Applied Causal Inference Powered by ML and AI

Victor Chernozhukov* Christian Hansen† Nathan Kallus‡
Martin Spindler§ Vasilis Syrgkanis¶

February 5, 2026

Publisher: Online
Version 0.1.2

* MIT
† Chicago Booth
‡ Cornell University
§ Hamburg University
¶ Stanford University

0 Sneak Peek: Powering Causal Inference with ML and AI

A primary question we will be concerned with in this book is: What is the *causal effect* of an action on an outcome? For example, we may want to know what the effect of setting a product's price is on the volume of its sales.[1] To consider this question we scraped data on toy cars from Amazon.com. Figure 0.1 shows a log–log scatter plot of the 30-day average price at which each car was offered and the reciprocal of its sales rank, a publicly available surrogate for sales volume.[2] We let D denote the log of the price and Y the negative log of the sales rank of a toy car randomly drawn from the population of toy cars sold on Amazon.com. We will use this running example to preview how the book's pieces interlock, and why modern ML and AI turn causal inference from a fragile craft into a disciplined toolkit.

1: This effect may be referred to as the price *elasticity* of demand for the product.

2: Were the reader to do such an analysis using internal company data they would use actual sales volumes.

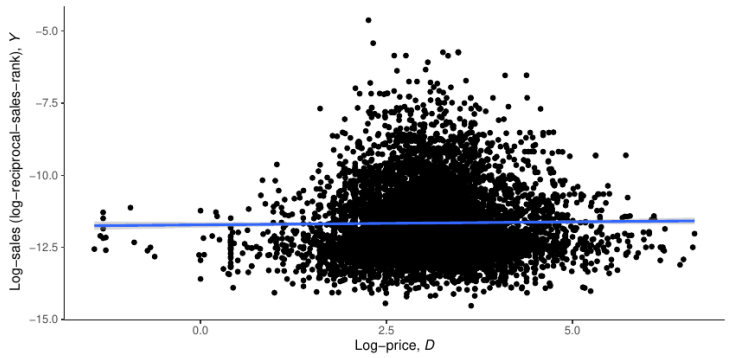

Figure 0.1: Log-prices and log-reciprocal-sales-rank of toy cars on Amazon.com along with a linear fit.

If we throw ordinary least squares at Figure 0.1—the way we do in Chapter 1—it whispers something unsettling: the fitted slope is so close to zero that one cannot even rule out a slightly positive relationship between price and sales. That is a classic false lead. If we let ourselves believe it, we would conclude that raising the price of a toy car barely changes its sales volume, or even increases it. Yet basic economic logic insists that the unobserved *potential* log-sales $Y(d)$ of a given product should *decrease* when we raise its log-price d.

The tension here is the tension that animates the whole book: prediction is easy to validate, but causality is treacherous. Prices are not assigned by a benevolent randomized experimenter; they are chosen in the market, and those choices are entangled with latent forces that also shape demand. Branding, licensing, visibility, product quality, and a hundred other factors move

both the price tag and the sales rank. When those factors are left in the shadows, naive regression makes the wrong suspect look innocent.

To bring the right suspects into the light, we build the causal vocabulary first. In Chapter **??** we introduce potential outcomes and the logic of randomized variation; in Chapter 5 we formalize confounding and what it means to "control" for it; in Chapter 6 we step into structural equations so that the word *intervention* stops being a metaphor; and in Chapter 7 we widen the frame to systems and nonlinear structure. In a simple linear structural equation,

$$Y(d) = \alpha d + U, \tag{0.0.1}$$

the parameter α is the causal effect: the change in Y produced by changing d while holding everything else fixed. That effect is generally not recovered from regressing the observed outcome on the observed treatment, because the observed price is selected by the market and is plausibly related to the unobserved idiosyncrasy U.

The key modeling move is to articulate what would make α learnable. If we observe a set of confounders W rich enough to account for all confounding, then the observed data obey a partially linear regression,

$$Y = \alpha D + g(W) + \varepsilon, \qquad \mathbb{E}[\varepsilon \mid D, W] = 0, \tag{0.0.2}$$

for some (possibly complicated) function g. At that point, the causal problem becomes an inferential problem: we must estimate α reliably even when g is high-dimensional, nonlinear, and learned by machine learning.

This is where the first twist arrives. If W is modest and the model is linear, classical tools work. But modern datasets rarely show mercy: W can be thousands of features, and the best predictors are nonlinear. In Chapter 3 we introduce regularization that stabilizes prediction in high dimensions; in Chapter 4 we show how to repair the bias that regularization injects into coefficients; in Chapter 8 we unleash nonlinear learners—trees, ensembles, neural nets—that predict superbly but refuse to hand us an interpretable slope; and in Chapter 9 we resolve the impasse by developing *orthogonalization* and *double machine learning* (DML). The idea is to learn the nuisance parts well enough to remove them, and to do it in a way that preserves valid inference for the low-dimensional causal target.

Now the plot thickens again. The scatter plot in Figure 0.1 is

a single snapshot; the market, of course, is a moving picture. In the demand-analysis study that accompanies this chapter, we follow $N = 7{,}226$ toy-car products over $T = 12$ adjacent four-week periods from March 2023 to January 2024, tracking prices and sales ranks over time.[3] Once you watch the series unfold, the data reveal a signature pattern: prices move in *sticky* steps, while sales ranks jump and surge like a seismograph. Figure 0.2 shows two examples.

3: The empirical study uses the log inverse sales rank as a proxy for log quantity and the log price as a price signal.

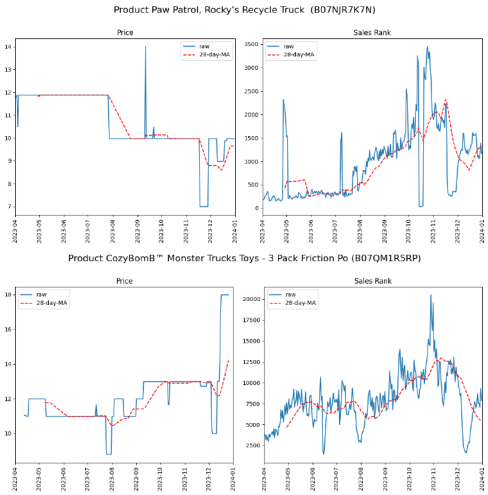

Figure 0.2: Price and sales-rank series for two example products. Prices typically evolve in piecewise-constant steps, while ranks react quickly.

This temporal structure changes what "confounding" looks like. A product's past sales are not just a lagged outcome; they are a proxy for visibility and latent quality, and they can drive both future prices and future demand. The demand study therefore uses a simple dynamic causal model in which the relevant state includes lagged quantity and price along with product characteristics. Conditioning on that state turns the causal problem back into a regression problem, and DML again provides the inferential engine.

If the market is the crime scene, then the product page is the dossier. Every toy car comes with rich evidence: text descriptions, images, and tabular facts. Figure 0.3 shows one such page.

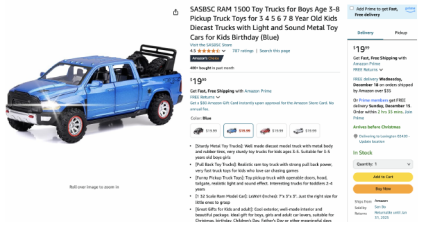

Figure 0.3: A typical product "dossier" consists of an image, a text description, and structured attributes.

The catch is that none of this evidence arrives in the neat columns that regression likes. We therefore will teach the practical art

of turning rich modalities into numerical representations that a causal pipeline can use without becoming a house of cards. We start from transformer encoders: models trained by self-supervision to read text, and vision transformers trained to see images. We then show how these encoders can be connected to downstream prediction heads, and how *fine-tuning* can align the representation with the nuisance regressions that DML needs. In the demand study, this means fine-tuning multimodal embeddings to predict both price and quantity signals, because those predictions are exactly what we must partial out to isolate price shocks.

When we project the resulting embeddings into low dimensions, they organize products into coherent neighborhoods: clusters that separate by visual style, branding cues, and latent "type." Figure 0.4 gives a glimpse. The point is not the picture itself; the point is that the representation has learned enough structure that it can serve as a control, an effect modifier, or both.

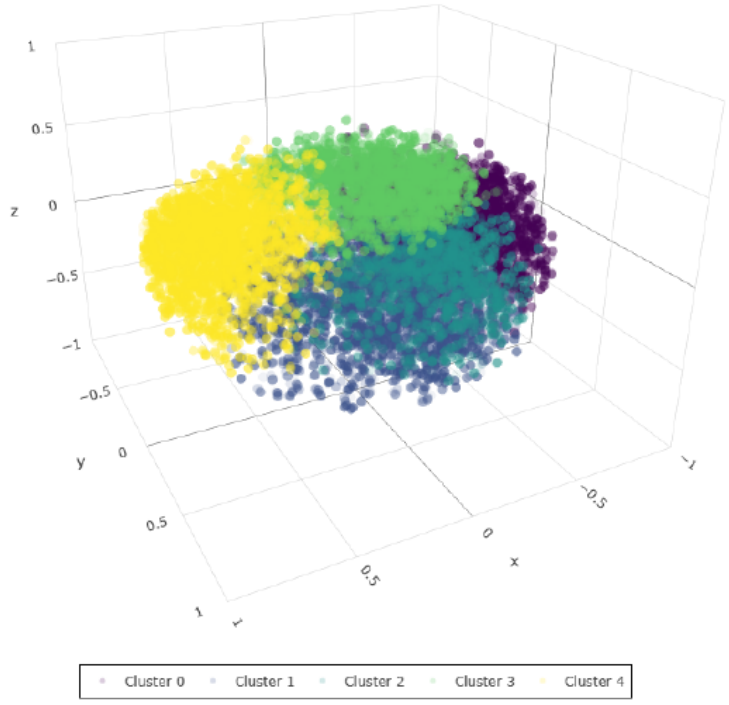

Figure 0.4: A 3D projection of multimodal product embeddings (text+image), colored by a simple clustering. Even in three dimensions, the representation separates latent product "types" that matter for demand.

At this stage, the big win from embeddings is often expected to be "better confounding control." The demand study delivers a subtler and more exciting twist: embeddings matter most as *effect modifiers*. Average price elasticity is informative, but it is also a blunt instrument; some products are almost immune to small price changes, while others are exquisitely sensitive. Once we let elasticity vary with product characteristics, the model uncovers a wide spread of price sensitivities that is not

statistical noise but systematic, predictable structure.

Figure 0.5 shows the estimated elasticity function sorted across products, together with confidence bands. The line climbs from strongly negative effects to near zero, and in some specifications even positive values for the least price-sensitive products, while the average effect sits in the middle like a red dashed "alibi." This is the moment when the case file turns into a story about *who* responds to price, not merely whether price matters.[4]

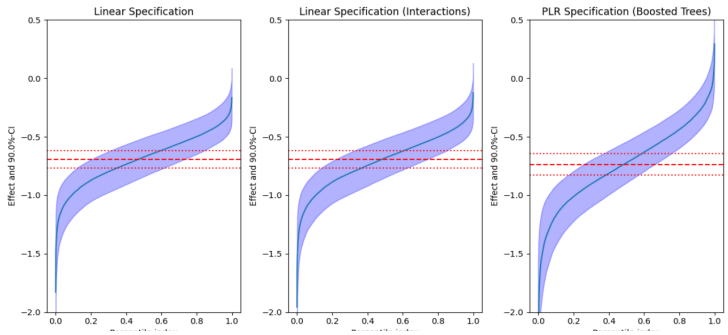

Figure 0.5: Sorted elasticity estimates as a function of effect modifiers, with pointwise confidence bands, under several specifications.

This book is built around that kind of turn: we begin with simple estimands, then show how the world tries to fool them, and then show how to defend them without giving up flexibility. The "BERT chapter"—Chapter 10—is therefore not about memorizing a brand name; it is about learning the encoder–decoder logic that turns text, images, and other unstructured data into usable features, and then learning how to plug those features into orthogonal causal estimators. Along the way we connect BERT-style masked-language modeling to modern transformer encoders, connect vision transformers to image embeddings, and show how fine-tuning for prediction can be made compatible with valid inference for causal targets.

If you want the story to keep escalating after that, the Advanced Topics do exactly that. In Chapter 12 and Chapter 13 we confront the cases where we do *not* believe we have measured all confounders, and we develop tools that exploit extra structure—sensitivity analysis, instruments, proxy controls—without surrendering the ability to use rich AI features. In Chapter 14 and Chapter 15 we move from "the average effect" to heterogeneous effects and personalization, and in Chapter 16 and Chapter 17 we show how DML can be fused with difference-in-differences and regression discontinuity designs. Finally, in Chapter 7 and Chapter 11 we explain why directed acyclic graphs are the right map for all of these journeys: they tell you, before you estimate anything, which arrows you are betting your credibility on.

4: The full story with many more background information can be looked up in the paper " Adventures in Demand Analysis Using AI" by Bach et. al. (2026).

We close the preview by returning to the guiding promise. It is impossible to definitively validate a causal claim from observational data alone, because identification rests on assumptions. But we can make those assumptions explicit, we can use the richest available information to make them more plausible, and we can avoid needless functional-form commitments. DML on top of modern encoders and decoders gives us a rare combination: the power of AI representations and the discipline of inferential guarantees. If you read the book, you will not only learn how to fit models that look impressive, but how to extract conclusions that can survive cross-examination.