Causal Inference in Social Science
An elementary introduction

Hal R. Varian
Google, Inc

Jan 2015
Revised: March 21, 2015

Abstract

This is a short and very elementary introduction to causal inference in social science applications targeted to machine learners. I illustrate the techniques described with examples chosen from the economics and marketing literature.

1 A motivating problem

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.[1] (The machine learning textbook by James et al. [2013] that describes a problem of this sort on page 59.)

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 TV ads for that movie. There are only two cities in the data set: Honolulu, Hawaii and Fargo, North Dakota.

[1] We assume all data has been centered, so we can ignore the constant in the regression.

Suppose that the data set indicates that the advertiser spent 10 cents per capita on TV advertising in Fargo and observed \$1 in sales per capita, while 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.

But 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 seems unlikely, so what is wrong with our regression model?

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's more, the marketing executives that determines ad spend presumably *know* this, and they choose to advertise more where interest is high and less where it is low. So 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

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

The regression coefficient will be unbiased when $\mathrm{cov}(x, e) = 0$.[2]

If we are primarily interested in predicting sales as a function of spend *and the advertiser's behavior remain constant,* this simple regression may be just fine. But usually simple prediction is not the goal; what we want to know is how box office receipts would respond to a *change* in the data generating behavior. The choice of ad expenditure was based on many factors observed by the advertiser; but now we want to predict what the outcome would have been if the advertiser's choice had been different—without observing the factors that actually influenced the original choices.

To put it slightly more formally: we have 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 change to a data generating process such as "increase your spend everywhere by x percent."

[2] Note that 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 is important to understand that the problem isn't simply that there is a missing variable in the regression. There are always missing variables—that's what the error term represents. The problem is that the missing variable, "interest in surfing," affects both the outcome (sales) and the predictor (ads), so the simple regression of sales on ads won't 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 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. But this means that it is very 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.[3]

In this simple example, we have described a *particular* confounding variables. But 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 since the relevant predictors are generally correlated with each other. Nevertheless, we seem comfortable in assuming that the predictors we don't observe—those in the error term—are magically orthogonal to the predictors we do observe!

The "ideal" set of data, from the viewpoint of the analyst, would be data from an advertiser with a totally incompetent advertiser who allocated advertising expenditures totally randomly across cities. If ad expenditure is truly random, then we don't have to worry about confounding variables since the predictors will automatically be orthogonal to the error term. But 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.

[3] It wouldn't 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 doesn't.

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 doesn't 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 James et al. [2013], p.283 for a machine learning approach to this problem and Card [1999] for an econometric approach.)

How does health care affect income? Those who have good jobs tend to have health care, so 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.* But the data was 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 was generated *mutatis mutandis.*

2 Experiments

As Box et al. [2005] 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 some cases are actually infeasible. 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.

But 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" that can shed light on such issues without requiring explicit intervention.

In an experiment, one applies a '*treatment* to some set of *subjects* and observes some *outcomes.* The outcomes for the treated subjects can be compared to the outcomes for the untreated subjects (the control group) to determine the causal effect of the treatment on the subjects.

One may be interested in the "impact of the treatment on the population," in which case one would like the subjects to be a representative sample from the population. Or one might be interested in the how the treatment affected those who actually were treated, in which case one is concerned with the "impact of the treatment on the treated." Or you 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.

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 implement involves voluntary participation, then one may be interested in the impact of the treatment on those who choose (or agree) to be treated.

In marketing, we are often interested the how a change in advertising policies affects a particular firm—the impact of a treatment on a subject that chooses to be treated. This impact may well be different from a subject where treatment is imposed.

3 Fundamental identity of causal inference

Following Angrist and Pischke [2009] we can decompose the observed outcome of a treatment into two effects.

Outcome for treated − Outcome for untreated
= [Outcome for treated − Outcome for treated if not treated]
+ [Outcome for treated if not treated − Outcome for untreated]

= Impact of treatment on treated + selection bias

The first bracketed term is the *impact of the treatment on the treated* while the second bracketed term is the *selection bias*—the difference in outcome between the treated if they were not treated, compared to the outcome for those who were, in reality not treated.

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 to the counterfactual (what would have happened if they were not treated), an insight that goes back to Neyman [1923] and Rubin [1974]. As Rubin emphasized, we cant actually observe what *would have happened* to the treated if they hadn't been treated, so we have to estimate that counterfactual some other way.

As an example, think of our Fargo/Honolulu data set. The true model is

$$y_c = a + x_c b + s_c d + e_c,$$

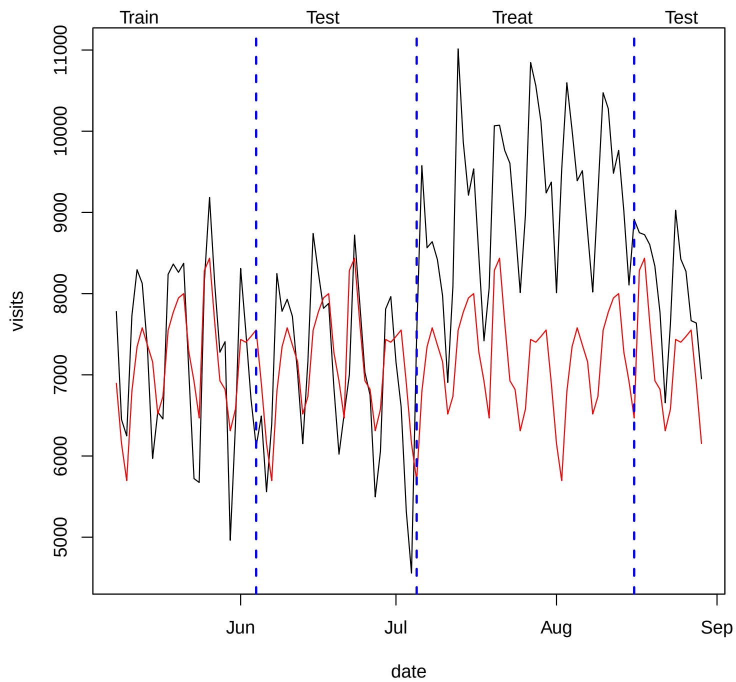
where s_c is a variable that measures "interest in surfing". If the counterfactual is *no* ad expenditure at all, we would still see variation in revenue across cities due to s_c. To determine the causal impact of additional ad expenditure on revenue, we have to compare the observed revenue to a counterfactual revenue that would associated with some default ad expenditure.

By the way, the basic identity nicely shows why randomized trials are the gold standard for causal inference. If the treated group are 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.

4 Impact of an ad campaign

Angrist and Pischke [2014] describe what they call the "Furious Five methods of causal inference:" random assignment, regression, instrumental variables, regression discontinuity, and differences in differences. We will outline these techniques in the next few sections, though we organize the topics slightly differently.

As a baseline case for the analysis, let us consider a single firm that is running a randomized experiment to determine whether it is beneficial to increase its ad spend. We could imagine applying the increase in ad spend to



some consumers and not others, to some geographic location but not others, or at some time but not at other times.

In each case, the challenge is to predict what *would have happened* if the treatment had not been applied. This is particularly difficult for an experiment, since the likelihood that a randomly chosen person buys a particular product during a particular period is typically very small. As Lewis and Rao [2013] have indicated, estimating such small effects can be very difficult.

The challenge is here is something quite familiar to machine learning specialists—predictive modeling. We have time-tested ways to build such a model. In the simplest case, we divide the data into a training set and a test set and adjust the parameters on the training set until we find a good predictive model for the test set. Once we have such a model, we can apply it to the treated units to predict the counterfactual: what would have happened in the absence of treatment. This train-test-treat-compare process is illustrated in Figure 4.

The train-test-treat-compare cycle is a generalization of the classic treatment-control approach to experimentation. In that model, the control group provides an estimate of the counterfactual. However, if we can build a predictive model that improves on predictions of what happens in the absence of treatment, all the better.

The train-test-treat-compare cycle I have outlined is similar to the synthetic control method described by Abadie et al. [2010].[4] 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, machine learning offers a variety of other modeling techniques which may lead to better predictions on the test set and, therefore, better predictions of the counterfactual.

One important caveat: we don't want to use predictors that are correlated with 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. So 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 seasonality as an additional predictor.

5 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 particularly 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. So if we are interested in the causal effect of the threshold the threshold, comparing subjects on each side of the threshold is appealing.

For example, Angrist and Lavy [1999] observes 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. But 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

[4]See also the time-series literature on interrupted regression, intervention analysis, structural change detection, etc.

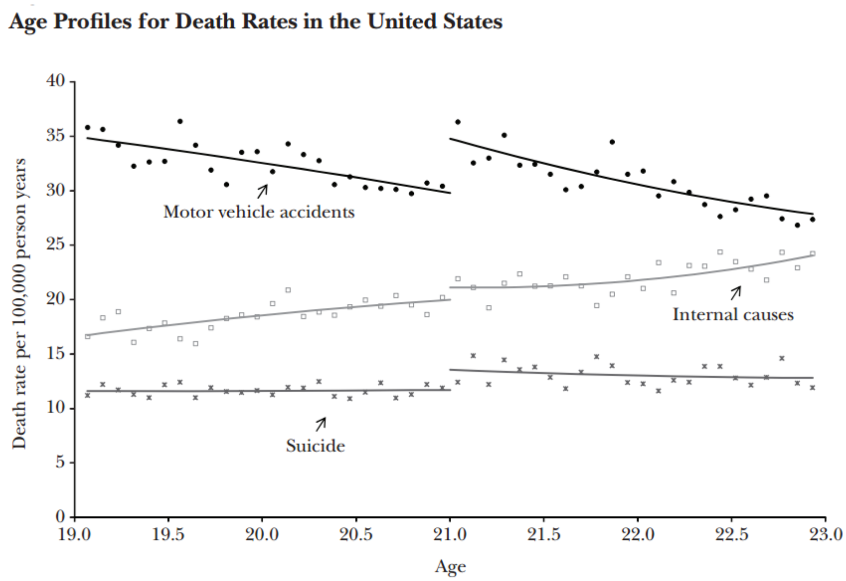


Figure 1: Death rates by age by type.

with (say) 41 initial students (who end up with 20-person classes), thereby teasing out the causal effect of class size on educational performance. Since it is essentially random which side of the threshold a particular subject ends up on, so this is almost as good as random assignment to different sized classes.[5]

Another nice example is the study by Valletti et al. [2014] that aims to estimate the impact of broadband speed on housing values. Just looking at the observational data is will not resolve this issue since houses in newer areas may be both more expensive and have better broadband connections. But 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 [2011], who examine the impact of the minimum legal drinking age on mortality. The story is told in Figure 5, which is taken from this paper.[6] As you can see, there is a major jump in motor vehicle accidents at the age of 21. Someone who is 20.5 years old isn't that different from someone who is 21 years old, on average,

[5]The actual policies used are a bit more complicated than I have described; see the cited source or Angrist and Pischke [2009] for a more detailed description.

[6]See also the helpful discussion in Angrist and Pischke [2014].

but 21 year olds have much higher death rates from automobile accidents, suggesting that the minimum drinking age *causes* this effect.

Regression discontinuity design is very attractive when algorithms are used to make a choice. 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 Narayanan and Kalyanam [2014] 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 to the status quo since only cases close to the threshold are impacted.

6 Natural experiments

If there is a threshold involved in making a decision, by focusing only on those cases close to the threshold we may have something that is almost as good as random assignment to treatment and control. But 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, as compared to 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 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 as compared to 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. The results are shown in Figure 2. As can easily be seen, those extra ad impressions made a big difference!

Details of the analysis are available in Stephens-Davidowitz et al. [2014]. Hartmann and Klapper [2014] independently applied the same idea to sales of soft drinks and beer that were advertised in Super Bowls.

7 Instrumental variables

Let us compare the Super Bowl example to the motivating example that started this paper, $y_c = a + bx_c + e_c$. The advertiser may well determine ad expenditure based on various factors that also influence outcomes, so we can't expect x_c to be orthogonal to e_c. However, *part* of ad expenditure is essentially randomly determined since it depends on which teams actually end up playing in the Super Bowl. So some 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, since it effectively increases viewership in two essentially randomly chosen cities.

We can express this mathematically using the following two equations:

$$y_c = bx_c + e_c \qquad (1)$$

$$x_c = az_c + d_c \qquad (2)$$

Letting $y, x, e \ldots$ be the population analogs of which $y_c, x_c, e_c \ldots$ are the realizations, we face the confounding variable problem when $\text{cov}(x, e) \neq 0$. But if we can find an instrument z such that $\text{cov}(z, x) \neq 0$ (z affects x) but $\text{cov}(z, e) = 0$ then we can still estimate the casual effect of x on z.

In fact, in this case the IV estimate is simply

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

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

$$
\begin{aligned}
b^{iv} &= \frac{\text{cov}(z, bx + e)}{\text{cov}(z, x)} && (3) \\
&= \frac{b\ \text{cov}(z, x) + \text{cov}(z, e)}{\text{cov}(z, x)} && (4) \\
&= b && (5)
\end{aligned}
$$

This calculation is correct only for the population. However, it can be shown that the sample analog of these computations gives you a good estimate of the casual effect for large sample sizes.

To take another example suppose you want to estimate how the demand for air travel responds to a change in ticket prices. Let y_c be number of tickets sold, p_c the price of the tickets, and e_c an error term. The natural regression to run is

$$y_c = bp_c + e_c.$$

But by now we should be familiar with the problem: the ticket prices are chosen by the airlines and will generally depend on factors in the error term. For example, if the economy is booming airlines might increase prices and if the economy is slow they might decrease prices. But the state of the economy affects not only the price of tickets but also the amount of air travel, so it is a confounding variable.

One solution is to figure out some proxy for the state of the economy and add that as a predictor in the regression. Another solution is to find a variable that affects ticket price but is uncorrelated with the error term. For example, a change in the taxes on air travel could provide such an instrument.

8 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 DMA (Designated Marketing Areas). Prior to a campaign there is zero spend, during the campaign there is spending at some level in certain DMAs but not in others.

In the simplest case, the outcome is x_{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 could consider comparing the treated groups before and after: $x_{TA} - x_{TB}$. However, it may be that something else happened while the experiment was progressing. To control for this, we compare the before-after change to the treated group to the before-after change of the control group: $x_{CA} - x_{CB}$. If the change in the treated group was the same as the change in the control group, it would suggest that there was no effect. Here the control group is simply estimate of the counterfactual: what would have happened to the treatment group if they weren't treated.

The final estimate is then the "difference in differences,"

$$[x_{TA} - x_{TB}] - [x_{CA} - x_{CB}],$$

which is simply the difference between what actually happened and an estimate of the counterfactual—what happened to those who were not treated.

8.1 Example of difference-in-differences

Let us consider the Fargo-Honolulu example described earlier. Suppose that Some DMAs were exposed to an ad (treated), some were not.

- s_{TA} = sales after treatment in treated groups
- s_{TB} = sales before treatment in treated groups
- s_{CA} = sales after treatment in control groups
- s_{CB} = sales before treatment in control groups

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

	treatment	control	counterfactual
before	s_{TB}	s_{CB}	s_{TB}
after	s_{TA}	s_{CA}	$s_{TB} + (s_{CA} - s_{CB})$

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 treatment we then compare the counterfactual to the actual:

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

This is a difference in differences. It might be more natural in this example to estimate a multiplicative model, which would then involve a "ratio of ratios" or a difference-of-differences in the logs of sales.)

This is, of course 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 and additional predictors such as weather, news events, and other exogenous factors of this sort which impact box office revenue in addition to the ad expenditure.

Note that the difference-in-differences calculation is giving us the impact of the treatment on the treated, unless, of course, the treatment is applied to a randomly chosen sample of the population.

Differences in differences is in the same spirit as the train-test-treat 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.

There are many examples of diff-in-diff in the economics literature. For a recent application to online advertising, see Black et al. [2015].

9 Guide to further reading

There are other more advanced approaches to causal modeling. Economists are fond of "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 Reiss and Wolak [2007] for a detailed survey.

There is also a large literature on propensity scores, which is a way to estimate the probability that a particular subject is chosen for treatment. Such models can allow estimates of the "treatment on the treated" to be extrapolated to estimates of the treatment on the population. See Rubin and Imbens [2013] for an up-to-date review.

There is also an emerging literature on causal methods for high-dimensional data that is motivated by genomics applications. See ETH [2015] and Institute [2015] for a selection of papers in this area. The considerations described in this paper do not seem relevant to this literature, though I could be mistaken.

Finally, there are graphical methods pioneered by Pearl [2009, 2013] that allow one to analyze complex models to determine when and how various causal effects can be identified.

With respect to the econometrics literature, Angrist and Pischke [2014] provides a very accessible introduction and Angrist and Pischke [2009] provides a somewhat more advanced description of the methods outlined here.

References

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of Californias tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505, 2010.

Joshua D. Angrist and Victor Lavy. Using Maimonide's rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics*, 114(2):533–575, 1999.

Joshua D. Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics*. Princeton University Press, 2009.

Joshua D. Angrist and Jörn-Steffen Pischke. *Mastering 'Metrics: the Path from Cause to Effect*. Princeton University Press, 2014.

Tom Black, Chris Nosko, and Steve Tadelis. Consumer heterogeneity and paid search effectiveness: A large scale field experiment. *Econometrica*, 83 (1):155–174, 2015.

George E. P. Box, J. Stuart Hunter, and William G. Hunter. *Statistics for Experimenters*. Wiley-Interscience, New York, 2005.

David Card. The causal effect of education on earnings. In Orley Ashenfelter and David Card, editors, *Handbook of Labor Economics*, volume 3, pages 1801–1863. Elsevier, 1999.

Chrisopher Carpenter and Carlos Dobkin. The minimum legal drinking age and public health. *Journal of Economic Perspectives*, 25(2):133–156, 2011. URL https://www.aeaweb.org/articles.php?doi=10.1257/jep.25.2.133.

ETH. Challenges in machine learning, 2015. URL http://www.causality.inf.ethz.ch/cause-effect.php?page=help.

Wesley R. Hartmann and Daniel Klapper. Super bowl ads. Technical report, Stanford Graduate School of Business, 2014.

Max Planck Institute. Causal inference at the max-planck-institute for intelligent systems, 2015. URL http://webdav.tuebingen.mpg.de/causality/.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R*. Springer, New York, 2013.

Randall A. Lewis and Justin M. Rao. On the near impossibility of measuring the returns to advertising. Technical report, Google, Inc. and Microsoft Research, 2013. URL http://justinmrao.com/lewis_rao_nearimpossibility.pdf.

Sridhar Narayanan and Kirhi Kalyanam. Position effects in search advertising: A regression discontinuity approach. Technical report, Stanford University, 2014. URL http://faculty-gsb.stanford.edu/narayanan/documents/search.pdf.

Jerzy Neyman. On the application of probability theory to agricultural experiments. *Statistical Science*, 5(4):465–472, 1923.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Judea Pearl. Linear models: A useful microscope for causal analysis. Technical report, UCLA, 2013. URL http://ftp.cs.ucla.edu/pub/stat_ser/r409.pdf.

Peter C. Reiss and Frank A. Wolak. Structural econometric modeling: Rationales and examples from industrial organization. In *Handbook of Econometrics*, volume 6A. Elsevier, 2007. URL https://web.stanford.edu/

group/fwolak/cgi-bin/sites/default/files/files/Structural%20Econometric%20Modeling_Rationales%20and%20Examples%20From%20Industrial%20Organization_Reiss,%20Wolak.pdf.

Donald Rubin. Estimating causal effects of treatment in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):689, 1974.

Donald L. Rubin and Guido Imbens. *Causal Inference in Statistics*. Cambridge University Press, New York, 2013.

Seth Stephens-Davidowitz, Hal R. Varian, and Michael D. Smith. Super returns from the Super Bowl? Technical report, Google, Inc., 2014.

Tommaso Valletti, Gabriel M. Ahfeldt, and Pantelis Koutroumpis. Speed 2.0. evaluating access to universal digital highways. Technical report, SERC, July 2014. URL http://www.spatialeconomics.ac.uk/textonly/serc/publications/download/sercdp0161.pdf.