended examination, we can now handle the missing values. First, we fill in the missing values of the constant variables which already have an observation for another examination by using the

`fill` function. This is mainly for the rows, we have just added and for the variables `Hdlc` and `Ldlc` which only have an observation at the last attended examination. Moreover, the variables `Age` and `Time` can be calculated based on their values for the attained examinations. Notice that we here assume that for each individual the time between the first and second examination is the same as the time between the second and third examination. Then we apply `missForest` with parallel programming for imputation of the remaining missing values.

Now what remains is to change the format of the imputed data set from long to wide. First, we combine the *time-varying variables* of Table 6.1 into two columns using `gather` where the first column contains the variable names and the second column contains their respective values. Since we want the data set in wide format, we want to be able to distinguish between for example `Age` from the first examination and `Age` from second examination. Thus, we use the `unite` function where we unite the column containing the variable names of the time-varying variables with `Period` such that for example `Age` from the first examination will be denoted as `Age_1` and, analogously, `Age` for the second and third examinations will be denoted `Age_2` and `Age_3`. Having done so, we now have a column of all the variables which we want to convert to wide format. Hence, we then use the `spread` function on the column with the variable names of the time-varying variables and the column with the corresponding values. The resulting data set is then in wide format and contains only missing value which are due to right censoring.

Having cleaned the data set, we in the following sections analyse the causal questions introduced in the beginning of this chapter using the TMLE, CV-TMLE and the L-TMLE methods, respectively.

6.3 TMLE and CV-TMLE

In this section, we analyse the causal question *"does smoking at the time of the first examination have a causal effect on stroke within a period of* 24 *years"*. In order to analyse this causal question, we use a subset of the data set obtained in Section 6.2. The data set consists of the fixed variables, the variables measured for 24 years and the variables corresponding to the first examination. We apply the TMLE and CV-TMLE methods in order to determine if there is an average causal effect of smoking at the time of the first examination on stroke within a period of 24 years. Specifically, we use the `tmle3` package and the usage of this package is based on [van der Laan et al., 2022].

By using the super learner package `sl3`, we define the methods which we want the super learner method to use in order obtain an initial estimate and an estimate of the clever covariate. As earlier mentioned, we do not use parametric methods in the super learner method. Thus, by using the `make_learner` function, we use the `xgboost`, `randomForest` and `mean` methods from the `sl3` package and the `ipredbagg` method from the `SuperLearner` package. Finally, we use the function `Lrnr_sl` to define the `learner_list` which is an argument in the `ltmle3` function specifying the list of methods used for the super learner method.

First, we use the CV-TMLE method. In order to do so, we apply the `tmle3` function which has CV-TMLE as default where we in particular use 10-fold cross validation. Specifically, we create a `spec` object which specifies the parameter of interest, that is, the average causal effect which is done by using the function `tmle_ATE`. We then use the `tmle3` function to estimate the average causal effect.

Then to use the TMLE method, we manually disable the additional layer of cross validation used in the CV-TMLE method. This is done by using the `ate_spec` function in combination with the `Targeted_Likelihood` function with the argument `updater=list(cvtmle=FALSE)`. We then use the `fit_tmle3` function to obtain the estimate. In the following table, the estimates and 95% confidence intervals obtained by the CV-TMLE and TMLE methods are shown.

Table 6.3: CV-TML and TML estimates with associated 95% confidence intervals.

Method	Estimate	95% Confidence Interval
CV-TMLE	-0.0020	[-0.0053, 0.0013]
TMLE	-0.0009	[-0.0029, 0.0011]

Notice that the 95% confidence intervals are calculated as explained in Section 4.3.

We observe in Table 6.3 that by applying the CV-TMLE and TMLE methods there is no average causal effect since zero is contained in both confidence intervals. Hence, we conclude that based on the TMLE and CV-TMLE methods, when considering the 95% confidence intervals, that there is no average causal effect of smoking at the time of the first examination on the outcome of stroke within a 24 years period.

6.4 L-TMLE

In this section, we analyse the causal question *"does smoking at the time of each of the examinations have a causal effect on stroke within a period of* 24 *years compared to not smoking at any of the examinations"*. In particular, we are interested in examining the two treatment strategies, presented in Section 5.1, $\bar{a}_T = (1, 1, 1)$ and $\bar{a}_C = (0, 0, 0)$. Thus, we examine the value of

$$\mathbb{E}[Y(\bar{A} = \bar{a}_T)] - \mathbb{E}[Y(\bar{A} = \bar{a}_C)]. \tag{6.3}$$

In order to analyse this causal question, we apply the L-TMLE method presented in Section 5.3. This method requires a particular time ordering. Thus, we first rearrange the variables such that we have the ordering $O = (W, A_1, L_1, \ldots, A_3, L_3, Y)$ where W denotes the fixed covariates, A_k the treatment at examination k, L_k the time-varying covariates at examination k and Y is the outcome.

However, as mentioned previously, the data set is subject to right censoring. The `ltmle` function which is used estimate the average causal effect by applying the LTMLE method can handle right censoring but in order to do so, we need to construct a time-varying censoring variable which specifies which individuals are censored at the three examinations. That is, this censoring variable only concerns with right censoring due to an individual being lost during follow up and not the individuals for which the outcome at the end of follow up has not occurred.

Hence, we construct the variables `c2` and `c3` denoting which individuals are censored before the second and third examinations, respectively. Notice that no individuals are censored before the first examination and thus, we only need to define the censoring variables corresponding to censoring prior to the remaining examinations. Since missing values now only occur due to right censoring, we can construct `c2` and `c3` based on whether or not any variables in L_2 and L_3, respectively, have missing values. This results in two binary variables for which we then use the `BinaryToCensoring` function such that the levels of these two variables are `censored` and `uncensored`. This is required for the censoring variables used in the `ltmle` function.

Returning to the time ordering $O = (W, A_1, L_1, \ldots, A_3, L_3, Y)$ then we want to include the censoring variables appropriately. Notice that, the `ltmle` function requires that missing values only occur after an individual has been censored, that is, the corresponding censoring variable must be placed prior to the censorings. Thus, it is important that the censoring variables are placed prior to the corresponding treatment variables since the treatment values also are missing if an individual has missing values for the corresponding time-varying variables since all these variables are measured at the examinations. Hence, if we denote the censoring variables as C_2 and C_3, we have the following time ordering $O = (W, A_1, L_1, C_2, A_2, L_2, C_3, A_3, L_3, Y)$.

Furthermore, we need to specify the algorithms used in the super learner method to obtain the initial estimate and an estimate of the clever covariate. We choose the methods `RandomForest` and `xgboost`. In Table 6.4, we present the L-TML estimate and the corresponding 95% confidence interval.

Table 6.4: LTML estimate with associated 95% confidence interval.

Method	**Estimate**	**95% Confidence Interval**
L-TMLE	-0.0009	$[-0.0342, 0.0325]$

By applying the L-TMLE method, when considering the 95% confidence interval in Table 6.4, we conclude that there is no average causal effect of smoking at the time of each of the examinations on stroke within 24 years compared to not smoking at any of the examinations.

7 | Conclusion

To summarise, we have in this master's thesis presented theory of causal inference for both non-longitudinal and longitudinal studies as well as examined various methods for estimating causal effects in practice. That is, we defined a causal effect, average causal effect and presented causal DAGs, (conditional) randomised experiments and the identifiability conditions. Then we considered the inverse probability weighting and standardisation methods which under the identifiability conditions can be used to identify average causal effects in observational studies.

Then we presented structural causal models which we used to define the target parameter where we in this master's thesis focused on the average causal effect. Afterwards, we described the TMLE method which consists of two steps where we in the first step obtain an initial estimator by applying the super learner method and in the second step we target the initial estimator such that we obtain an optimal bias-variance trade-off for the target parameter.

Furthermore, we considered the CV-TMLE method which is an extension of the TMLE method where we apply an extra layer of cross validation to the method. Then we showed that the CV-TML estimator is an asymptotic linear estimator where we also presented influence functions. Moreover, we extended the definition of a causal effect and an average causal effect and presented the identifiability conditions for observational longitudinal studies. This was done in order to present the L-TMLE method which is an extension of the TMLE method such that the method can be applied to longitudinal studies.

At last, we considered a subset of the Framingham Heart Study. First, we cleaned the data set where we applied the miss forest imputation method to adjust for missing values and converted the data to wide format. This was done in order to apply the TMLE, CV-TMLE and L-TMLE methods. In particular, we then applied the TMLE and CV-TMLE methods in order to analyse the causal question *"does smoking at the time of the first examination have a causal effect on stroke within a period of 24 years"*. At last, we applied the L-TMLE method in order to analyse the causal question *"does smoking at the time of each of the examinations have a causal effect on stroke within a period of 24 years compared to not smoking at any of the examinations"*. We then concluded from the confidence intervals for the TMLE and CV-TMLE methods that there was no causal effect of smoking at time of the first examination on the on outcome stroke within a 24 years period. Furthermore, we concluded from the confidence intervals for the L-TMLE method that there was no causal effect of smoking at the time of each of the examinations on stroke within 24 years compared to not smoking at any of the examinations.

8 | Bibliography

Jeremy R. Coyle. tmle3: The extensible TMLE framework. https://github.com/tlverse/tmle3, 2021. URL https://doi.org/10.5281/zenodo.4603358. [Cited 02-06-2022]. R package version 0.2.0.

Jeremy R. Coyle, Nima S. Hejazi, Ivana Malenica, Rachael V. Phillips, and Oleg Sofrygin. *sl3: Modern Pipelines for Machine Learning and Super Learning*. https://github.com/tlverse/sl3, 2021. URL https://doi.org/10.5281/zenodo.1342293. [Cited 02-06-2022]. R package version 1.4.2.

Aaron Fisher and Edward H. Kennedy. *Visually Communicating and Teaching Intuition for Influence Functions*, 2018. URL https://arxiv.org/abs/1810.03260. [Cited 25-04-2022].

Ralf Herbrich. *Learning Kernel Classifiers: Theory and Algorithms*. The MIT Press, 2002. ISBN 0-262-8306-X.

Søren Højsgaard, David Edwards, and Steffen Lauritzen. *Graphical Models with R*. Springer, 2012. ISBN 978-1-4614-2298-3.

Steffen L. Lauritzen. *Graphical Models*. Oxford University Press, 1996. ISBN 0-19-852219-3.

Samuel D. Lendle, Joshua Schwab, Maya L. Petersen, and Mark J. van der Laan. *ltmle: An R Package Implementing Targeted Minimum Loss-Based Estimation for Longitudinal Data*. *Journal of Statistical Software*, 81(1):1–21, 2017. doi: 10.18637/jss.v081.i01. URL https://www.jstatsoft.org/index.php/jss/article/view/v081i01. [Cited 18-05-2022].

James M. Robins Miguel A. Hernán. *Causal Inference: What If*. Chapman & Hall/CRC, 2020.

Brady Neal. *Introduction to Causal Inference*, 2020. URL https://www.bradyneal.com/Introduction_to_Causal_Inference-Dec17_2020-Neal.pdf. [Cited 30-03-2022].

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009. ISBN 978-0-521-89560-6.

Judea Pearl. *An Introduction to Causal Inference*. *The International Journal of Biostatistics*, 6 (2):Article 7, 2010. doi: 10.2202/1557-4679.1203.

James M. Robins and Miguel A. Hernán. *Longitudinal Data Analysis*. Chapman & Hall/CRC, 2009. ISBN 978-1-58488-658-7.

Michael Schomaker, Miguel Luque-Fernandez, Valeriane Leroy, and Mary-Ann Davies. *Using longitudinal targeted maximum likelihood estimation in complex settings with dynamic interventions*. *Statistics in medicine*, 38(24):4888–4911, 2019. doi: 10.1002/sim.8340. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6800798/pdf/nihms-1043657.pdf. [Cited 26-05-2022].

D. J. Stekhoven and P. Buhlmann. *MissForest–non-parametric missing value imputation for mixed-type data*. *Bioinformatics*, 28(1):112–118, oct 2011. doi: 10.1093/bioinformatics/btr597. URL https://doi.org/10.1093%2Fbioinformatics%2Fbtr597. [Cited 20-05-2022].

Framingham Heart Study Longitudinal Data Documentation. National Heart, Lung and Blood Institute, 2021. URL https://biolincc.nhlbi.nih.gov/media/teachingstudies/FHS_Teaching_Longitudinal_Data_Documentation_2021a.pdf?link_time=2022-02-20_19:01:29.259994. [Cited 24-05-2022].

Mark van der Laan, Jeremy Coyle, Ivana Malenica Nima Hejazi, and Alan Hubbard Rachael Phillips. *Targeted Learning in R: Causal Data Science with the tlverse Software Ecosystem*, 2022. URL https://tlverse.org/tlverse-handbook/. [Cited 19-5-2022].

Mark J. van der Laan and Sherri Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA, 2011. ISBN 978-1-4419-9781-4.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer-Verlag New York, Inc., 1996. ISBN 0-387-94640-3.

Wenjing Zheng and Mark J. van der Laan. *Asymptotic Theory for Cross-validated Targeted Maximum Likelihood Estimation*. *U.C. Berkeley Division of Biostatistics Working Paper Series*, Working paper 273, 2010. URL http://biostats.bepress.com/ucbbiostat/paper273. [Cited 21-03-2022].