# Chapter 22 – Causal Inference

## Correlation does not imply causation

### 22.1 – Introduction

Open any textbook on statistics, and you’ll find in it, somewhere, the statement that ‘correlation does not imply causation’. Whenever we see two phenomena A and B fluctuate in unison, our instinct has us belief that one (A) is somehow exerting an influence on the other (B). In other words, A causes B. Our brains are wired to take advantage of this because it makes evolutionary sense. If you are a caveman Grog and you see your fellow caveman Gurr eat some berries and subsequently die, it is not a bad assumption that the eating those berries caused Gurr to die. But it is not a given. Perhaps Gurr died of something else, and the berries had nothing to do with it. Two phenomena A and B can correlate without one having a direct influence on the other. Perhaps just by chance. Or, more common, some third phenomena C influencing A and B. In this scenario a change in C does cause a chance in A and B, but if we don’t observe C, we just see A and B correlate and falsely assume A is exerting its influence on B or vice versa. Here, C is referred to as a confounding variable, and the resulting correlation between A and B is called a spurious correlation. Tyler Vigen has collected a few bizarre correlations that very clearly demonstrate that things in our world can correlate that surely have no direct causal effect on each other. Figure 22.1 shows three of my favorites but check out his website and his book for a lot more of them (links in the references section).

### 22.2 –Experiments

Those who are experimental scientists are acutely aware of this issue and make a clear distinction between collecting data through a study or an experiment. In the news, these two words are typically used interchangeably, but in science the difference is crucial in how the results can be interpreted. In an experiment, scientists are typically interested in the effect or influence of an independent variable on a dependent variable. The independent variable is the thing the experimenter varies, the dependent variable is what they measure. For example, we can give subjects a certain amount of caffeine (the independent variable) and measure the speed at which they type as measured in words per minute (the dependent variable). Some subjects get a low dose of caffeine, some get a high dose of caffeine. And the experimenter, beforehand, posits their hypothesis: that a higher dose of caffeine will result in a statistically higher typing speed. This is the alternative hypothesis and what the experimenter wants to proof. Typically, in statistics, it is phrased in the opposite way: the experimenter wants to find statistical evidence that allow them to reject the null hypothesis. The null hypothesis is the opposite of the alternative hypothesis: there is no statistically significant effect of caffeine on typing speed and any variation seen in typing speed merely represents measurement noise and chance. But to so do, the experimenters must make sure the condition of ‘ceteris paribus’ is met. This Latin for ‘with all other things remaining equal’. It means that during the experiment, every other variable that might reasonably influence typing speed is kept the same for those receiving a high or a low dose of caffeine. In this way, it is guaranteed that we can attribute the difference in typing speed to the caffeine, and the caffeine alone. And especially important is that we must make sure we don’t accidentally introduce a bias in who gets the high dose of caffeine and who gets the low dose. If for some reason the subjects in the high dose group already have a higher typing speed to begin with, the experiment can no longer claim that typing speed is influence by the amount of caffeine. You might wonder how that would even happen, and this is often the focus of many scientific discussions on all experimental studies. Tiny details inadvertently overlooked by the experimenters can still invalidate an experiment. In our case, imagine this. The experimenter sends out an email to all the students at their university, asking them to participate in the study. Students are eager to participate because it earns them valuable credits. And to join the study, they need to reply to the email with a short paragraph about themselves and why they want to participate. After a few minutes, the experimenter’s inbox blows up with students willing to participate (that’s how eager they are…). So, the experimenter decides to take the first 100 replies, and divides them into the two groups. The first 50 replies go in the high caffeine dose condition, the remaining 50 go in the low dose condition. Can you spot the problem? If all students started replying as soon as they all got the email (at the same time), the 50 first repliers are likely to represent people that are on average faster typists than the 50 next repliers, they managed to get that paragraph written faster. And now all of them are assigned to the high dose caffeine condition and the experiment is no longer unbiased. You might find this example a bit contrived, and it is. But there many examples of these subtle effects that unfortunately did invalidate experiments. The solution, in the end is simple, and it is randomization. What the experimenter should have done is two randomly assign the 100 participants to one of the two groups, by flipping a coin, or using a random number generator. Randomization, per definition, will ultimately make the two groups of participants indistinguishable from each other, as we increase the number of participants further and further. And this is key to a properly conducted science experiment.

### 22.3 – Studies

However, the scientific method described above is often simply not an option. In the real world, people decide for themselves what group they are in. You choose to play Farmville, and you choose to buy a certain product. If we want to test whether playing Farmville versus not playing Farmville influences some other measurable outcome, like overall happiness levels, we cannot randomly assign them to two unbiased groups. What if happy people naturally gravitate toward playing Farmville? In that way we would prove the opposite hypothesis: happiness causes Farmville, not vice versa. Of course, we could invite a group of people who have never played Farmville and divide them into two random groups and make half of them play Farmville for a while and measure the outcome as a true experiment. But this isn’t practical, and why waste all the data that is already out there without the need for expensive experimentation? In clinical research, an additional problem arises. If we want to test the effectiveness of a new drug for a certain disease, it is simply unethical to divide a group of people into those who receive the drug and those who will receive a placebo.

### 22.3 – Digital Advertisement as an example

We have all seen a lot of them. The clutter of digital advertisements on the websites we visit. Some of them are random, but most of them are eerily relevant to what products or services we as consumers are interested in. This is obviously not a coincidence. Everything we do online is tracked, mostly by still using cookies, although a variety of other methods are used as well. Furthermore, this data can be joined with our offline activities, like charges to our credit cards, our subscriptions, our rewards programs at supermarkets and other retailers, geolocation data from our smartphones, the things we post on social media apps like Twitter Facebook, and TikTok. And that, unfortunately, is just the tip of the iceberg. But regardless of the data used, the objective remains the same: deliver advertisements to people that are more likely to be influenced by it. For digital ads, this is the job of the publishers. Publishers act on behalf of a brand to target a specific audience that is more likely to convert. Conversion in digital advertising is the term used for any action that results in a favorable outcome for the brand advertised. Favorable outcomes are typically ranked according to their position within the so-called funnel. A funnel is the process of a consumer going through the steps necessary to reach the desired outcome. This can be the purchase of the product advertised or subscribing to a service. These actions represent the end of the funnel and or obviously the most lucrative type of conversion from the perspective of the brand. Before that, the user first must visit the website, browse the various offerings, and put them in their shopping cart. These still are considered conversions because they raise brand awareness and might influence future consumer behavior in favor of the brand. So, from a game theory perspective, the publishers seek out consumers to show ads to that they feel have a higher chance of converting. Different publishers have different strategies to do so, and brands often have several publishers work on a single campaign (the promotion of a new product or service). This, however, creates the problem of attribution. If multiple publishers have targeted the same consumer, and the consumer does convert, who gets the credit for that? But more importantly, the strategy of targeting users that are likely to convert creates a slew of unwanted confounds that make it impossible to say anything meaningful about the effect of any of the ads we are shown.

### 22.3 – Confounds and Covariates

As we discussed already, confounds are variables that make it appear two other variables have a causal relationship, while it is the confounding variable that causes them simply to correlate. Staying within the world of digital advertisement, what we hope is that our independent variable of showing someone and ad while effect the dependent variable of that same person converting. However, since people are actively targeted because we think they are more likely to convert than others, confounding variables come into play, hard and fast. Imagine a publisher working for a large gaming studio to advertise their latest video game. So, the publisher starts to model people in terms of their gaming enthusiasm. There are many different signals one could use for that, like their credit card data showing they have bought games in the past or their demographics for example. But publishers will have to merge and match 3rd party data with the data they themselves have, which is notoriously difficult to do. What publishers do see is users visiting websites, and the type of content we consume on these websites are a reasonable proxy of who we are and what we like. So, the publisher starts targeting users who visit game related websites, such as game forums, and review websites. And herein lies the problem. If a user visits game related websites, isn’t it reasonable to assume they are more likely to buy the game, regardless of whether they have been shown the ad for it. Being a gamer is confounding variable, it makes it more likely you buy the game, and it makes it more likely you are being shown an ad for it. There can still be variables that influence a person being targeted and not whether they would purchase the game, perhaps a piece of demographics the publisher included that has no predictive value. Similarly, there are variables that influence whether users buy the game, but not whether they will be targeted. In this case, perhaps the publisher forgot or had no access to that variable and therefore wasn’t used to target users. These are two examples of so-called covariates, which do not confound the relationship between ads and conversions.

### 22.5 – Directed Acyclical Graphs

This problem is by no means exclusive to digital advertising and new. The medical world has long faced the same issue. When somebody develops a disease for which a drug is being developed, they will the ones to be treated with it. But the reason for having the disease in the first place also probably effects the prognosis or outcome. Again, where it might appear that the drug is working for people with the disease, it is the confounding effect of the patient on both getting the drug and the prognosis. In recent years, a lot of work has been done on trying work out the statistics of causation, or causal inference. The confounding situations described above are typically visualized in what is called a Directed Acyclical Graph (DAG). It is directed because we assume some direction of the influence. An ad is supposed to lead to conversion, and a treatment is supposed to a cured patient. The standard notation used in the most basic of these DAGs is A for treatment, Y for outcome, and W for the confounding variables. Note that we are leaving out the covariates, as they are not essential for how we turn the DAG into a model we can train using machine learning and subsequently use a method called propensity score matching to remove the confounding effects W in our data as best we can, and hopefully leaving us with just the true causal effect of A on Y.

### 22.6 – A possible solution

Before we describe propensity score matching (PSM), it is worth pointing out that in some situations, it might be possible to turn this confounding situation into a more controlled experiment. One effort I was involved in when working at a digital advertisement metrics agency aimed to circumvent the necessity for PSM by introducing randomization into the DAG. If somehow, a randomization can be introduced such that users who are considered more likely to convert are never exposed to the ad, we break the confounding tie between A and Y. One classical way of doing so is AB testing. In AB testing, not all targeted users are shown the ad that they are targeted for. Instead, they receive an ad for totally different product, or a public service announcement. But AB testing is costly, because you’ll have to sacrifice some potential converters to a control group, and it is extremely messy, as keeping track of users as they are targeted by multiple publishers is hard if not impossible. Instead, we argued that while we can’t randomize who will receive an add, that doesn’t mean they saw the ad. Many webpages have ads just about everywhere and depending on how the user interacts with the webpage the ads are on; they might not ever see most of them. If a user doesn’t scroll down to bring the ad in view, they were technically shown the ad (A=1), but they didn’t see it (V=0). Since we could track whether this was the case, it thus represented an extra step in the original DAG, between treatment A and outcome. At first, we reasoned that this constituted a random process, and as such, it would have removed the confounding effect. Unfortunately, we did quickly figure out that even seeing the ad is influenced by the same variables used for targeting and whether a user converts. To return to the gamer example. Not only is the gamer targeted because they visit game related websites (A), but it also makes it more likely they see the ad, because they spend more time on these websites, and are more likely to scroll down further than people with less enthusiasm for gaming. So, back to propensity score matching.

### 22.7 – Propensity Score Matching - Theory

Propensity Score Matching is a technique to remove confounding effects W on A and B by means of stratification of the data based on the propensity of W on A. That is a mouth full, but it is a relatively straightforward process in this its original form. Imagine that the publisher did target specific individuals at all. Instead, they would, like in a true experiment, flip a coin for each user to device whether to show them an ad. Done like this, the propensity of each user to be shown an ad is the same, as measured by the probability of receiving the add. In more mathematical terms, each user has a p(A) = 0.5, a 50% percent chance of being shown an ad. If that were the case, we would end up with two truly randomized groups, one in which an ad was shown (A=1), and one where the ad wasn’t shown (A=0), without any bias introduced by the confounding variables W that also effect Y. Now we simply compare the conversion rate between the two groups. If group (A=1) has a significantly higher conversion rate than group (A=0) then we have shown that the ad had a measurable positive effect on whether a user converted on it. Furthermore, it doesn’t matter what the p(A) is, as long as it is the same for every user. This is key in PSM. If every user has a p(A) = 0.1, it just means we end up with randomized groups, although now one group is obviously much bigger, as in this scenario only 10% of users will receive the ad. But despite the difference in size, there shouldn’t be any other differences between the users in the two groups, because nothing but chance is deciding what group a user ends up in. But that is not what happens. Instead, we know that ads are targeted. Some people receive ads, some don’t. But this targeting isn’t perfect. The models used by the publisher produce some probability of a user have a high propensity to convert, and the publisher uses a threshold to decide whom to target or not. Therefore, each user has their own p(A), based on whatever model the publisher uses predicts. And since it is predicting p(A) based on the confounding W that also predict Y, we find ourselves in the situation whereby default p(A) correlates with p(Y), just because of W, and not for causal reasons. However, this correlation isn’t perfect. And it is this imperfection we take advantage of in PSM. This starts by reverse engineering the probability p(A) of each user. And to that we will need a machine learning model.

### 22.8 – Propensity Score Matching – The Kitchen Sink

In most situations we have some idea of what the confounding variables in our study could have been. In digital advertisement, we can assume the publishers have used certain features. In clinical research, we can assume that demographics, phenotypes, and genotypes are likely to confound our results. Usually, it is best to use what I refer to as the kitchen sink approach. Build your model W with as many variables as you can find, if you don’t create too many features compared to the number of actual data points you have. In the digital advertising scenario this could mean we build a large matrix W of users by websites they visited and populate each element *Wij* with either a binary value (this user has visited this website), a count (the number of times this user has visited this website) or some more complex feature, such as the TFiDF metric we discussed in the chapter on NLP. Only now it is not the Term Frequency - inverse Document Frequency, it is the Per User Website Visit Count Frequency - inverse Overall Website Frequency measure (Only that doesn’t make for a particularly nice acronym: PUWVCFiOWF, so let’s stick with TFiDF). If we had additional information, we could include these as well. For example, an internet user comes with a so-called User Agent String, which at some level of granularity specifies their operating system and browser. In addition, we could map their IP address to a specific geographical region and use that geographical region to further deduce the probabilities of the users’ demographics.

### 22.9 – Propensity Score Matching – Predicting p(A)

The next step is to build a machine learning model using our carefully crafted matrix W to predict A (whether an individual was shown the ad). Specifically, we will need a model that predicts the probability of that. Most machine learning algorithm doing classification will, and traditionally it is usually good old fashioned logistic regression. After training the model, hopefully with some but not perfect accuracy (more on that later) we use it to predict the probability of A=1 for each user.

**22.10 – Stratification, Coverage, and Computing Lift**

Remember how we stated earlier that users with the same propensity p(A) to be targeted essentially represented a randomized group? That is what we will try to artificially create by combining users that have very similar probabilities associated for their A=1. Stratification means we divide the users up into buckets of roughly the same p(A). For example, we can divide our users into buckets of p(A)>=0 to p(A)<=0.1, p>0.1 to p<=0.2, up to p(A)>0.9 and p(A)<=1.0. And for each of these buckets, we compute what is called lift in digital advertising: the actual causal influence of the ad on conversion behavior, with the confounding effects removed. So, for each bucket, we compute the ratio between p(Y | A = 1) and p(Y | A = 0). Since, within each bucket, the propensity to be targeted p(A) is the same, this ratio represents relative impact of A on Y. But wait, you say, if somebody has a low p(A) and they should have a low p(Y), correct? If the users were targeted using the same variables W that also drove conversions p(Y), anyone with a low p(A) has a low p(Y). Yes, this is true. But this is where the concept of coverage comes in. Coverage refers to how well the publishers have been able to model the users p(A), compared to their actual p(Y). When we see that W is a set of confounding variables, that doesn’t mean they perfectly and completely link p(A) and p(Y). Those covariates we discussed briefly also play a part, and the publisher likely modeled certain aspects of the user that actual do not correlate p(Y). And aspects the drive p(Y) may have been overlooked by the publisher so those don’t affect p(A). In the two extremes we have one scenario where the publisher has captured nothing about the user that predicts p(Y). and therefore p(A) and p(Y) are completed decoupled. At the other extreme, the publisher magically captured everything there is to predict p(Y) and therefore it is perfectly coupled with p(A). The first extreme is the best-case scenario, because it essentially means that we have a randomized experiment. The second extreme the worst-case scenario, because whatever we manage to remove as an influence of W on p(A), we remove from p(Y). The result: no measurable remaining causal effect of p(A) on p(Y).

### 22.9 – Demos

### 22.10 – References

* Pearl, J. (2009). Causality. United Kingdom: Cambridge University Press.
* Cunningham, S. (2021). Causal Inference: The Mixtape. United Kingdom: Yale University Press.