Applied Causal Inference Powered by ML and AI

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

February 28, 2024

Publisher: Online
Version 0.1.1

* 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 9,212 toy cars from Amazon.com. Figure 0.1 shows a log-log-scale scatter plot of the 30-day average price at which each 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 example to preview the book's chapters and how they come together to enable the reader to power applied causal inference on modern datasets using ML and AI.

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.

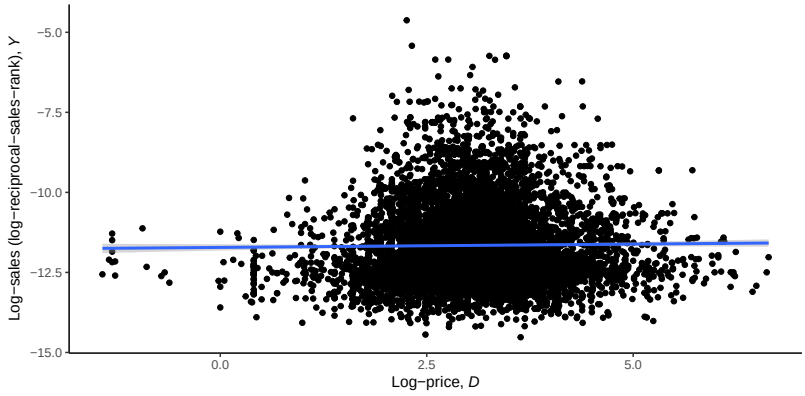

Figure 0.1: Log-prices and log-reciprocal-sales-rank of 9,212 toy cars on Amazon.com along with a linear fit.

In Chapter 1, we present linear regression by ordinary least squares (OLS), which can help us understand the relationship between these two variables. Here it suggests that a unit increase in D is associated with anything between a -0.008 and a 0.050 unit change in Y on average over toy cars; that is, $(-0.008, 0.050)$ is the 95% confidence interval on the slope of the best linear predictor. In words, it suggests one cannot rule out small negative or even slightly positive association between price and sales. It would be incorrect, however, to infer that arbitrarily increasing the price on any one toy car would cause almost no effect on its sales volume, or even increase it.

Instead, economic theory would suggest that the unobserved *potential* log-sales, $Y(d)$, of any *one* toy car should in fact *decrease* as the log-price that one sets, d, increases. In Chapter 2, we

present this notion of potential outcomes and study inference on their averages when actions are *randomized* (or, *exogenous*). For example, we may be interested in the average sales if price were set to a certain level. Unlike the randomized controlled trial (RCT) setting discussed in that chapter, here prices are *not* actually set at random; that is, prices are *endogenous*. Thus, the reason we may see no or a slightly positive association is *confounding* factors that affect both the potential sales at any one price and the particular price that is set. For example, whether a toy car is produced by a brand name or incorporates characters from a popular TV show might increase sales at any one price as well as lead the seller to choose a higher price, whether in anticipation of higher demand or because of higher production or licensing costs.

We formalize this notion of confounding in Chapter 5 and consider causal inference on averages of potential outcomes when one observes *all* confounding variables, W. In Chapter 6, we go on to consider a *linear* structural equation,

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

which posits that, on average, log-sales at any one log-price is a linear function of the log-price, aside from the idiosyncrasies U of each one toy car. Within this structural equation, we interpret α as the causal effect of d on Y; that is, the effect of a change in d on Y produced by intervening in the system to change d while holding all other determinants of sales constant. This causal effect is generally not recovered from regression of observed Y on observed price, D, as observed price is set in the market and plausibly related to unobserved factors U.

In our simple linear structural equation, the assumption that W accounts for all confounding leads us to conclude that we have

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

for some function $g(W)$. Thus, after all of our causal modeling and assumptions, what remains is inference on a coefficient in a possibly complex regression model of Y on D and W, all of them *observed* variables. That is, under our causal modeling and assumptions, making statistical inferences (such as constructing estimates and confidence intervals) on α in Eq. (0.0.2) from data on (Y, D, W) would be *causal inference*. (0.0.1) is the simplest of structural equations – to understand more complex structures we consider *systems* of equations, and in Chapter 7 even *nonlinear* structural equations.

To explain how we power such causal inference with ML and AI, let us now return to the question of *what is W in the first place*? There are many features we can observe about each toy car on Amazon.com in addition to its price and sales: all the text on the product page such as name and description, the product subcategory (beyond being a toy), the brand, the color, and the dimensions and weight of both the item and its packaging. What features can we make use of, and how?

Classical methods, like OLS, (Chapter 1) allow us to conduct inference on α when Eq. (0.0.2) is a linear regression with moderately high dimensions, that is, when W is a p-dimensional random vector, $g(W) = \beta_1 + \beta_2' W$, and p is much smaller than the number of observations we have (here, 9,212). Letting $g(W) = \beta_1 + \beta_2' W$ in Eq. (0.0.2) we obtain a *linear model*. There are 243 product subcategories for our toy cars. Consider identifying each with a number in $1, \ldots, 243$ and letting W be a 243-dimensional vector with a 1 in the index corresponding to the product's subcategory and 0 elsewhere. OLS regression of Y on D and this particular W explains 7.5% of Y's variance (as measured by adjusted R^2) and gives a 95% confidence interval on α of $(-0.026, 0.036)$. These results are not very different from what we inferred in the observed association between Y and D without adjusting for any confounding effects, but at least the upper bound is smaller – we indeed do not believe a positive effect is realistic.

Perhaps we need to control for more confounding effects than just subcategory membership. However, even without departing from linearity, OLS no longer provides reliable inference if we include too many features in W. Letting $g(W) = \beta_1 + \beta_2' W$ in Eq. (0.0.2) with a high-dimensional W, that is, where d is comparable to or bigger than the number of observations, we obtain a *linear model with high-dimensional controls*. In Chapter 3, we present more advanced ML methods than OLS: predictive inference in high dimensions using regularized linear regression. The use of regularized linear regression may improve prediction relative to OLS but introduces biases that imperil inference on coefficients. In Chapter 4, we show how to remedy this bias when making inferences on any one coefficient. In the context of causal inference, this setup allows us to potentially handle very many confounders, and the hope is that we can then more reliably justify having accounted for *all* confounders. In a nutshell, in the setting of Eq. (0.0.2), if we take $\tilde{Y}$ and $\tilde{D}$ to be the *residuals* from a modern high-dimensional linear regression of Y on $(1, W)$ and of D on $(1, W)$, respectively, then OLS regression of $\tilde{Y}$ on $\tilde{D}$ yields valid inference on α even when

W is high-dimensional.

Consider letting W be a 11546 dimensional vector including not only the indicator of subcategory but also the item's physical dimensions, transformed by log and expanded up to third power of the logarithms, missingness indicators, the interaction of these dimension features with subcategory, the indicator of brand (among 1827 brands). In this case, p is greater than the number of observations n. Using the methods we present in Chapter 4 to leverage this high-dimensional W in this particular set up, we obtain a 95% confidence interval on α of $(-0.10, -0.029)$. The confidence interval including only negative values is in concordance with the intuition that intervening to increase price would decrease demand. At the same time, we may still worry that a linear model is too restrictive, in essence allowing us only to control linearly for pre-specified confounders.[3]

In Chapter 9, we present nonlinear ML methods for regression: trees, ensembles, and neural nets. Compared to predicting log-price and log-sales with LASSO, using these methods (with a 2083-dimensional feature vector omitting the expansions and interactions needed for linear models) increases the R^2 by 25-53% and 89-189% (evaluated using 5-fold cross-validated R^2). Clearly these methods offer significant predictive improvements in this dataset. However, such nonlinear methods have no clear parameter to extract, no coefficient to inspect. While making excellent predictions, it is not immediately clear how to use them to make valid statistical inferences on finite-dimensional parameters, like average effects. We tackle that question in Chapter 10. Letting $g(W)$ be an arbitrary nonlinear function in Eq. (0.0.2) gives rise to what is called the *partially linear model*, which strikes a nice balance between structure and flexibility: the causal-effect part of the model is simple and interpretable – for each unit increase in action we get α increase in outcome – while the confounding part, which we have no interest in interpreting, can be almost-arbitrarily complex.[4] In the setting of Eq. (0.0.2), it turns out we can keep the method of residual-on-residual OLS inference, but using residuals from advanced *nonlinear* regressions, as long as we fit these regressions on parts of the data that exclude where we use them to make predictions and produce the residuals. This is *double machine learning* or *debiased machine learning* or *double/debiased machine learning*[5] for the partially linear model. Using DML together with gradient-boosted-tree regression to make inferences on the price elasticity α in this example yields a confidence interval of $(-0.139, -0.074)$, suggesting an effect whose direction agrees

3: One may include *pre-specified transformations* of confounders as well as discussed in Chapter 1.

4: Luckily, even if the partially linear assumption fails, estimates still reflect some average of the *causal* effects of increasing *all* prices by a small amount, provided we have accounted for all confounding effects in in W. See Remarks 10.2.2 and 10.3.3.

5: We will use these terms interchangeably and abbreviate them with *DML*.

even more strongly with our intuition, which can be attributed to these more powerful predictive methods being able to better account and correct for the confounding effects that pushed the apparent direction upward.

It is still unclear, however, whether the numeric features we observe can reliably capture all of the confounding effects – if they cannot, then no regression, no matter how flexible, can help. This problem – getting the right data to enable causal inference – is a common challenge when dealing with observational data. It is in using all the available data, where modern AI along with the tools we develop in this book come together to uniquely enable powerful causal inferences using modern observational data sets. Modern data sets are rich, containing far more than just numeric features. This data set, for example, contains text on each product – descriptions that capture many important features about each product that are not clearly tabulated but must be inferred by reading the text. Luckily, modern AI has made great inroads in recent years in machine cognition of text, images, videos, and other rich data.

In Chapter 11, we discuss how these powerful tools can be used in concert with DML. BERT is a large language model leveraging a deep learning architecture known as *transformers* and achieving impressive performance on natural-language-processing benchmarks. Using neural-net-based predictive models for log-price and log-sales built on top of BERT results in a 12-37% and 4-59% increase in cross-validated R^2, respectively, relative to the nonlinear models using only numeric features in the data. The non-numeric features in the data therefore seem to account for more than the baseline numeric factors of products in predicting price and sales. Using DML for the partially linear model together with these models that use the non-numeric features, we are able to make causal inferences that account for confounding factors reflected in the rich text on the product page for each toy car. Proceeding in this way, as we explain in greater detail in Chapter 11, we obtain a confidence interval on α of $(-0.21, -0.13)$. That we get a more negative estimate here again suggests that there were residual confounding effects inducing a spurious positive relationship between price and sales that we could only have controlled for and counteracted by using AI to account for the rich text data.

While it is relatively easy to validate predictive models' performance by using held-out test sets and cross-validation, it is difficult – impossible, even – to definitively validate a causal effect, as it will inevitably rest on fundamentally untestable assumptions. Nonetheless, we can have greater confidence in

estimates that correctly and fully leverage the available data and do not rely on unnecessary parametric assumptions. Estimates based on DML on top of AI allow us to do just that. We can use rich data without imposing strong functional form restrictions and importantly can do so without imperiling guarantees on valid statistical inference. The Core material outlines the basic ideas and provides fundamental results for using DML with AI learners to estimate and do inference for low-dimensional causal effects.

The Advanced Topics section includes chapters that expand upon the basic material from the Core chapters. In the Core material, we discuss more complex structures than the partially linear model introduced in this preview, but do inference essentially only when all relevant variables are observed. In Chapter 12, we present alternative ways to identify causal effects when we do not believe we observe all confounders – techniques such as sensitivity analysis, instrumental variables, and proxy controls, and we provide methods for causal inference in such settings in Chapter 13. These tools allow us to have confidence in causal estimates that leverage special structure like instruments or proxies without additionally making unnecessary parametric assumptions and with the ability to leverage rich data using powerful AI. In many examples, one may wish to understand heterogeneity in causal effects such as how causal effects differ across observed predictors. Chapter 14 covers DML inference on quantities that characterize this heterogeneity, and Chapter 15 goes beyond inference on low-dimensional causal parameters and discusses learning heterogeneous causal effects from rich individual-level data and even personalizing treatments based on such data. Finally, we consider application of DML in conjunction with two popular methods for identifying causal effects – difference-in-differences and regression discontinuity designs – in Chapter 16 and Chapter 17 respectively.

After studying the book, the reader should also be able to understand and employ DML in many other applications that are not explicitly covered. In the toy car example we focused on sales, but sales may not reflect demand when we reach the limits of on-hand inventory, something known as right-censoring. Censoring is an example of data coarsening, and mathematically it is not too dissimilar from the missingness of potential outcomes for actions not taken. Similarly, we may want to look at distributional effects beyond averages, like effects on the quantiles of sales. DML can often be applied to these problems and there is active research on applying it to ever more intricate problems.

There are also topics beyond our scope. We started by saying we focus on the causal effect of an action on an outcome – a broader yet much more challenging question is, among multiple variables, discovering which have causal effects on which. While we *do* discuss the use of directed acyclic graphs in Chapter 7 and Chapter 8, we only use them to represent assumed structure and only briefly mention how one might try to learn causal structure directly from data, which is the subject of *causal discovery*.

Our aim is rather focused: present the building blocks of predictive inference and of causal inference and illustrate their effective and correct use in concert in a way that allows readers to employ them in real, practical settings. The book interweaves the two kinds of inference, with many real-data examples with code notebooks. We hope the outcome is that we reach an endpoint where the reader is ready to power causal inferences with ML and AI and be able to draw valid, reliable inferences in practice using rich modern data.