Causal Modeling: A Conceptual Framework

1 Introduction

Causal modeling aims to estimate the effect of an intervention on an outcome, distinguishing true causal effects from statistical associations observed in data. Unlike standard regression approaches, which describe relationships conditional on observed variables, causal models explicitly address counterfactual questions such as what would have happened under an alternative treatment assignment. This document presents a conceptual overview of causal modeling using the potential outcomes framework.

2 Potential Outcomes Framework

The foundation of causal inference is the potential outcomes framework. For each individual, two counterfactual outcomes are defined:

$$Y(1) : \text{the outcome if the individual receives the treatment,}$$

$$Y(0) : \text{the outcome if the individual does not receive the treatment.}$$

At any given time, only one of these outcomes is observed, while the other remains unobserved. As a result, individual-level causal effects cannot be directly identified from observed data.

3 Causal Estimand

Because individual causal effects are unobservable, causal inference focuses on population-level quantities known as causal estimands. The most commonly studied estimand is the average treatment effect (ATE), defined as

$$\text{ATE} = \mathbb{E}[Y(1) - Y(0)].$$

The ATE represents the expected difference in outcomes if the entire population were exposed to the treatment compared to if the entire population were not exposed.

3.1 Average Treatment Effect (ATE)

The Average Treatment Effect (ATE) measures the expected difference in outcomes between treatment and control across the entire population:

$$\text{ATE} = \mathbb{E}[Y(1) - Y(0)].$$

The ATE represents the causal effect of a hypothetical intervention applied universally.

3.2 Average Treatment Effect on the Treated (ATT)

The Average Treatment Effect on the Treated (ATT) focuses on individuals who actually received the treatment:

$$\text{ATT} = \mathbb{E}[Y(1) - Y(0) \mid T = 1].$$

ATT answers the question of how much the treated group benefited from receiving the treatment compared to if they had not been treated.

3.3 Conditional Average Treatment Effect (CATE)

The Conditional Average Treatment Effect (CATE) captures heterogeneity in treatment effects across subpopulations defined by covariates X:

$$\text{CATE}(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x].$$

CATE allows causal effects to vary with individual characteristics and is central to personalized decision-making.

Real-World Example. Consider a medical study evaluating the effect of a new drug (T) on blood pressure reduction (Y). Let X include patient age. The overall average treatment effect (ATE) may suggest a modest benefit, but the effect may differ by age group.

For example,

$$\text{CATE}(\text{Age} < 50) = -8 \text{ mmHg},$$

$$\text{CATE}(\text{Age} \geq 50) = -2 \text{ mmHg}.$$

This indicates that younger patients experience a substantially larger reduction in blood pressure from the treatment compared to older patients.

Interpretation. CATE answers the question: *"What is the causal effect of the treatment for individuals with characteristics $X = x$?"* Unlike ATE, which averages effects over the entire population, CATE enables subgroup-specific and personalized causal inference.

Relation to ATE. The average treatment effect can be expressed as an average of conditional effects:

$$\mathrm{ATE} = \mathbb{E}_X[\mathrm{CATE}(X)].$$

Thus, ATE summarizes overall impact, while CATE reveals effect heterogeneity across different subpopulations.

4 Identification Assumptions

Identification of causal effects from observational data requires several key assumptions:

- **Exchangeability**: conditional on observed covariates, treatment assignment is independent of the potential outcomes, implying no unmeasured confounding.
- **Positivity**: each individual has a positive probability of receiving every treatment level.
- **Consistency**: the observed outcome corresponds to the potential outcome under the treatment actually received.

When these assumptions hold, causal effects can be estimated from observed data.

5 Interventions and the do-Operator

A key distinction between causal and associational models is the representation of interventions. Causal models use the $\mathrm{do}(\cdot)$ operator to denote external interventions that set a treatment variable to a specific value. For a treatment variable T, an intervention is written as $\mathrm{do}(T = t)$. Causal questions are therefore expressed as comparisons between interventional distributions:

$$\mathbb{E}[Y \mid \mathrm{do}(T = 1)] \quad \text{vs.} \quad \mathbb{E}[Y \mid \mathrm{do}(T = 0)].$$

6 Relation to Regression Models

Standard regression models describe associations between covariates and outcomes based on the observed data distribution. These associations do not necessarily correspond to causal effects. Regression coefficients can only be interpreted causally if the causal assumptions are satisfied and all relevant confounders are properly controlled. Without these conditions, regression-based estimates should be interpreted as associational rather than causal.

7 Causal Model Types

Causal inference does not rely on a single model, but on a class of models that target well-defined causal estimands under explicit assumptions. Below are the main causal model types commonly used in practice.

7.1 Propensity Score Model

The propensity score is a model for the treatment assignment mechanism, defined as

$$e(X) = P(T = 1 \mid X),$$

where T denotes the treatment and X observed covariates. The propensity score itself is typically estimated using a logistic regression or other classification models. While it does not model the outcome directly, it plays a central role in causal inference by balancing covariates between treatment groups.

Propensity scores are used in several causal modeling strategies, including matching, stratification, covariate adjustment, and weighting.

7.2 Inverse Probability Weighting (IPW)

Inverse probability weighting constructs a pseudo-population in which treatment assignment is independent of covariates. Each individual is assigned a weight based on the inverse of the propensity score:

$$w_i = \frac{T_i}{e(X_i)} + \frac{1 - T_i}{1 - e(X_i)}.$$

In this weighted population, the average treatment effect (ATE) can be estimated as a weighted contrast of outcomes. IPW directly targets interventional quantities rather than conditional associations.

7.3 Outcome Regression (G-Formula)

Outcome regression models the conditional expectation of the outcome given treatment and covariates:

$$\mathbb{E}[Y \mid T, X].$$

Under causal assumptions, the interventional mean is identified via the g-formula:

$$\mathbb{E}[Y(t)] = \mathbb{E}_X[\mathbb{E}[Y \mid T = t, X]]\,.$$

This approach relies on correct specification of the outcome model.

7.4 Marginal Structural Models (MSMs)

Marginal Structural Models directly parameterize causal effects through a model for interventional expectations:

$$\mathbb{E}[Y^{\mathrm{do}(T)}] = \alpha_0 + \alpha_1 T.$$

The parameters of MSMs are estimated using inverse probability weights. Unlike standard regression models, MSM coefficients represent causal effects under the assumed intervention.

7.5 Doubly Robust Models

Doubly robust estimators, such as augmented inverse probability weighting (AIPW) and targeted maximum likelihood estimation (TMLE), combine an outcome model and a propensity score model. These methods are consistent if either the outcome model or the propensity score model is correctly specified, making them particularly attractive in applied causal analysis.

7.6 Key Distinction

Propensity score models are not causal estimands themselves; they are models for the treatment assignment mechanism. Causal effects emerge only when propensity scores are combined with outcome information through weighting, matching, or structural models.

8 Model Comparison in Causal Inference

Model comparison in causal inference differs fundamentally from predictive or associational modeling. Causal models are not compared based on goodness-of-fit or predictive accuracy, but on their ability to estimate a well-defined causal estimand under plausible assumptions.

8.1 Estimand Alignment

The first requirement for causal model comparison is that all models target the same causal estimand. Models estimating different quantities, such as the average treatment effect (ATE),

$$\text{ATE} = \mathbb{E}[Y(1) - Y(0)],$$

and the average treatment effect on the treated (ATT),

$$\text{ATT} = \mathbb{E}[Y(1) - Y(0) \mid T = 1],$$

cannot be meaningfully compared. Only models targeting the same estimand are eligible for comparison.

8.2 Covariate Balance Diagnostics

For propensity score–based models, comparison is primarily based on covariate balance rather than likelihood. Balance is commonly assessed using standardized mean differences (SMDs):

$$\text{SMD} = \frac{\bar{X}_1 - \bar{X}_0}{s},$$

where values close to zero indicate good balance. A causal model that achieves better balance across confounders is preferred, even if it fits the data less well.

8.3 Stability of Causal Estimates

Causal models are compared by examining the stability of estimated treatment effects across reasonable model specifications. Large changes in the estimated ATE or ATT when modifying the propensity score model, trimming extreme weights, or altering covariate sets indicate model sensitivity and weaker causal credibility.

8.4 Weight Diagnostics

For inverse probability weighting (IPW) and marginal structural models, the distribution of weights is critical. Models producing extreme or highly variable weights are generally disfavored, as they increase variance and reduce effective sample size:

$$\mathrm{ESS} = \frac{(\sum_i w_i)^2}{\sum_i w_i^2}.$$

8.5 Doubly Robust Comparisons

In doubly robust approaches, such as augmented inverse probability weighting or TMLE, comparison focuses on robustness. A model is preferred if the estimated causal effect remains stable when either the outcome model or the propensity score model is perturbed.

8.6 Sensitivity Analysis

Causal model comparison often includes sensitivity analyses to unmeasured confounding or violations of assumptions. Models whose conclusions are less sensitive to these perturbations are considered more credible.

8.7 Key Principle

In causal inference, the best model is not the one with the best fit, but the one that most plausibly represents the intervention of interest, achieves adequate covariate balance, and yields stable estimates of the target causal estimand.

9 Conclusion

Causal modeling provides a principled framework for estimating the effects of interventions using counterfactual reasoning. By clearly defining causal estimands, assumptions, and interventions, causal models establish a clear distinction between association and causation and enable meaningful causal interpretation when their assumptions are met.