Notes on Causality

Tomás Paiva de Lira

December 16, 2025

Abstract

The subject of causality has accompanied men on the journey of *understanding* since its inception: from Aristotle's postulation of the four causes that delineate physical phenomena, to Hume's framework of constant conjunctions. A paradigm shift occurred at the turn of the 20th century, a time when the philosophy of Positivism held the most adherence in the scientific community. As a result, the founders of modern statistics, such as Karl Pearson and Francis Galton, rejected the inquiry into that which had no empirical grounding. Restricting the scope of statistics to answering "what" rather than "why". However, motivated to mitigate the missing data problem, Rubin proposed a model based on potential outcomes that was inherently causal, but restricted its scope to measuring effects. Rubin's Causal Model joined causality back with statistics. The objective of these notes is to elucidate the modern applications of causality through the agnostic perspective that measure theory permits. Moreover, once proven, these causal techniques are put to the test with real examples to understand their limitations and successes.

Contents

1 Potential Outcomes

1.1 Answering a Different Type of Question

Correctly using any causal framework requires understanding the types of questions and premises assumed by standard statistical procedures, such as linear regression, and knowing how to overcome their limitations. In light of this, our exposition of causality will be presented by contrasting the different styles of *thought* between these two paradigms by applying them to real-world data.

It is required to understand what a *causal effect* is. In regression, our goal is to maximize the likelihood of the response variable given explanatory variable to obtain a good estimate for $\mathbb{E}[Y \mid X = x]$. On the other hand, in causality we are interested in measuring the contrast, in this context known as the effect, in the response variable by asking: what if $X = x'$ were the case, instead of $X = x$. This seemingly subtle change of inquiry completely overhauls how we ought to think of statistics. For starters, the question requires a leap of faith: we must assume that we could observe two states for the same unit simultaneously. In other words, it is as if we have run two procedures on two literally identical units and registered their responses

Such assumption is empirically impossible to verify, known as making a counterfactual presupposition. It is tempting to question the necessity of this claim. After all, after fitting a regression line through the data, one can easily calculate the difference: $\mathbb{E}[Y \mid X = x] - \mathbb{E}[Y \mid X = x']$, by selecting the output from the adjusted line with respect to the points of interest.

Unfortunately, this line of reasoning is fallacious. By simply conditioning on observed data, we risk comparing individuals from fundamentally different populations. Consider an example where Y represents salary and X indicates whether an individual attended college. If we simply compare the expected salary of those who attended ($X = 1$) versus those who did not ($X = 0$), the observed difference is likely spurious. An individual who attends university often comes from a family with greater financial resources or an environment that incentivizes education. This is why the counterfactual reasoning is important for studying effects. More specifically we'll use potential outcomes to solve the proposed problem.

1.2 Formalization of Potential Outcomes

Suppose that we are working with an i.i.d. sample $\{(Y_i, D_i)\}_{i=1}^n$, where $D_i \in \{0, 1\}$. With no further assumptions, the best we can do is study the joint distribution $P(Y_i, D_i) = P(Y_i \mid D_i) \cdot P(D_i)$ to calculate the conditional expectation $\mathbb{E}[Y_i \mid D_i = d]$.

Definition 1 (Potential Outcomes)**.** *The potential outcomes for unit i are $Y_i(1)$ and $Y_i(0)$, where the former denotes the outcome observed if the unit is treated, and the latter denotes the outcome observed if the unit is not treated.*

Definition 2 (Individual Causal Effect)**.** *Let $Y_i(1)$ and $Y_i(0)$ denote the potential outcomes for unit i. The individual causal effect, denoted by τ_i, is defined as the difference:*

$$\tau_i := Y_i(1) - Y_i(0)$$

Key Distinction: Homogeneous vs. Heterogeneous

Homogeneous Treatment Effect: A situation in which τ_i is the same for all objects.

Heterogeneous Treatment Effect: A situation in which τ_i differs across objects.

Definition 3 (Fundamental Problem of Causal Statistics)**.** *It's impossible to observe $Y_i(1)$ and $Y_i(0)$ at the same time on the same unit, making it impossible to observe the effect of the treatment over the placebo.*

In this context, the role of D is not restricted to that of explanatory variable. Instead of an assignment mechanism that dictates the treatment for an unit, which results in the realization of a potentical outcome.

Observation (Heterogeneity): Examining effects at the individual level leads to conflicting results due to treatment effect heterogeneity. For example, suppose we have access to both potential outcomes for four units with respect to their hourly salary with and without a degree, denoted as $(Y_i(1), Y_i(0))$:

- $u_1 = (10, 5) \implies \tau_1 = 5$
- $u_2 = (20, 4) \implies \tau_2 = 16$
- $u_3 = (12, 7) \implies \tau_3 = 5$
- $u_4 = (4, 100) \implies \tau_4 = -96$

Based on the first three units, we might conjecture that going to college has a universally positive effect on salary. However, unit u_4 contradicts this pattern.

When dealing with heterogeneous effects, we average out the individual effects thorugh the expectation operator. Formally, our *Average Treatment Effect* is:

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

We emphasize the linearity of the expectation operator because, theoretically, it allows us to calculate potential outcome. By rewriting the equation as $\mathbb{E}[\tau_i] + \mathbb{E}[Y_i(0)] = \mathbb{E}[Y_i(1)]$. In this context, ATE is our causal parameter that we want to estimate. Although not explicitly stated previously, our framework relies on two assumptions aobut the stability of the system. First, we assume *No Interference*: the potential outcomes of any specific unit i are unaffected by the treatment assignments of other units. In other words, the treatment of one individual does not spill over to influence the result of another. Second, we assume *Consistency*: the observed outcome Y corresponds exactly to the potential

outcome associated with the assigned treatment D This guarantees that the treatment is well-defined and consistent across all units. Together, these conditions form what is known as the Stable Unit Treatment Value Assumption (SUTVA). In short our sample (Y_i, D_i) satifies the following property:

$$Y_i = Y_i(D_i) \quad \text{(SUTVA)}$$

instead of $Y_i = Y_i(D_1, \ldots, D_n)$.

1.3 Applying Potential Outcomes on Real Data

The dataset on which we will base our analysis for this section is the LaLonde dataset. The choice isn't arbitrary. The author intentionally designed the dataset to evaluate if experiments required the expensive randomized controlled trial, rather than the cheaper alternative: an observational study.

Key Distinction: Experimental vs. Non-Experimental

Randomized Control Trial. A study where assignment to treatment is determined solely by random chance.

Observational Study. A study utilizing existing data where assignment is determined by human choice, circumstance, or via a non-arbitrary selection.

The care shown in the design of the data itself reveals the idiosyncrasies of the causal framework, where our paradigm is present from the very beginning: from how data is collected to the final causal inference. But why is this distintion so important?