Causal inference in economics and marketing

Hal R. Varian[a,1]

[a]Economics Team, Google, Inc., Mountain View, CA 94043

Edited by Richard M. Shiffrin, Indiana University, Bloomington, IN, and approved May 25, 2016 (received for review May 28, 2015)

This is an elementary introduction to causal inference in economics written for readers familiar with machine learning methods. The critical step in any causal analysis is estimating the counterfactual—a prediction of what would have happened in the absence of the treatment. The powerful techniques used in machine learning may be useful for developing better estimates of the counterfactual, potentially improving causal inference.

causal inference | economics | machine learning | marketing

Suppose you are given some data on ad spend and product sales in various cities and are asked to predict how sales would respond to a contemplated change in ad spend. If y_c denotes per capita sales in city c and x_c denotes per capita ad spend in city c, it is tempting to run a regression of the form $y_c = bx_c + e_c$, where e_c is an error term and b is the coefficient of interest. (We assume all data have been centered; therefore, we can ignore the constant in the regression.) The machine-learning textbook by James et al. describes a problem of this sort (ref. 1, p. 59).

Unfortunately, such a regression is unlikely to provide a satisfactory estimate of the "causal" effect of ad spend on sales. To see why, suppose that the sales, y_c, are per capita box office receipts for a movie about surfing and x_c are per capita television ads for that movie. There are only two cities in the dataset: Honolulu, Hawaii and Fargo, North Dakota.

Suppose that the dataset indicate that the advertiser spent 10 cents per capita on television advertising in Fargo and observed $1 in sales per capita, whereas in Honolulu, the advertiser spent $1 per capita and observed $10 in sales per capita. Hence, the model $y_c = 10x_c$ fits the data perfectly.

However, here is the critical question: Do you really believe that increasing per capita spend in Fargo to $1 would result in box office sales of $10 per capita? For a surfing movie? This outcome seems unlikely, so what is wrong with our regression model?

A Motivating Problem

The problem is that there is an omitted variable in our regression, which we may call "interest in surfing." Interest in surfing is high in Honolulu and low in Fargo. What is more, the marketing executives that determine ad spend presumably know this, and they choose to advertise more where interest is high and less where it is low. Therefore, this omitted variable—interest in surfing—affects both y_c and x_c. Such a variable is called a "confounding variable."

To express this point mathematically, think of (y, x, e) as being the population analogs of the sample (y_c, x_c, e_c). The regression coefficient is given by $b = \mathrm{cov}(x,y)/\mathrm{cov}(x,x)$. Substituting $y = bx + e$, we have

$$b = \mathrm{cov}(x, xb + e)/\mathrm{cov}(x,x) = b + \mathrm{cov}(x,e)/\mathrm{cov}(x,x).$$

The regression coefficient will be unbiased when $\mathrm{cov}(x,e) = 0$.*

If we are primarily interested in predicting sales as a function of spend, and the advertiser's behavior remains constant, the simple regression described in ref. 1 may be just fine. However, usually a prediction of past behavior is not the goal; what we want to know is how box office receipts would respond to a change in the advertiser's behavior.

To put it slightly more formally: we have historical observations that were generated by a process such as "choose spend based on factors you think are important," and we want to predict what would happen if we switch to a data generating process such as "increase your spend everywhere by some amount."

It is important to understand that the problem is not simply that there is a missing variable in the regression. There are always missing variables—that is what the error term represents. The problem is that the missing variable, "interest in surfing," affects both the outcome (sales) and the predictor (ads); therefore, the simple regression of sales on ads will not give us a good estimate of the causal effect: what would happen to sales if we explicitly intervened and changed ad expenditure across the board.

This problem comes up all of the time in statistical analysis of human behavior. In our example, the amount of advertising in a city, x_c, is chosen by some decision makers who likely have some views about how various factors affect outcomes, y_c. However, the analyst is not able to observe these factors—they are part of the error term, e_c. It is therefore unlikely that x_c and e_c are uncorrelated. In our example, cities with high interest in surfing may have high ad expenditure and high box office receipts, meaning a simple regression of y_c on x_c would overestimate the effect of ad expenditure on sales.†

In this simple example, we have described a particular confounding variable. However, in realistic cases, there will be many confounding variables—variables that affect both the outcome and the variables we are contemplating changing.

Everyone knows that adding an extra predictor to a regression will typically change the values of the estimated coefficients on the other predictors because the relevant predictors are generally correlated with each other. Despite this well-known phenomenon, many analysts seem comfortable in assuming that the predictors we do not observe—those in the error term—are magically orthogonal to the predictors we do observe.

The "ideal" data, from the viewpoint of the analyst, would be data from an incompetent advertiser who allocated expenditures randomly across cities. If ad expenditure is truly random, then we do not have to worry about confounding variables because the predictors will automatically be orthogonal to the error term. However, statisticians are seldom lucky enough to have a totally incompetent client.

There are many other examples of confounding variables in economics. Here are a few classic examples.

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Drawing Causal Inference from Big Data," held March 26–27, 2015, at the National Academies of Sciences in Washington, DC. The complete program and video recordings of most presentations are available on the NAS website at www.nasonline.org/Big-data.

Author contributions: H.R.V. wrote the paper.

Conflict of interest statement: H.R.V. is a full-time employee of Google, a private company.

This article is a PNAS Direct Submission.

[1]Email: hal@sims.berkeley.edu.

*Note that the problem is not inherently statistical in nature. Suppose that there is no error term, so that the model "revenue = spend + interest in surfing" fits exactly. If we only look at the variation in spend and ignore the variation in surfing interest, we get a misleading estimate of the relationship between spend and revenue.

†It would not have to be that way. Perhaps surfing is so popular in Honolulu that everyone already knows about the movie, and it is pointless to advertise it. Again, this is the sort of thing the advertiser might know but the analyst does not.

www.pnas.org/cgi/doi/10.1073/pnas.1510479113

How Does Fertilizer Affect Crop Yields? If farmers apply more fertilizer to more fertile land, then more fertilizer will be associated with higher yields and a simple regression of fertilizer on outcomes will not give the true causal effect.

How Does Education Affect Income? Those who have more education tend to have higher incomes, but that does not mean that education caused those higher incomes. Those who have wealthy parents or high ability tend to acquire both more education and more income. Hence, simple regressions of education on income tend to overstate the impact of education. (See ref. 1, p. 283, for a machine-learning approach to this problem and ref. 2 for an econometric approach.)

How Does Healthcare Affect Income? Those who have good jobs tend to have health care; therefore, a regression of health care on income will show a positive effect but the direction of the causality is unclear.

In each of these cases, we may contemplate some intervention that will change behavior.

How would crop yields change if we change the amount of fertilizer applied? How would income change if we reduce the cost of acquiring education? How would income change if we changed the availability of health care?

Each of these policies is asking what happens to some output if we change an input and hold other factors constant. However, the data were generated by parties who were aware of those other factors and made choices based on their perceptions. We want an answer to a *ceteris paribus* question, but our data were generated *mutatis mutandis*.

In the next section, we describe the gold standard for estimating causal effects: controlled experiments. Controlled experiments are not always feasible; therefore, in the following sections, we examine four techniques that have been used in economics that sometimes enable identification of causal effects with observational data. These methods are (*i*) natural experiments, (*ii*) instrumental variables, (*iii*) regression discontinuity, and (*iv*) difference in differences.

Controlled Experiments

As Box et al. (3) put it, "To find out what happens when you change something, it is necessary to change it." As we will see, that may be slightly overstated, but the general principle is right: the best way to answer causal questions is usually to run an experiment.

However, experiments are often costly and in many important cases are not feasible. Consider the example of the impact of education on income. An ideal experiment would require randomly selecting the amount of education students acquire, which would be rather difficult.

However, this is an extreme case. Actual education policies being contemplated might involve things like student loans or scholarships and small scale experiments with such policies may well be feasible. Furthermore, there may be "natural experiments" where there is some natural randomization that is that can shed light on such issues without requiring explicit intervention.

In a classic clinical experiment, one applies a "treatment" to some set of "subjects" and observes some "outcomes." The outcomes for the treated subjects can be compared with the outcomes for the untreated subjects (the control group) to determine the causal effect of the treatment on the subjects. In effect, an experiment is simply a small scale version of a policy that you are contemplating implementing.

One may be interested in the "impact of the treatment on the population"; in which case, one would like the experimental subjects to be a representative sample from the population. Additionally, one might be interested in how the treatment affected those who were actually treated but were not necessarily randomly chosen for treatment. This is the case of "impact of the treatment on the treated." Furthermore, one might be interested in those who were invited to be treated, whether or not they actually agreed to be treated; this is called an "intention to treat" analysis.

When we are interested in the "impact of treatment on the treated" and controlled experimentation is not feasible, there are two modeling approaches. In one case, "selection on observables," the researcher attempts to build a predictive model of who received treatment. In the other case, "selection on unobservables," the research attempts to find natural experiments that are "as good as random" and can overcome the confounding variable problem described earlier. Our focus is on the latter case, which is common in economic examples. Both approaches are carefully described and compared in ref. 4.

If the proposed policy is going to be applied universally to some population, then one is likely interested in the impact of the treatment on the population. If the proposed policy to be implemented involves voluntary participation, then one may be interested in the impact of the treatment on those who choose (or agree) to be treated.

Basic Identity of Causal Inference

Following ref. 5, we can decompose the observed outcome of a treatment into two effects:

$$\begin{aligned}&\text{Outcome for treated} - \text{Outcome for untreated}\\ &= [\text{Outcome for treated} - \text{Outcome for treated if not treated}]\\ &\quad + [\text{Outcome for treated if not treated}\\ &\qquad - \text{Outcome for untreated}]\\ &= \text{Impact of treatment on treated} + \text{selection bias.}\end{aligned}$$

This "basic identity of causal inference" shows that the critical concept for understanding causality is the comparison of the actual outcome (what happens to the treated) compared with the counterfactual (what would have happened if they had not been treated), an insight that goes back to refs. 6 and 7. As Rubin (7) emphasized, we cannot actually observe what would have happened to the treated group if the individuals had not been treated; therefore, we have to estimate that counterfactual some other way.

The basic identity nicely shows why randomized trials are the gold standard for causal inference. If the treated group is a random sample of the population, then the first term is an estimate of the causal impact of the treatment on the population, and if the assignment is random, then the second term has an expected value of zero.

Impact of an Ad Campaign

Angrist and Pischke (8) describe what they call the "Furious Five methods of causal inference": random assignment, regression, instrumental variables, regression discontinuity, and differences in differences. We will give a brief introduction to these methods in the next few sections, although we organize the topics slightly differently.

In marketing, we may be concerned with the impact of an ad exposure on a consumer. In this case, the classic experimental treatment-control framework described earlier can be applied, where we estimate the counterfactual (no ad exposure) using a control group.

Another approach is to estimate impact of an ad treatment on the advertiser. For example, an advertiser might ask "if I increase my ad expenditure by some amount, how many extra sales do I generate?" Of course, the answer depends on how the consumers respond to the ad, but we do not necessarily have to model that in detail to answer this question. Instead, the advertiser can simply increase spend for a limited period, and we can compare the outcome of that experiment to an estimate of the counterfactual—what would have happened during the limited period without that increase in spend.

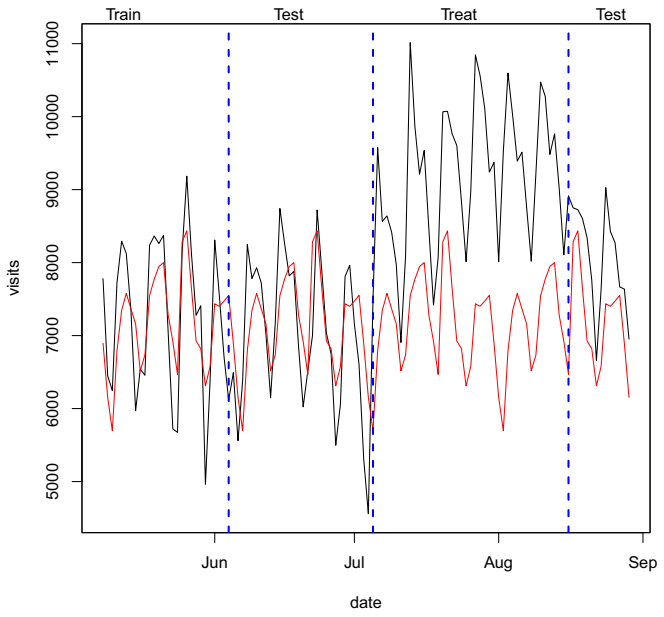

Fig. 1. Hypothetical TTTC process. The model is estimated during the training period and its predictive performance is assessed during the test period. The extrapolation of the model during the treat period (red line) serves as a counterfactual. This counterfactual is compared with the actual outcome (black line), and the difference is the estimated treatment effect. When the treatment is ended, the outcome returns to something close to the original level.

However, where does the counterfactual come from? Answer: it is a predictive model developed using data from before the experiment was run.

In the classic experiment design described earlier, we compare treated and untreated subjects. Here, we treat all of the subjects for a limited time and measure their aggregate response. Our counterfactual is a prediction of what would have happened during the limited period of spend increase.

The classic design is appropriate when one is primarily interested in how the treatment affects the subjects. For example, it is important to know whether some observed change in health is attributable to a drug effect or a placebo effect. However, in many marketing experiments, the primary interest is often in how the treatment of the subjects affects the "experimenter." The experimenter/advertiser may be interested simply in how much visits increase; whether the increase is attributable to ad clicks, search clicks, or direct navigation may be of secondary importance.

This predictive model could depend on behavior of a control group, if such a group is available, but this is not the only way to develop such a prediction. Brodersen et al. (9) describe a case where the outcome of interest was visits to a particular website and the treatment was ad spend. It turns out that the visits to this website could be well predicted by the number of "searches" about topics related to the subject matter of the website; therefore, counts of these searches could contribute to the construction of the predictive model.

In building the predictive model, we can use standard machine learning tools such as cross-validation to tune parameters. Once we are satisfied with our model, we can apply it to a test set to determine how well it performs. We can then apply the model during the treatment period to predict the counterfactual and compare what actually happened to the treated to the prediction of our model of what would have happened without the treatment. This train-test-treat-compare (TTTC) process is illustrated in Fig. 1.

TTTC is a generalization of the classic treatment–control approach to experimentation. In that model, the control group provides an estimate of the counterfactual, which is the gold standard for causal inference. However, even if we do not have a true control group, we still may be able to develop a predictive model of the counterfactual using other methods.

Note that approach gives us the "impact of the treatment on the treated" because we are interested in the impact of the spend change on this particular advertiser, not on advertisers in general.

The TTTC cycle that I have outlined is similar to the synthetic control method described by Abadie et al. (10). (See also the time-series literature on interrupted regression, intervention analysis, structural change detection, etc.) Synthetic control methods use a particular way to build a predictive model of to-be-treated subjects based on a convex combination of other subjects outcomes. However, in principle, other modeling techniques could be used to develop predictions of the counterfactual.

One important caveat about building the predictive model: we do not want to use predictors that may be affected by the treatment, otherwise we run into the confounding variable problem described earlier. For example, during the "Holiday Season," we commonly observe both an increase in ad spend and an increase in sales. Therefore, the Holiday Season is a confounding variable, and a simple regression of spend on sales would give a misleading estimate. The solution here is simple: pull the confounder out of the error term and model the Holiday period as an additional predictor.

We have seen that causal inference involves comparing actual outcomes to counterfactual outcomes. The standard approach is typically a cross-section model to compare treated subjects to untreated subjects. In this case, the counterfactual is a prediction of the outcome for those treated if they had not been treated, which is typically based on the outcome for the control group (sometimes with an adjustment for other factors).

As the above example illustrates, one can also examine a single subject before, after, and during treatment. In this case, the counterfactual is the forecast of the outcome for the subject constructed using data from before the experiment. To implement this approach, one would normally build a model using time series methods such as trend, seasonal effects, autocorrelation, persistence of treatment effect, and so on.

Regression Discontinuity

As I indicated earlier, it is important to understand the data-generating process when trying to develop a model of who was selected for the treatment. One common selection rule is to use a threshold. In this case, observations close to, but just below, a threshold should be similar those close to, but just above, the threshold. Therefore, if we are interested in the causal effect of passing the threshold, comparing subjects close to the threshold but on different sides is appealing.

For example (11), observe that in Israel, class sizes for elementary school students that have 40 students enrolled on the first day remain at that size throughout the year. However, classes with 41 or more students have to be divided in half, or as close to that as possible. This allows them to compare student performance in classes with 40 initial students to that with (say) 41 initial students (who end up with 20-person classes), thereby teasing out the causal effect of class size on educational performance. Because it is essentially random which side of the threshold a particular subject ends up on, this administrative policy is almost as good as random assignment to different sized classes.[‡]

Another nice example is the study by Valletti et al. (12) that aims to estimate the impact of broadband speed on housing values. Just looking at the observational data will not resolve this

[‡]The actual policies used are a bit more complicated than I have described; see the cited source or ref. 5 for a more detailed description.

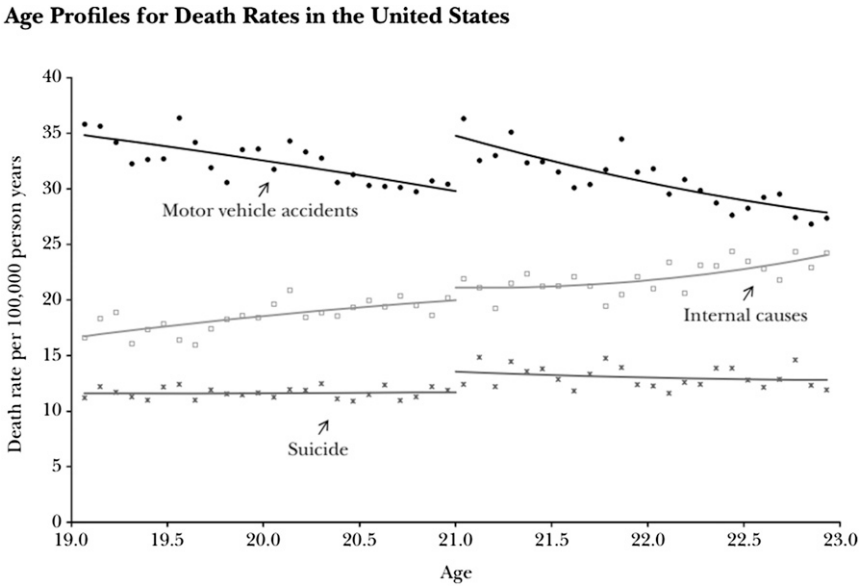

Fig. 2. Death rates by age by type of death. Reprinted with permission from ref. 13.

issue because houses in newer areas may be both more expensive and have better broadband connections. However, looking at houses that are just on the boundary of internet service areas allows one to identify the causal effect of broadband on house valuation.

As a final example, consider Carpenter and Dobkin (13), who examine the impact of the minimum legal drinking age on mortality. The story is told in Fig. 2, which is taken from this paper (see also the helpful discussion in ref. 8).

As you can see, there is a major jump in motor vehicle accidents at the age of 21. Someone who is 20.5 y old is not that different from someone who is 21 y old, on average, but 21-y-olds have much higher death rates from automobile accidents, suggesting that the minimum drinking age causes this effect. For example, suppose you were asked to determine what would happen to the mortality–age relationship if the drinking age were raised to 22 y old. Once you have seen Fig. 2, it is not hard to come up with a reasonable answer.

Regression discontinuity design is very attractive when algorithms are used to make a choice of treatment. For example, ads may receive some special treatment such as appearing in a prominent position if they have a score that exceeds some threshold. We can then compare ads that just missed the threshold to those that just passed the threshold to determine the casual effect of the treatment. Effectively, the counterfactual for the treated ads are the ads that just missed being treated. See ref. 14 for an example in the context of ranking search ads.

Even better, we might explicitly randomize the algorithm. Instead of a statement like "if (score > threshold), do treatment" we have a statement like "if (score + e > threshold), do treatment," where e is a small random number. This explicit randomization allows us to estimate the causal effect of the treatment on outcomes of interest. Note that restricting e to be small means that our experiment will not be very costly compared with the status quo because only cases close to the threshold are impacted.

Of course, software engineers at companies like Google, Microsoft, and Facebook write code that implements this sort of experimentation. However, this requires a lot of discipline: engineers have to recognize in advance that they will want to do experiments when the code is productionized. Writing such code is harder than it should be; it would be nice to have libraries and tools that would make it easy to generate learning code. Even better would be a way to automatically modify legacy code to become "learning code." For examples of such tools, see refs. 15 and 16.

Natural Experiments

If there is a threshold involved in making a decision, we have seen that by focusing on those cases close to the threshold, we may have a procedure that is "almost as good" as random assignment to treatment and control. However, we may be able to find a "natural experiment" that is "as good as random."

Consider, for example, the Super Bowl. It is well known that the home cities of the teams that are playing have an audience about 10–15% larger than cities not associated with the teams playing. It is also well known that companies that advertise during the Super Bowl have to purchase their ads months before it is known which teams will actually be playing. The combination of these two facts implies that two essentially randomly chosen cities will experience a 10% increase in ad impressions for the movie titles shown during the Super Bowl. If the ads are effective, we might expect to see an increase in interest in those movies in the treated cities, compared with what the interest would have been in the absence of a treatment.

We measure interest in two ways: the number of queries on the movie title for all of the movies and the opening weekend revenue, which could be obtained only for a subset of the movie titles. We use data for the cities whose teams are not playing to estimate the boost in query volume after being exposed to the ad compared with before, and use this to estimate the counterfactual: what the boost would have been without the 10–15% additional ad impressions in those cities associated with the home teams.[§]

Instrumental Variables

Let us compare the Super Bowl example from the previous section to the motivating example that started this paper, $y_c = a + bx_c + e_c$. The advertiser may well determine ad expenditure (and hence exposures) based on factors that also influence outcomes; therefore, we cannot expect x_c to be orthogonal to e_c.

[§]Hartmann and Klapper (17) were the first to recognize the potential of the Super Bowl as a natural experiment and applied this insight to sales of soft drinks and beer. Stephens-Davidowitz et al. (18) independently developed the same idea several months later and applied it to movies.

However, part of ad expenditure is essentially randomly determined because it depends on which teams actually end up playing in the Super Bowl. Therefore, some potentially observable part of x_c is independent of the error term and thus allows us to see how an essentially random variation in spend (or viewership) affects outcomes.

A variable that affects y_c only via its effect on x_c is called an "instrumental variable." Think of this variable as a physical instrument that moves x_c around independently of any movements in e_c. In the Super Bowl example, winning the playoffs is such an instrument, because it effectively increases viewership in two essentially randomly chosen cities.

We can express this mathematically using the following two equations:

$$y_c = bx_c + e_c \quad [1]$$

$$x_c = az_c + d_c. \quad [2]$$

Let $y, x, e \ldots$ be the population analogs of which $y_c, x_c, e_c \ldots$ are the realizations. Here, we face the confounding variable problem when $\text{cov}(x, e) \neq 0$. However, if we can find an instrument z that affects x but, at the same time, is uncorrelated with e, then we can still estimate the casual effect of x on z.

In this simple case, the instrumental variables estimate is simply

$$b^{iv} = \frac{\text{cov}(z,y)}{\text{cov}(z,x)}.$$

To see why this works, substitute the definition of y:

$$b^{iv} = \frac{\text{cov}(z, bx+e)}{\text{cov}(z,x)} \quad [3]$$

$$= \frac{b\ \text{cov}(z,x) + \text{cov}(z,e)}{\text{cov}(z,x)} \quad [4]$$

$$= b. \quad [5]$$

There is an equivalent way to estimate the model described in Eqs. **1** and **2**, known as "two-stage least squares." In this method, we estimate the "first-stage regression" (Eq. **2**) and use that to get the predicted value of x_c. We then plug the predicted value into the "second-stage regression" (Eq. **1**) to determine how the outcome responds to the changes in x that are driven by the instrument. It can be shown that this two-stage approach gives the same results as the instrumental variable approach.

Difference in Differences

In estimating causal effects, it is helpful to have longitudinal data—data for individual units across time. For example, we might have data on advertising expenditures across designated marketing areas (DMAs). Before a campaign, there is zero spend, whereas during the campaign, there is spending at some level in certain DMAs but not in others.

In the simplest case, the outcome is y_{td} at time t in DMA d. Time is labeled B for "before" and A for "after," as in "after the experiment commences." If we think that the experiment will only have a temporary effect, this could also be called "during the experiment." The DMAs are divided into two groups, indexed by T for treatment and C for control.

We are interested in the comparison between the actual outcome for the treated group, y_{TA}, and the counterfactual outcome, which we denote by y_{TF} (what would have happened to the treated group if it had not been treated). Our assumption is that without treatment, the change in the treated group would be the same as that in the control group:

$$y_{TF} - y_{TB} = y_{CA} - y_{CB}.$$

The effect of the treatment is simply the comparison between the actual outcome and the counterfactual:

$$y_{TA} - y_{TF} = [y_{TA} - y_{TB}] - [y_{CA} - y_{CB}],$$

which is easily seen to be a "difference in differences."

An alternative closely related model would be to assume that the percentage change was the same:

$$y_{TF}/y_{TB} = y_{CA}/y_{CB}.$$

This is just a difference in differences in logs.

In addition, there could be other covariates that could help predict y. Letting x_{td} denote a vector of such covariates, our goal is to predict the counterfactual outcome as some function of the observables. Most econometric work uses a linear regression, but one could well be a more complex nonlinear function such as a random forest regression. One can then train a model using subsets of the control group, test the model on the remainder of the control group, and then use the resulting model to predict the counterfactual.

For an example of difference in differences, let us consider the Fargo–Honolulu example described earlier. Suppose that some DMAs were exposed to an ad (treated), and some were not.

s_{TA} = sales after ad campaign for treated groups

s_{TB} = sales before ad campaign for treated groups

s_{CA} = sales after ad campaign for control groups

s_{CB} = sales before ad campaign for control groups

We assemble these numbers into a 2×2 table and add a third column to show the estimate of the counterfactual.

The counterfactual is based on the assumption that that the (unobserved) change in purchases by the treated would be the same as the (observed) change in purchases by the control group. To get the impact of the ad campaign, we then compare the predicted counterfactual sales to the actual sales:

Period	Treatment	Control	Counterfactual
Before	s_{TB}	s_{CB}	s_{TB}
After	s_{TA}	s_{CA}	$s_{TB} + (s_{CA} - s_{CB})$

$$\text{effect of treatment on treated} = (s_{TA} - s_{TB}) - (s_{CA} - s_{CB}).$$

This is, of course a very simple case. We can get an estimate of the sampling variation in sales using a bootstrap, or we can express this as a regression model as described above and additional predictors such as weather, news events, and other exogenous factors of this sort that impact box office revenue in addition to the ad expenditure.

This example illustrates that differences in differences is in the same spirit as the TTTC example described earlier. There, we built a predictive model for the outcome when no treatment was applied. Here, we can build a predictive model for those units where no treatment was applied. We then apply this model to the treated units to get the counterfactual and then compare the actual outcome to the counterfactual.

Note that the difference-in-differences calculation gives us the impact of the treatment on the treated. If there is reason to believe that assignment to the treatment or control groups is

random, then we may be able to interpret the results as the impact of the treatment on the population.

There are many examples of difference in differences in the economics literature. For a recent application to online search advertising, see ref. 19.

Summary

We have described four techniques for estimating causal effects from observational data. In each case, techniques for predictive modeling from machine learning may be useful.

Experiments. With either a designed or natural experiment, it is important to have an estimate of the counterfactual—what would have happened in the absence of the experiment. This is essentially a problem of predictive modeling, an area where machine learning offers several powerful techniques.

Regression Discontinuity. When a treatment is applied depending on some threshold, one can estimate causal effects by comparing outcomes for experimental units on each side of the threshold. In this case, one wants to build a predictive model for behavior near the threshold. We can then use the predicted outcome for the treated group estimated using the training data from the untreated units.

Instrumental Variables. The first stage in an IV model involves predicting treatment as a function of instrumental variables (variables that are thought to be independent of potential confounders) along with other helpful covariates. There are good reasons why the instruments should enter the predictive model linearly, but the other covariates could easily be nonlinear. See ref. 5, pp. 190–192, for a discussion of the problem and some recommendations.

Difference in Differences. In this case, we have two groups, the treated and the untreated, and two time periods, before treatment and after treatment. We also have a number of predictors that may affect the observed values of the outcome for each group. The goal is to estimate a predictive model of what the outcome would be for the treated group if it were not treated. To accomplish this goal, one can use a model, possibly nonlinear, of the observed outcomes of the untreated group in the posttreatment period.

In each of these cases, building a predictive model is a key step in identifying the causal impact. Machine-learning tools offer powerful methods for predictive modeling that may prove useful in this context.

Guide to Further Reading

If you know nothing about machine learning and would like to read an elementary introduction to it written for economists, see ref. 20. The references below are for those familiar with machine learning who want to learn more about the econometric approach to causal inference.

A more advanced approach to causal modeling involves "structural equation modeling," which involves building a specific model of the data generating behavior. For example, in the Honolulu–Fargo example, we might build a model of how marketing managers choose to allocate ad spend across cities and estimate the behavioral effects along with the responses. See ref. 21 for a detailed survey.

There is also a large literature on propensity scores, which are estimates of the probability of treatment as a function of "observed" characteristics. This approach can be contrasted with the confounding variables framework described earlier, which involves selection on "unobservables." See ref. 4 for an up-to-date review.

With respect to the econometrics literature, ref. 5 provides a very accessible introduction and ref. 8 provides a somewhat more advanced description of the methods outlined here. Ref. 23 describes the historical development of these methods.

Finally, there are graphical methods pioneered by Pearl (23, 24) that allow one to analyze complex models to determine when and how various causal effects can be identified.

1. James G, Witten D, Hastie T, Tibshirani R (2013) *An Introduction to Statistical Learning with Applications in R* (Springer, New York).
2. Card D (1999) The causal effect of education on earnings. *Handbook of Labor Economics*, eds Ashenfelter O, Card D (Elsevier, Amsterdam), Vol 3, pp 1801–1863.
3. Box GEP, Hunter JS, Hunter WG (2005) *Statistics for Experimenters* (Wiley-Interscience, New York).
4. Imbens G, Rubin DL (2015) *Causal Inference in Statistics, Social and Biomedical Sciences: An Introduction* (Cambridge Univ Press, New York).
5. Angrist JD, Pischke JS (2009) *Mostly Harmless Econometrics* (Princeton Univ Press, Princeton).
6. Neyman J (1923) On the application of probability theory to agricultural experiments. *Stat Sci* 5(4):465–472.
7. Rubin D (1974) Estimating causal effects of treatment in randomized and nonrandomized studies. *J Educ Psychol* 66(5):688–701.
8. Angrist JD, Pischke JS (2014) *Mastering 'Metrics: The Path from Cause to Effect* (Princeton Univ Press, Princeton).
9. Brodersen KH, Gallusser F, Koehler J, Remy N, Scott SL (2015) Inferring causal impact using Bayesian structural time-series models. *Ann Appl Stat* 9(1):247–274.
10. Abadie A, Diamond A, Hainmueller J (2010) Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *J Am Stat Assoc* 105(490):493–505.
11. Angrist JD, Lavy V (1999) Using Maimonide's rule to estimate the effect of class size on scholastic achievement. *Q J Econ* 114(2):533–575.
12. Valletti T, Ahfeldt GM, Koutroumpis P (2014) *Speed 2.0. Evaluating Access to Universal Digital Highways. Technical Report* (Spatial Economics Research Centre, London).
13. Carpenter C, Dobkin C (2011) The minimum legal drinking age and public health. *J Econ Perspect* 25(2):133–156.
14. Narayanan S, Kalyanam K (2014) *Position Effects in Search Advertising: A Regression Discontinuity Approach. Technical Report* (Stanford University, Stanford, CA).
15. Bakshy E, Eckles D, Bernstein MS (2014) Designing and deploying online field experiments. *Proceedings of the 23rd ACM Conference on the World Wide Web* (Association for Computing Machinery, New York), pp 283–292.
16. Sivaraman A, Winstein K, Thaker P, Balakrishnan H (2014) An experimental study of the learnability of congestion control, *Proceedings of the 2014 ACM Conference on SIGCOMM* (Association for Computing Machinery, New York), Vol 44, Issue 4, pp 479–490.
17. Hartmann WR, Klapper D (2014) *Super Bowl Ads. Technical Report* (Stanford Graduate School of Business, Stanford, CA).
18. Stephens-Davidowitz S, Varian HR, Smith MD (2014) *Super Returns from the Super Bowl?* (Google, Inc., Mountain View, CA).
19. Black T, Nosko C, Tadelis S (2015) Consumer heterogeneity and paid search effectiveness: A large scale field experiment. *Econometrica* 83(1):155–174.
20. Varian HR (2014) Big data: New tricks for econometrics. *J Econ Lit* 28(2):3–28.
21. Reiss PC, Wolak FA (2007) Structural econometric modeling: Rationales and examples from industrial organization, *Handbook of Econometrics* (Elsevier, Amsterdam), Vol 6A.
22. Panhans MT, Singleton JD (2015) *The Empirical Economists Toolkit: From Models to Methods. Technical Report* (Duke University, Durham, NC).
23. Pearl J (2009) *Causality* (Cambridge Univ Press, Cambridge, UK).
24. Pearl J (2013) *Linear Models: A Useful Microscope for Causal Analysis. Technical Report* (University of California, Los Angeles).