Chapter 4

Causal Modeling with Observational Data

4. Introduction

Causal inference focuses on causal relationship identification and estimation, which is an important concern for research conducted within the fields of economics and social sciences, given that policy interventions and theories can be explained based not only on associations but also causal relations. In an ideal scenario, causal effects can be identified using experimental data, i.e., through experiments that use the concept of randomized controlled trials, so that comparison can be made without any kind of bias between groups and the control group [65]. But, in economics and most social sciences, it is not possible to perform experiments due to high costs. Thus, it is often the case that researchers have to rely on observational data, which are collected in a manner that is non-experimental, i.e., they are collected in a non-random manner, which gives rise to challenges such as confounding, selection, and reverse causality [66]. Yet, the main interest in observational data stems from the nature of the challenges, as they are a reflection of real-world data collected over a wide range of populations over long periods of time, allowing for the estimation of effects of events not amenable to manipulation. Hence, more effective methodologies are formulated for estimating causal effects from non-experimental studies.

4.1 Why Observational Data Matters

4.1.1 Limitations of Randomized Experiments

Even though randomized experiments are considered the gold standard for making causal inferences, some of the drawbacks of randomized experiments have to be addressed in detail. Ethical conditions associated with randomized experiments include the inability of researchers to randomly expose people to harmful interventions as well as to withhold beneficial interventions in domains like health, educational settings, and policy [67]. Secondly, randomized experiments can be expensive and sometimes challenging to conduct on a large scale. Even when the experiment is conducted with a degree of success, it remains a problem that such experiments often lack external validity, which means that while they are conducted in a controlled atmosphere, the results obtained in such a controlled situation cannot be generalized well to a larger or varied population, institutions, or time period [68]. In fact, while results from an experiment done with randomization assure a fair degree of internal validity, it cannot truly reflect the complex nature of a real-world scenario.

4.1.2 Advantages of Observational Data

Observational data also possess several important advantages that make it particularly valuable as a tool for economic and social science research. These include its usually large scale and real-world nature, allowing researchers to analyze behavior and outcomes as they occur in real-life settings and across varied settings and populations. Also important is the fact that observational data is usually not subject to the limitations set by time and history, which is evident in most experiments. More particularly, observational data allows researchers to analyze rare events and long-run effects such as financial crises, intergenerational mobility, and early-life events, which would otherwise be impossible or extremely difficult through random assignment-based experiments [69]. All these make observational data very central in most economic and related social science research, despite its challenges for causal inference.

4.1.3 Challenges with Observational Data

The observational data, while extremely useful, also presents important challenges in terms of constructing causal inference. An important challenge is that of confounding variables, or those that are related at least to some degree to our treatment and our outcomes, such that it becomes difficult to disentangle our desired causal effect. The use of observed treatment effects, as found generally in these types of observational data, must confront another critically important challenge: that of selection bias, where individuals or groups tend to select themselves into treatment or control groups on the basis of characteristics that are, at least partially, unobservable by our researchers. Finally, issues of measurement error or data missingness, either one or both, raise important questions about our results, casting doubt on our established empirical relationship.

4.2 Causal Graphs and Directed Acyclic Graphs (DAGs)

4.2.1 Introduction to Causal Graphs

A causal graph is a formal tool that is often employed to depict causal associations among variables within a system. The key aim for employing a causal graph is to ensure the presentation of a clear picture of how distinct variables are assumed to influence each other causally. Further, a causal graph has its variables represented by nodes, while arrows are employed to show the direction of influence from the cause to the effect [71]. Essentially, a causal graph is considered instrumental in helping authoritatively understand the potential causes of confounding when drawing causal inferences among variables within a system.

4.2.2 Directed Acyclic Graphs (DAGs)

Directed Acyclic Graphs, abbreviated as DAG, is a special form of causal graphs that is used in representing cause-effect relationships in a system in a well-structured and logically consistent manner. The term directed is used in the context of directed arrows, which reflect causal influences from a cause to its actual effect, and acyclic relates to the absence of feedback cycles in a causal framework, where a variable can neither directly nor indirectly cause itself. The reason why DAGs have significant application and usage in economics is that they explicitly discuss the causal assumptions, showing how one variable is affected through various connections. Using basic economic concepts, a DAG might have education as one variable affecting another variable, income, which then affects another variable, health (education → income → health), demonstrating direct and indirect causal effects [72].

4.2.3 Confounders, Mediators, and Colliders

In causal analysis, variables that are considered as confounders, mediators, and colliders play important roles, and it is always important to identify the roles of these variables in understanding their influence in specifying models correctly as well as in obtaining correct causal analyses. Confounders are variables that affect both the outcome and treatment and, unless properly controlled, end up creating a spurious association. Mediators, on the other hand, explain how to carry out a causal effect and thus, when controlled, they block the effect of interest. The appearance of collider bias is also possible when variables influence other two or more variables and one is controlled, which results in artificial correlation with its causes [73]. Identifying the roles played by these variables, although using causal graphs, helps one establish which variables should always be controlled and which variables should never be controlled in analysis.

4.2.4 Using DAGs for Causal Reasoning

DAGs are very powerful tools for causal reasoning because they explicitly and transparently lay out the structure of causal relationships. Mapping out variables and the directed paths between them, DAGs help researchers identify the causal paths-a treatment takes to affect an outcome as well as non-causal paths that may introduce bias. They also make visible potential sources of spurious correlations, such as confounding or collider bias, which may not be immediately obvious from data alone [74]. Using this graphical approach, a researcher determines appropriate adjustment strategies, explicates their underlying assumptions more clearly, and strengthens the validity of the causal interpretations from observational data. Key Elements in Causal Graphs and DAGs is illustrated in table 4.1.

Table 4.1: Key Elements in Causal Graphs and DAGs

Element	Definition	Role in Causal Analysis	Example
Node	A variable represented in the graph	Represents treatments, outcomes, or covariates	Education, Income, Health
Edge	A directed arrow indicating causal influence	Shows direction of causality from cause to effect	Education → Income
Confounder	A variable that influences both the treatment and outcome	Can create spurious associations; must adjust for it to get unbiased estimates	Parental Socioeconomic Status affecting both Education and Health
Mediator	A variable lying on the causal pathway between treatment and outcome	Explains how or through what mechanism a causal effect operates; adjusting may block part of the effect	Income mediating Education → Health
Collider	A variable influenced by two or more other variables	Conditioning on it can introduce bias by creating artificial correlations	Health behaviors influenced by Smoking and Exercise
Pathway	A sequence of edges connecting nodes	Determines causal and non-causal relationships	Education → Income → Health
Acyclic Structure	No cycles exist; no variable can cause itself	Ensures logical consistency of causal assumptions	Education cannot cause itself directly or indirectly

4.3 Backdoor and Frontdoor Criteria

4.3.1 The Backdoor Criterion

The backdoor criterion basically is a central concept within causal inference, providing a systematic approach toward identifying variables in the case of which the control is necessary to take from an unbiased estimation of a causal effect. Intuitively, it requires blocking all "backdoor paths" between a treatment and a result, which means those

paths that create spurious associations because of common causes rather than under the influence of actual causation. In a DAG, backdoor paths can be detected by tracing all the non-causal routes that enter the treatment variable from behind and lead to the outcome. Conditioned on an appropriate set of confounding variables by which these backdoor paths are blocked, researchers will then be isolating the true causal effect of the treatment [75]. For example, when estimating the impact of education on income, adjusting for family background or innate ability may block backdoor paths and help ensure that the estimated relationship reflects causality rather than confounding influences.

4.3.2 The Frontdoor Criterion

The front door criterion is a framework for causal identification that is employed when direct control of confounding is either not feasible or is considered inadequate, but an intermediate outcome is correlated. It is a viable option when the outcome is affected only through the intermediate outcome, provided that the intermediate outcome is not subject to the confounding which affects the relationship between the treatment and outcome. It requires that the treatment is causally affecting the mediator, and the mediator in turn causally affects the outcome, whilst all back-door paths from mediator to outcome need to be blocked by the observed variables [76]. For instance, if education affects health through income, and income is well-measured and not confounded with health by unobserved factors, researchers can use income as a mediator to identify the causal effect of education on health even when direct confounding between education and health cannot be fully controlled.

4.3.3 Selecting Adjustment Sets

Covariate adjustment sets selection is a vital step in making causal inference since such selection influences the results obtained in terms of causal effects estimation. To make the best out of the situation and come up with results that are less likely to be influenced by confounding elements, various methods such as causal graphs and the backdoor criterion are employed. Balancing the risk of bias reduction and overadjustment is the biggest issue affecting the decision to use one variable for adjustment over the use of another. Overadjustment might influence the results obtained in such a way that the causal effect of interest is eliminated [77]. Backdoor and Frontdoor Criteria for Causal Adjustment is depicted in Table 4.2.

Table 4.2: Backdoor and Frontdoor Criteria for Causal Adjustment

Criterion	Purpose	Key Conditions	Adjustment Strategy	Example
Backdoor Criterion	Identify confounders that create spurious paths between treatment and outcome	1. All backdoor paths must be blocked 2. No variable in adjustment set is a descendant of the treatment	Condition on confounders that block non-causal paths	Estimating effect of Education on Income: adjust for Family Background, Innate Ability
Frontdoor Criterion	Identify causal effect via a mediator when direct confounder adjustment is not possible	1. Treatment affects mediator 2. Mediator affects outcome 3. All backdoor paths from mediator to outcome are blocked	Adjust for mediator to estimate causal effect	Estimating effect of Education on Health via Income: use Income as mediator if unobserved confounding exists between Education and Health
Adjustment Set Selection	Choose variables to include in regression or matching to avoid bias	1. Block confounding paths 2. Avoid conditioning on colliders or mediators unnecessarily 3. Use subject-matter knowledge and DAG guidance	Use DAGs and backdoor/frontdoor logic to select covariates	For Education → Health: include Family Background but exclude Income if interested in total effect

4.4 Matching and Weighting Methods

4.4.1 Propensity Score Matching (PSM)

Propensity Score Matching is an increasingly popular approach to causal inference with observational data that tries to reduce the selection bias by constructing comparable treatment and control groups. The propensity score is defined as the probability of receiving the treatment given a set of observed covariates, summarizing multiple confounding variables into a single measure. This procedure typically consists

of three steps: first, estimation of the propensity scores using a model such as logistic regression; second, matching treated and untreated units with similar propensity scores; and third, checking the balance of the covariates to make sure the matched groups are comparable [78]. While PSM seems very attractive for its intuitive framework and high-dimensional covariates handling, it has all its limitations due to its reliance only on observed variables and cannot account for unmeasured confounding, which makes careful model specification and diagnostic checks important. It can be mathematically represented as

$e(X) = P(W = 1 \mid X) = Pr(\text{Treatment} \mid \text{Covariates})$

Where:

- **$e(X)$** = propensity score (probability of treatment assignment)
- **W** = treatment indicator ($W = 1$ if treated, $W = 0$ if control)
- **X** = vector of observed covariates (confounding variables)
- **$P(\cdot)$** = probability

Logistic Regression Specification (most common estimation method):

$e(X) = 1 / (1 + \exp(-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k)))$

Or equivalently in logit form:

$\text{logit}(e(X)) = \log(e(X)/(1-e(X))) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$

Average Treatment Effect on the Treated (ATT) using PSM:

$\tau\_ATT = E[Y(1) - Y(0) \mid W = 1] = E[Y(1) \mid W = 1] - E[Y(0) \mid W = 1]$

Estimated using matched pairs:

$\hat{\tau}\_ATT = (1/N_1) \Sigma_{i \in \{treated\}} [Y_i - \hat{Y}_i(0)]$

where $\hat{Y}_i(0)$ is the average outcome of control units matched to treated unit i based on similar propensity scores.

4.4.2 Inverse Probability Weighting (IPW)

Inverse Probability Weighting (IPW) is a method of making causal inferences based on data collected under observational studies by creating a pseudo-population in which the treatment received by the individuals in the data is independent of other observed covariates, similar to what occurs in experiments. This is accomplished by weighing each individual in the data by the inverse of the probability of receiving the treatment

received by the individual, where probability is usually calculated using the propensity score model. Here, the treated units will be weighted by the inverse of the calculated propensity score, and the control units will be weighted by the inverse of one minus the calculated scores, with greater weight on underrepresented units [79]. IPW is frequently used in causal inference for effect estimation in average treatment effects and can be broadly used in policy analyses and epidemiology but with the condition that the model used for estimation will be correctly specified with all confounders being observed. IPW can be formulated as below

$$\hat{\tau}\_ATE = (1/N) \Sigma_{i=1}^{N} [W_i Y_i / e(X_i) - (1-W_i) Y_i / (1-e(X_i))]$$

Where:

- $\hat{\tau}\_ATE$ = estimated average treatment effect
- N = total sample size
- W_i = treatment indicator (1 if treated, 0 if control)
- Y_i = observed outcome for unit i
- $e(X_i)$ = propensity score (probability of treatment given covariates X)
- $1/e(X_i)$ = weight for treated units (inverse of propensity score)
- $1/(1-e(X_i))$ = weight for control units (inverse of probability of being in control)

Individual Weights:

For treated units ($W_i = 1$): $w_i = 1/e(X_i)$

For control units ($W_i = 0$): $w_i = 1/(1-e(X_i))$

4.4.3 Other Methods

Besides the simple methods of matching and weighting, several other techniques have also been developed to enhance covariate balance in the process of making cause-and-effect analysis with observational data. Covariate balancing techniques focus primarily on the optimal balance of covariates between the treated and control groups rather than purely on the estimation of well-behaved propensity score models. Genetic matching builds on the basic approach to matching by making use of a search algorithm to generate the optimal weighting schemes. Doubly robust estimation combines outcome modeling and adjustment for covariates using the propensity score, but only requires that either the first or second stage is specified correctly. Doubly robust estimation provides better flexibility and reliability for causal inference, especially in the presence of possible model misspecifications. Matching and Weighting Methods for Causal Inference is briefly described through table 4.3.

Table 4.3: Matching and Weighting Methods for Causal Inference

Method	Key Idea	Main Steps	Advantages	Limitations
Propensity Score Matching (PSM)	Matches treated and control units with similar propensity scores	1. Estimate propensity scores using a model (e.g., logistic regression) 2. Match treatments and controls based on Propensity scores 3. Assess the covariate balance	Intuitive and reduces bias from observed confounders; handles high-dimensional covariates	Cannot adjust for unmeasured confounders; sensitive to model specification
Inverse Probability Weighting (IPW)	Weights observations to create a pseudo-population mimicking randomization	1. Estimate propensity scores 2. Compute weights: treated = 1/PS, control = 1/(1-PS) 3. Use weighted sample to estimate treatment effects	Handles multiple treatment groups; flexible for policy evaluation	Large weights can increase variance; relies on correct model specification
Covariate Balancing Methods	Optimizes balance of covariates directly rather than relying on propensity scores	1. Define balance criteria 2. Adjust weights or select units to achieve optimal balance	Reduces reliance on model specification; improves balance	Computationally intensive; less common in practice
Genetic Matching	Uses search algorithms to find optimal weights for balance across covariates	1. Specify covariates and matching algorithm 2. Iteratively adjust weights to minimize imbalance	Automated optimization for better balance	Computationally heavy; complex implementation

		3. Match treated and control units		
Doubly Robust Methods	Combines outcome modeling with propensity score adjustment	1. Model treatment assignment (propensity score) 2. Model outcome 3. Combine estimates to achieve robustness	Consistent if either treatment or outcome model is correct; more reliable	

4.5 Sensitivity Analysis

4.5.1 Motivation

The rationale for sensitivity analysis arises from the awareness that observational data might be subject to differing levels of unmeasured confounding, which cannot be directly addressed through control in empirical estimation. The validity of even sophisticated approaches to causal estimation ultimately relies on the assumption that any such confounders are low in effect. Hence, it is critical that sensitivity analysis be conducted to check how robust a particular estimate of causality is, even in the face of violations of these types of assumptions.

4.5.2 Approaches

A number of methods are frequently employed to critically review the dependability of causal inferences in observational studies via sensitivity analysis. By applying Rosenbaum Bounds, researchers can critically review the strength of hidden confounding influencing treatment effects to determine whether a statistically significant effect could be reversed. Furthermore, simulations and scenario analyses widen Rosenbaum Bounds by taking into account hypothetical confounding variables to determine their effect on the results of a study [80]. The critical use of the aforementioned sensitivity analysis techniques by researchers helps to critically review the strength of confounding bias to determine whether the obtained results are reliable or highly sensitive to confounding variables.

4.6 Example Application: Healthcare Spending and Outcomes

4.6.1 Problem Setup

The question is specified as a problem, and it is defined around a research question of whether an increase in healthcare spending and expenditure improves health outcomes for patients. To answer this research question, an observational dataset is considered, as it is typically collected rather than an experiment. The dataset, for example, will have various factors for healthcare spending, health outcomes, and various covariates such as age, income, pre-existing conditions, and access to medical facilities, among others, which could impact health outcomes or health spending. The research question and research strategy are important aspects that define and narrow down a problem and ensure that various aspects, such as data sources and key elements of the research, are properly defined.

4.6.2 DAG Representation

The DAG is used as a tool for visually describing the causal pathways associated with the relationship between healthcare spending and patient health outcomes. In its structure, the amount spent on healthcare is considered the treatment, health outcomes are considered the outcome, and variables like income, age, initial state of well-being, and healthcare service provision are considered to be potential confounding variables associated with both healthcare spending and health outcomes [81]. With its use, it is possible to identify backdoor variables that produce spurious correlations between treatment and outcome through various confounding variables. Using its structure, it is possible to identify which variables must be controlled for in order to prevent backdoor variables and provide an unbiased estimate of the effect of healthcare spending on health outcomes.

4.6.3 Applying Matching and Weighting

Methods for matching and weighting are used in the observational health care data set to account for confounding and to approximate a randomized comparison. PSM or IPW techniques balance the observed covariates of age, income, and baseline health between those patients with higher and lower health care spending [82]. Once acceptable balance has been achieved, outcome comparisons are made between the treated and control groups to estimate the causal effect of increased spending on patient health. This method allows researchers to reduce selection bias and make more credible conclusions concerning the impact of health care spending using non-experimental data.

4.6.4 Sensitivity Analysis

Sensitivity analysis also helps to assess how robust the estimated effects of healthcare spending on the health of patients are. It examines whether these effects are robust to some factors that are not observed. Such factors may include patients' lifestyles and their behaviors. Hence, examining these effects serves to help determine whether they are strong and believable, or whether they are sensitive to potential biases. Thus, it helps to either strengthen or qualify the outcome of the causal analysis of patients' health [83]. Factors in Healthcare Spending Causal Analysis are presented in table 4.4.

Table 4.4: Variables in Healthcare Spending Causal Analysis

Variable	Type	Role in Analysis	Notes / Example
Healthcare Spending	Continuous	Treatment	Total spending per patient over a defined period; primary independent variable
Patient Health Outcome	Continuous / Binary	Outcome	Could be overall health score, mortality, hospital readmission, or quality-adjusted life years
Age	Continuous	Confounder / Covariate	Older patients may spend more and have worse outcomes; must be adjusted for
Income	Continuous / Categorical	Confounder / Covariate	Socioeconomic status may influence access to care and health outcomes
Baseline Health Status	Continuous / Categorical	Confounder / Covariate	Pre-existing conditions affect both spending and outcomes
Access to Healthcare Facilities	Binary / Categorical	Confounder / Covariate	Proximity or availability of clinics/hospitals may affect spending and health
Lifestyle Factors	Continuous / Categorical	Potential Unobserved Confounder	Not always measured (diet, exercise, smoking), assessed in sensitivity analysis

4.7 Summary and Key Takeaways

In summary, observational data might be extremely useful for causal analysis when randomized experiments are not feasible or ethical or, indeed, practical; when opportunities for such research abound with regard to economic and social conditions. Most importantly, DAGs) provide a rigorous framework for modeling causality explicitly, thus highlighting the underlying assumptions and enabling the identification of bias in causal inference. More notably, the backdoor and frontdoor criteria provide a set of clear requirements for choosing appropriate adjustments for causal estimation. Matching and weighting are also used to mimic randomized experiments for observational data; this increases the credibility of causal effects for the research. Notably, sensitivity analysis plays a vital part; success stories such as healthcare spending provide examples for the applicability of the proposed ideas and concepts with regard to causal inference.