Causality in Statistics: Development, Methods, and Applications

Alex Ho

May 2022

Faculty Mentor: Professor Chad Shaw, Dept. of Molecular and Human Genetics, Baylor College of Medicine, and Dept. of Statistics, Rice University
Editors: Professor Elizabeth Jennings McGuffey, Dept. of Statistics, Rice University; Xin Tan, Dept. of Statistics, Rice University

Abstract

Statistical inference has long been used for describing underlying distributions of observed data, but it is ultimately limited in its capabilities. Causal inference was subsequently developed to provide stronger capabilities by estimating causal relations between variables. Building upon structural causal model and the potential outcomes frameworks, many methods have been developed to address problems in modern data science and machine learning applications. While estimating causal relationships is a nonstandard part of statistics, it holds great promise for applications with critical outcomes due to a richer understanding of how variables may interact. This project aims to describe to statisticians and data scientists why causal inference is needed, a brief history leading up to the development of rigorous causal inference methods, the foundations and applications of modern causal inference methods, and the implications and future directions of causal inference with regard to data science and analytics.

1 Introduction

Introductory statistics courses frequently remind students that "correlation does not imply causation." However, it is not standard practice to teach these students what does in fact imply causation, even if estimating causal relationships may be useful. For example, causal relationships are essential to how people and organizations make informed decisions. Knowing the difference between "smoking and lung cancer are correlated" and "smoking causes cancer" give vastly different amounts of information about how to act. Moreover, with data science and machine learning becoming pervasive in automated decision making processes, it is clear that correlation relationships are insufficient for critical applications with severe consequences.

We will start this paper by outlining the motivation of why classical statistical inference is insufficient and which situations causal inference is required. Namely, we will focus on the problems of spurious correlation and conclusions from observational data when viewed through the lens of statistical inference in Sections 1.2 and 1.3 respectively. Then, we will continue by describing Simpson's Paradox as a motivating example in 1.4. Next, we present a brief history outlining the foundational papers of causal inference in 2.1 as well as how applied fields have developed their own forms of causal inference in 2.2. Later, we will go into more detail about the two main branches of causal inference: the potential outcomes model and the structural causal model, in Sections 3.1 and 3.2 respectively. Section 3 also covers common statistical methods adapted for causal inference through instrumental variables in 3.3 and Mendelian randomization in 3.4. Section 4 covers more recent, alternative methods developed for causality. Finally, we will end with a discussion in Section 5 about the implications of everything covered as well as what is next for causal inference. Additionally, Appendix A has more detail on instrumental variables by walking through a simulated case study.

1.1 Traditional Statistical Inference versus Causal Inference

We will first distinguish the differences between traditional statistical and causal inference. Traditional statistical inference consists of regression, estimation, and hypothesis testing in order to gain knowledge about the underlying assumed distribution from which our observed samples have been drawn from. With this knowledge, we can then predict what will happen in this system, but usually only if we assume these predictions are independently drawn from identical distributions, the $i.i.d$ assumption. Causal inference seeks to extend the capabilities of classical statistical inference; it can be seen as encapsulating everything that traditional statistical inference can, but also what happens when the system is subject to changed conditions.[1] For example, traditional statistics alone without the design of a randomized controlled trial cannot and does not claim to predict how a system will behave under interventions; prediction under altering conditions such as changing treatments, introducing new policies, or forcing behavior is outside of the scope of statistical inference. It is only *after* these conditions have been changed and data from the new data generating dis-

1. Changing conditions is different from dynamic parameters, like non-stationary variance, which can be estimated with traditional statistics.

tribution is collected can we approximate how the system will look. Causal inference builds on top of statistical inference with careful model design and assumptions so that analysis of observational data can lead to estimated causal conclusions. In the rest of Section 1, we will continue to show how traditional statistical inference alone can be insufficient in certain situations.

1.2 Spurious Relationships

A spurious relationship is a mathematical relationship in which two or more variables are associated but not causally related, because of either coincidence or the presence of another, unseen factor, often called a "confounding" or "lurking" variable. These relationships can be trivially avoidable, such as the relationship that honey-producing bee colonies inversely correlate with juvenile arrests for marijuana, as described by Vigen (2015), but other times can have more severe consequences if missed.

Silber and Rosenbaum (1997) describe such a case of potentially dangerous spurious correlation by studying hospital mortality and complication rates. When two outcome measures, in this case mortality and complication rates, are intended to measure the same underlying quantity, hospital quality of care, it is expected that there would be a high degree of correlation between them. Moreover, as data quality and quantity increases, the correlation should also increase. However, the authors show that these expectations are not always met by studying three predictive models which each adjusts for severity of illness in different ways. By comparing two hospital rankings, based on case mortality and complication rates, initial regression fits shows that they are highly correlated when not adjusted for severity. However, as more clinical data are added to the models, the correlation tends to disappear. In other words, unobserved variables left out of the regression adjustment spuriously strengthened the correlation between mortality and complication rates when both are predicted by that variable, and there exist partial correlations between mortality, complication rate, and severity covariates. Without properly controlling for confounding variables, users may be lulled into thinking that their model and variables are working as expected. However, it is not always so simple, as the authors note, and additional steps must be taken to ensure that metrics are truly measuring the desired quantity and are not a case of spurious correlation.

Khoury and Ioannidis (2014) also reflect on how big data may come with big errors in the context of epidemiology. When associations are posited between high-accuracy data, like genomic sequences, and poorly measured data, like administrative claims health data, the conclusions can only be drawn at the same level of rigour as the weakest link. Big data are largely observational and can often be contaminated by selection bias[2], confounding variables, and lack of generalizability. However, with tools developed through causal inference, we can test and correct for these problems. Once we can recognize and adjust for these problems, statisticians and analysts are one step closer to having their models reflect their true, usually causal, inquiries. With just traditional statistics, it would be a practically impossible task to try and find all the unaccounted for confounders and unknown biases.

2. Selection bias is the bias introduced into data by a selection process which does not have proper randomization. In this case, the sample obtained is not representative of the population intended to be analyzed, which is a common assumption in data analysis.

	Small	Large	Total
Treatment A	14.3% (200/1400)	30% (30/100)	**15.3% (230/1500)**
Treatment B	**10% (4/40)**	**16.7% (100/600)**	16.3% (104/340)

Table 1: Simpson's Paradox example. Hypothetical data for mortality rate among patients with large or small tumors given treatment A or treatment B.

1.3 RCTs and Experimental versus Observational Data

Randomized controlled trials (RCTs) take advantage of interventional data in a controlled way, and are usually deemed the gold standard for understanding the causal effect of a treatment or policy. In an RCT, participants are randomly assigned differing treatments. The random treatment allocation ensures that treatment status will not be confounded with either measured or unmeasured features. Therefore, the causal effect of treatment on outcomes can be estimated by comparing outcomes directly between treated and untreated subjects.

Grossman and Mackenzie (2005) give a description of how RCTs differ from observational studies. The overarching case for RCTs is their ability to properly balance unknown and confounding variables between treatment groups and a control group. Observational data is an extreme form of data which is often assumed but rarely abides by the assumption of *i.i.d.*, where each data point is independently sampled from the same distribution. Interventional data has known interventions, where we observe data sets sampled from multiple distributions each of which is the result of a known intervention.

However, there are many cases where we cannot sample interventional data. For example, treatments that are more invasive (e.g. surgery) may be impossible to replicate in the control group; too few people might have a certain disease and also be available for investigation in both treatment and non-treatment groups; recruitment of participants to a particular trial may be too difficult; or may be unethical to impose a treatment on a group or withhold the treatment from another group. Thus we come to the dilemma that we would like to have interventional data as much as possible to estimate causal effects, but are limited to observational data. However, causal inference allows us to approximate these causal effects without RCTs and interventional data, unlike traditional statistical inference.

1.4 Simpson's Paradox and Misleading Statistics

Beyond spurious correlation and lack of experimental data, there are cases when traditional statistics can provide misleading and confusing conclusions alone. Suppose a hospital wants to compare two treatments, A and B, for and their effectiveness on cancer. Treatment A is a standard and commonly used treatment for cancer, but Treatment B is new and there is a limited availability. After a while, the hospital compiles all the data and sees that Treatment A has a lower overall mortality rate than Treatment B. Discouraged, the researchers try conditioning different variables to understand the full picture, but they come across something peculiar: when accounting for the size of the tumor treated (small and large), Treatment B does better in both cases (see Table 1 for details). How could Treatment

A be better overall, but Treatment B be better in all cases when conditioning on another variable? This type of paradox was first described by Simpson (1951) and is difficult to solve with the summary statistics alone.

Pearl (2011) provides two causal structures for why this type of paradox may occur:

1. The treatment is a cause of both the tumor size and the mortality rate.

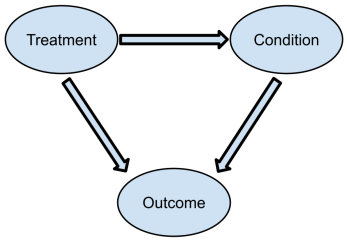
2. The tumor size is a cause of both the treatment and the mortality rate.

An illustration of the causal diagrams for each situation is shown in Figure 1. In Figure 1, we use a directed acyclic graph model to model causal relations between variables. The nodes are variables of interest, while the arrows demonstrate the directional causal relationship between two variables. For example, if there is an arrow pointing from treatment to condition, this can be interpreted as the treatment has a causal effect on condition. Further details on these causal models can be found in Section 3.2.

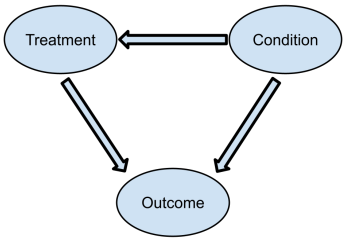
If causal structure 1 is true, then Treatment A would be the better choice. Because Treatment B is more scarce, the tumor may progress and worsen over the time period waiting for it. Thus, it could be that many of those who were initially prescribed Treatment B started with small tumors which then developed into large tumors as they were waiting for the treatment. Therefore, giving the treatment to a patient increases their probability of dying, and we should pay attention to the total counts to recommend Treatment A.

If causal structure 2 is true, then Treatment B would be the better choice. Because Treatment B is more scarce, the doctors may want to save it for the more severe cases. Thus, the vast majority of the easier cases with small tumors are given Treatment A. Moreover, the severe cases with large tumors are more often given Treatment B. Therefore, Treatment B is disproportionately being given to people who are more likely to die regardless of the treatment, which skews the total mortality rate down, and we should pay attention to the subgroup statistics.

Kügelgen, Gresele, and Schölkopf (2021) also describe a real world example of Simpson's Paradox with COVID-19 data. In this study, COVID-19 case fatality rates were measured in China and Italy. The authors found that case fatality rates were lower in Italy for every age group, but higher overall. Much of this confusion can be explained by how the countries were compared. Because case fatality rates are relative frequencies, the absolute magnitude and distribution of cases between age groups were hidden. In Italy, many of the confirmed cases were in older patients, while most of the cases in China were in the 30-59 year old age range. The authors note that there are several possible reasons why these sample distributions may be different, including the fact that the Italian population has a higher median age of 45.4 while China's population has a median age of 38.4. However, regardless of the true reason for the difference in age group distributions, this case of Simpson's Paradox arises because Italy had a proportionally large proportion of elderly people who had confirmed cases, and it is also this group which is at higher risk of fatality from COVID-19.



(a) Diagram for causal structure 1: the treatment is a cause of the tumor size and the mortality rate.



(b) Diagram for causal structure 2: the tumor size is the cause of the treatment and the mortality rate.

Figure 1: Causal diagrams for the two possible scenarios in our example of Simpson's Paradox.

2 Development of Causal Inference

2.1 In Statistics

Wright (1921) and Wright (1923) together generate the first attempt to formulate the notion of directional causal relations. Wright used equations with graphs to communicate causal relationships with linear relations. For example, if we have x which is a disease variable and y which is a symptom variable, Wright proposes that we should write

$$y = \beta x + u_y$$

which says that x is the severity of the illness, y is the severity of the symptom, and u_y is all other factors, the "exogenous variables," other than the disease x which could influence y. Finally, β is the "path coefficient" which quantifies the direct causal effect of x on y. However, there is still the problem that this equality can also be represented as

$$x = \frac{(y - u_Y)}{\beta}$$

which might be misinterpreted to mean that the symptom influences the disease.

To address this ambiguity, Wright supplements the equation with what is called a "path diagram" which depicts arrows being drawn from (perceived) causes to their (perceived) effects. It is also important to note that the absence of an arrow implies a strong assumption or indication that there is no relation between the two variables. Continuing this example with x, y, and u_Y, a potential path diagram is illustrated in Figure 2. Here we can see that both x and u_Y have a causal effect on y, however y does not have a causal effect on either x or u_Y. We can also note that x and u_Y are assumed to be causally independent, meaning that there is no relation between these two variables. Using this diagram as a supplement to either of the equations mentioned before, there is now no ambiguity in the direction of the effects between the symptoms and the disease and the relationships between the variables.

Wright's main contributions were that he introduced the notation of path diagrams with graphical rules for writing down the covariance of any pair of observed variables in terms of

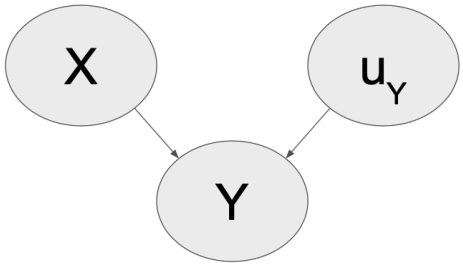


Figure 2: Path diagram example. Let x be the severity of the illness, y be the severity of the symptom, and u_y is all other factors other than the illness x which could influence y. If we are interested in how the severity of the illness affects the severity of the symptoms, then we would have the pictured path diagram.

path coefficients and covariances of the error terms.

Splawa-Neyman, Dabrowska, and Speed (1990) translated a copy of the original work by Neyman (1923), and describe the beginnings of the potential outcomes framework. This framework is based on the unit-based response variable, $Y_x(u)$, which is to be interpreted as "the value that outcome Y would have in experimental unit u had treatment X been x" where the unit u could be any individual subject. However, this notation has since been revised in the binary treatment or no treatment case among statisticians to read $Y_i(1)$ or Y_i^1 for the potential outcome of individual i under treatment, and $Y_i(0)$ or Y_i^0 represents the outcome under no treatment.

Thus, if X is binary, we can represent the overall value of Y as the following

$$Y = xY_1 + (1 - x)Y_0$$

and we can notice that when $x = 1$ or $x = 0$ one of the terms Y_0 or Y_1 respectively zeroes out, showing that the outcome Y must be realized as either Y_0 or Y_1.

2.2 In Other Fields

At the same time, researchers in econometrics, biostatistics, and other social sciences developed their own methods to draw causal conclusions, primarily Haavelmo (1943) and Simon (1977).

Prior to Haavelmo (1943), econometrics was subject to purely statistical inferences and had weak justification for causal interpretations. Haavelmo's work introduced three main contributions to economic thought:

1. A structural equation is not a statistical relationship, but rather a set of hypothetical experiments which are to be encoded in the systems of equations

2. Because of the first point, the model is capable of answering policy intervention questions without further input by the designer

3. A mathematical description which takes any arbitrary model and produces a quantitative response to policy intervention queries

Imbens (2020), who also won the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2021, covers the advantages and disadvantages of potential outcomes and graphical models for causal inference in economics. Imbens notes that there are several features of the potential outcomes framework that may contribute its current popularity in economics. Primarily are some assumptions, namely monotonicity, convexity, or concavity, that are easily captured in the potential outcomes framework relative to the graphical models approach which are critical in many identification strategies in economics. Additionally, it was natural that the potential outcomes adapts easily to economics because economic models for supply and demand settings developed the primitives of potential outcome functions independently. Thus, current econometrics techniques are latently baked with mathematics analogous to the potential outcomes framework because many of the inquiries for economics are causal.

3 Primary Tools for Causal Inference

Based on the prior work of Splawa-Neyman, Dabrowska, and Speed (1990) and Wright (1921), Rubin (1974) introduced the general potential outcomes framework which is based upon traditional methods of statistics, and Pearl (2009b) created a new framework called structural causal models (SCMs) based around his prior work on Bayesian networks and structural equation models[3]. Together, the potential outcomes method allows us to use statistics to calculate causal relationships, while SCMs allow us to model structural dependencies through directed acyclic graphs (DAGs).

The modern mathematical formulation for representing causal relations was only recently developed. Rubin (1974) and Pearl (2009b) are often credited with formalizing the modern methods for causal inference, but they were strongly influenced by the works described in Section 2. Other common methods for estimating causal effects that will be discussed in this section are instrumental variables, and a specific application of instrumental variables in Mendelian randomization.

3.1 Potential Outcomes

Rubin (1974) provides a more complete perspective for using the potential outcomes framework. Intuitively, potential outcomes seeks to infer the effect of some treatment or policy on some outcome. For example, will taking an aspirin fix my headache? Again, we will use the notation Y_i^1 to represent the potential outcome of individual i under treatment, and Y_i^0 to be the potential outcome of individual i under no treatment. Thus, the causal effect of the treatment can be defined as $\delta_i = Y_i^1 - Y_i^0$.

Unfortunately, δ_i is impossible to measure because of the fundamental problem of causal inference: we cannot go back to counterfactually change what has already happened to observe the other potential outcomes of the world. However, there are ways we can get around this fundamental problem.

3. Structural Equation Models are models that aim to explain the relationships between measured variables and latent variables, as well as the relationships between latent variables.

We now introduce the idea of average treatment effects (ATE). The subscript i is often omitted because it is implicit and easier notationally as we are now dealing with entire groups. These can thus be represented as the following:

$$\begin{aligned} ATE &= \mathbb{E}[\delta_i] \\ &= \mathbb{E}[Y^1 - Y^0] \\ &= \mathbb{E}[Y^1] - \mathbb{E}[Y^0]. \end{aligned}$$

We also have what is called the associational difference $\mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0]$, however we cannot say $\mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0] = ATE$ unless we have the assumption of *Ignorability*, that is, the treatment is independent of each potential outcome. This assumption means that in situations in which the treatment and the outcomes have a common, often unobserved, cause, the associational difference is not equal to the ATE. This assumption can also be seen from the perspective of exchangeability, or $\mathbb{E}[Y^1|T=0] = \mathbb{E}[Y^0|T=0]$. In words, exchangeability means that the treated and control group have the same expected outcome when given the same treatment. Under these assumptions, our causal metrics are *identifiable* if we can compute them from a purely statistical quantity. Namely for the ATE, we can use the adjustment formula for identifiability:

$$\begin{aligned} \mathbb{E}[Y^1 - Y^0] &= \mathbb{E}_X \mathbb{E}[Y^1 - Y^0|X] \\ &= \mathbb{E}_X[\mathbb{E}[Y|T=1,X] - \mathbb{E}[Y|T=0,X]]. \end{aligned}$$

As an example, we will refer to Luque-Fernandez et al. (2018). Here, we use simulated data and then estimate the effect of sodium intake on blood pressure using a traditional approach, and then after using the adjustment formula. The outcome Y is blood pressure, treatment T is sodium intake, and covariates X are age and amount of protein excreted in urine. Because this data is simulated, we know the true ATE is 1.05. The naive statistical estimate of the ATE, $\mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0]$, yields 5.33, which is 407% different from the true ATE. Now we will see how the adjustment formula can approximate the ATE.

First, we need identification, measuring our causal quantity with statistical quantities using the adjustment formula:

$$\mathbb{E}[Y^1 - Y^0] = \mathbb{E}_X[\mathbb{E}[Y|T=1,X] - \mathbb{E}[Y|T=0,X]].$$

Then, we can estimate the outer expectation by using an arithmetic mean:

$$\frac{1}{n}\sum_x [\mathbb{E}[Y|T=1,X] - \mathbb{E}[Y|T=0,X]].$$

Each conditional expectation can then be estimated using linear regression. For example, we could use the following model:

$$Y = \beta_0 + \beta_1 X + \beta_2 T + \beta_3 XT$$

so that when $T = 0$ the model reverts to

$$Y = \beta_0 + \beta_1 X$$

while when we have $T = 1$ the model is equal to

$$Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)X.$$

Then for each data point set $X = x$, and we can compute the difference in our expected values for when $T = 0$ and $T = 1$. By using this linear model for the adjustment formula, the ATE estimate is 0.85, which is much closer to our true ATE of 1.05 than the naive estimate.

Rosenbaum and Rubin (1983) introduce the notion of propensity scores in the context of estimating causal effects. A propensity score is loosely the conditional probability that an individual will be assigned a particular treatment given the observed covariates. By using conditional probabilities of being assigned a group given the covariates, the bias of these covariates can successfully be mitigated. Propensity score matching (PSM) is a statistical creation but allows for causal interpretation. Let X be a binary treatment and S be an arbitrary set of covariates. Then the propensity score $L(s)$ is the probability that action $X = 1$ is chosen by a participant with characteristics $S = s$:

$$L(s) = P(X = 1|S = s),$$

which allows us to view $L(s)$ as a function of the random variable S. Because of this, X and S are independent given $L(s)$. A simple proof for this independence follows directly from the Local Markov assumption discussed in Section 3.2 if we view s as generating both $L(s)$ and X, that is that s has a causal relationship with both the probability of the action and the action, then X and s are then independent because s is not a descendant of X. It can then be justified that the propensity score can be an efficient estimator of the adjusted estimand, meaning that it is an unbiased estimand of the causal effect.

3.2 Structural Causal Models

Pearl's structural causal models depend on the directed acyclic graph (DAG) notation and framework. In these graphs we have the following components:

1. Nodes are variables of interest and the edges are the causal relations between the variables

2. Causal relations must be directed so it must exclusively be that either $A \to B$ or $B \to A$ if there exists a relation between A and B.

3. Graphs must be acyclic, meaning that there cannot be the case where A causes B which causes C which causes A (the cycle of $\ldots \to A \to B \to C \to A \to B \to \ldots$)

Two examples for graphs which are not DAGs are shown in Figures 3 and 4. Figure 3 shows a graph which breaks the DAG requirement that they the relations are always

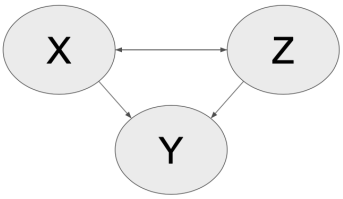


Figure 3: Example of a graph which is not directed. There exists a relation between X and Z which is not defined exclusively in one direction.

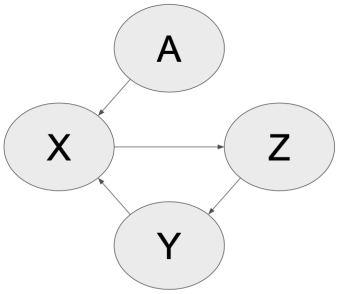


Figure 4: Example of a graph which is not acyclic. There exists a cycle between X, Y, Z

exclusively directed in one direction, while Figure 4 breaks the DAG requirement that the relations do not create a cycle.

Pearl (1998) introduces the advances in Bayesian networks and the causal interpretation of such graphs. Pearl also details the conditions for identifiability, the conditions that allow us to recover causal relations between variables from standard statistics without random experimentation. One important assumption from these is called the Local Markov assumption: *Given its parents in the DAG, a node X is independent of all its non-descendants.* This assumption is useful because it allows for easier computation of joint distributions if we have a DAG of the variables. For example, in statistical modelling, we know the joint distribution can be conditionally factorized as

$$P(x_1, x_2, \ldots, x_n) = P(x_1) \prod P(x_i | x_{i-1}, \ldots, x_1).$$

However, the conditional probability requires that we know the values of 2^{n-1} parameters if the conditioning variables only take on two values. Therefore, we come to a problem because the number of parameters we need to estimate for each conditional probability grows exponentially. The Local Markov assumption gives us Bayesian network factorization:

$$P(x_1, \ldots, x_n) = \prod_i P(x_i | pa_i),$$

which means that a joint distribution is equal to the product of each variable in the joint distribution conditioned only on its parent nodes. The parents of a node X are all other nodes which have a directed edge into X. However, we also need the minimality assumption for proper interpretation, which states that adjacent nodes in a DAG are dependent. Given the

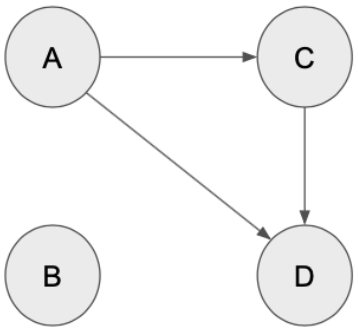


Figure 5: Example DAG to demonstrate Bayesian network factorization.

DAG between variables X and Y is $X \to Y$ means that by the Local Markov assumption alone $P(x, y) = P(x)P(y)$; however we know that Y is not independent of X. With the addition of the minimality assumption, we correctly get $P(x, y) = P(x)P(y|x)$.

For example, take the variables A, B, C, D with the DAG in Figure 5 and suppose we are interested in their joint distribution $P(A, B, C, D)$. A standard conditional factorization would require us to find $P(A)$, $P(B|A)$, $P(C|A, B)$, $P(D|A, B, C)$. However, with our Bayesian network factorization, we can see that the joint distribution is the following:

$$P(A, B, C, D) = P(A)P(B)P(C|A)P(D|A, C)$$

which drastically reduces the computational burden for calculating the joint distribution because we do not have to account for an exponential number of parameters in the conditional probability.

Pearl (2009a) covers the recent advances in causal inference, and highlights the changes necessary to move from traditional statistical analysis to causal analysis of multivariate data. He focuses on the assumptions required with causal inferences, the languages used in formulating those assumptions, the conditional nature of all causal and counterfactual conclusions, and the methods that have been developed for the assessment of such claims. These methods use a general theory of causation based on the structural causal model which is based on the work of Wright (1921) and Wright (1923). Pearl also claims that his model encapsulates the potential outcomes framework, and provides additional support in showing how we should condition on variables in the causal structure.

One of the major contributions of Pearl (2009b) is the popularization of the back-door criterion and its solution the back-door adjustment. Given a causal graph G with non-experimental data on a subset V of the observed variables, we wish to estimate the effect of interventionally setting $X = x$ on the response variables $Y \subset V$ where $X \subset V$. The back-door criterion tests whether a proposed subset $Z \subseteq V$ is sufficient for identifying the intervention's effects, and is defined as the following:

> *A set of variables Z satisfies the* back-door *criterion relative to an ordered pair of variables (X_i, X_j) in a DAG G if:*
>
> 1. *no node in Z is a descendant of X_i; and*

2. Z blocks every path between X_i and X_j that contains an arrow into X_i.

Similarly, if X and Y are two disjoint subsets of nodes in G, then Z satisfies the back-door criterion relative to (X, Y) if it satisfies the criterion relative to any pair (X_i, X_j) such that $X_i \in X$ and $X_j \in Y$.

Here, a "path" means any consecutive sequence of edges, without regard to their directionalities. By "blocked path" we mean a path that cannot be traced without traversing a pair of arrows that collide "head-to-head." Arrows that meet head-to-head do not constitute a connection for the purpose of passing information, so their meeting node is be called a "collider". For example with $X \to Z \leftarrow Y$ there is a blocked path from X to Y because Z, which is a collider, is along the only path between X and Y. The case of $X \leftarrow Z \to Y$ also has a blocked path because we must traverse arrows which do not flow the same direction, but this time Z is not a collider.

Now if we have a set of variables Z which satisfy the back-door criterion relative to (X, Y), then we can identify the causal effect of X on Y with the formula

$$P(y|\hat{x}) = \sum_z P(y|x, z)P(z)$$

where $\hat{x}$ is intervening to set $X = \hat{x}$. This estimation is the so called back-door adjustment, which is significant because it allows us to quickly and easily estimate causal effects within a DAG.

As an example, suppose we have the DAG shown in Figure 6, and we are interested in the effect of X on Y. If we were capable of intervening on this DAG, then we could set X to be some value and measure its effect on Y. This case is easy because by intervening on X, we control X, and not A so we effectively remove the incoming arrows to X. With these arrows removed, the backdoor paths flowing from X to A to X to Y are also removed, and our estimation of the effect of X and Y are not confounded.

However, in the case that we do not have access to intervention, we can use the back-door path criterion and adjustment. In this DAG, we also have A and Z in addition to our variables of interest X and Y. Then we can calculate the causal effect of X on Y if either A or Z satisfy the backdoor criterion. After inspection, we can see that both A and Z satisfy the backdoor criterion, so we can calculate $P(Y|\hat{X}) = P(Y|X, Z)P(Z) + P(Y|X, A)P(A)$.

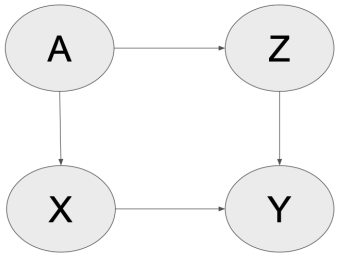


Figure 6: Example DAG to demonstrate the back-door criterion and adjustment.

3.3 Instrumental Variables

A common method for estimating causal effects with standard statistics is with instrumental variables. Assume we have some data generating process which follows

$$y_i = \alpha + \beta X_i + \epsilon_i.$$

A standard ordinary least squares (OLS) approach to this problem would require the assumption that $\mathbb{E}[\epsilon_i | X_i] = 0$. However this can be easily violated under the following circumstances:

1. Changes in Y change the value of at least one of the features of X (reverse causation).
2. There are omitted "exogenous" variables that affect both X and Y.
3. X has non-random measurement error.

Thus it may be the case that $\mathbb{E}[\hat{\beta}_{OLS}] \neq \beta$ and $\hat{\beta}_{OLS}$ is a biased and inconsistent estimator. This poses problems because we primarily wish to measure the relationship between X_i and y_i, but because $\mathbb{E}[\epsilon_i | X_i] \neq 0$, there is some relationship between ϵ_i and X_i which entangles our estimates of y_i by having an additional factor hidden in ϵ_i.

One way to account for this problem is with instrumental variables. An instrumental variable (IV) is a variable that satisfies the following assumptions:

1. The IV must be correlated with the endogenous explanatory variables, conditionally on the other covariates.
2. The IV cannot be correlated with the error term in the explanatory equation, conditionally on the other covariates.

For example, in our case of the linear model, a variable Z is an instrumental variable if it is correlated with X but not with ϵ.

As an example, suppose we were interested in the causal effect of smoking on health. Let Y be a metric for health, and X be smoking patterns. We know with prior knowledge that exogenous variables, such as depression, may affect both health and smoking, or health may affect smoking patterns. Now let Z be the tax rate for tobacco products. Tobacco tax can only be correlated with health through its effect on smoking. If there is a correlation between tobacco tax and health, this may be viewed as evidence that smoking causes changes in health. A traditional linear regression would only reveal that there is a correlation between smoking and health. This amount of information does little in terms of how we should intervene or make policy changes based on this data. However by using instrumental variables we can come to directed causal conclusions which can allow us to make decisions which influence this causal chain. A simulated case study to demonstrate this example is included in Appendix A.

3.4 Mendelian Randomization

Mendelian randomization is an example of an instrumental variable method used in biostatistics and epidemiology. The method uses the properties of genetic variation (usually single nucleotide polymorphisms) strongly associated with some trait as an instrumental variable to test for and estimate a causal effect of the trait on an outcome of interest from observational data. The genetic variation used has either well-understood effects on trait expression patterns, or effects that are similar to those produced by controllable exposures. It is important that the genetic variation must only affect the outcome of interest indirectly through its effect on the exposure of interest.

Mendelian randomization requires three assumptions on the instrumental variables:

1. The genetic variant(s) being used as an instrument for the exposure is associated with the exposure.

2. There are confounders of the genetic variant(s) and the outcome of interest.

3. There is no line of influence between the genetic variant(s) and the outcome other than through the exposure.

For example, suppose we are interested in measuring how BMI influences blood pressure. Individuals in the general public have many differences in their life (job, marriage status, diet, activity level) which may confound the relationship between BMI and blood pressure. However, these differences should not influence their *genetic* predisposition for having a high or low BMI. People with this gene which gives people a propensity towards a higher BMI will, on average, have a higher measured BMI level. If we can find that people with higher genetic propensities toward high BMI also have an increased rate of high blood pressure, then we have evidence that BMI causally increases risk of high blood pressure. This is because we can treat the genetic propensity as an instrumental variable, which itself can't be affected by unobserved confounders or hypertension itself (reverse causation).

Ference et al. (2012) showed an application of Mendelian randomization in a meta analysis to estimate the causal relationship between lower plasma low-density lipoprotein cholesterol on the risk of coronary heart disease. The study concluded that there was a stronger causal relationship between extensive early-life exposure to lower plasma low-density lipoprotein cholesterol to coronary heart disease than the standard methods used at the time.

4 Alternative Methods

In this section we will provide a brief overview of some alternative methods to causal inference and discovery which were more recently developed as a way to expose the readers to emerging approaches. Peters, Janzing, and Schölkopf (2017) provide a comprehensive overview of causal inference from a modern computational statistics and machine learning perspective, and has a different perspective in approaching causality. They focus on the cause-effect problem, as well as how causality can be inferred without random experiments, all assuming a basic probability or machine learning background.

4.1 Independent Component Analysis

Hyvärinen (2013) covers the modern advances of independent component analysis (ICA), which has become a useful method for identifiable causal analysis. By viewing causal analysis as a special case of blind source separation, the author covers recent work which has shown independent component analysis to be useful for structural equation modeling.

Khemakhem et al. (2020) creates a unifying theory of nonlinear ICA with variational autoencoders (VAEs). They introduce an identifiable framework which provably ensures disentanglement of latent variables by nonlinear ICA by assuming each observed variable x is modulated in some way by an auxiliary variable u, and the generating prior can be factorized conditionally on u.

ICA methods and their nonlinear counterparts give us an alternative way for both discovering causal variables simultaneously with their causal effects. The nonlinear ICA approaches are the most complex approach to causal inference and rely on deep neural networks with large datasets, but also allow for the modeling and discovering of complex causal relationships not possible in the basic potential outcomes or graphical model frameworks alone.

Now, a growing body of literature for nonlinear ICA is developing to discover additional conditions for identifiability spurred by the works of Hyvarinen and Morioka (2017) (see Yao et al. (2022) and Hyvärinen (2013)).

4.2 Deconfounding Data

Wang et al. (2020) aims to deconfound movie exposure data so that robust recommendations can be made downstream. They frame the problem of recommendation as a counterfactual problem, and define the potential outcomes reparameterization. They use a Poisson factorization method which assumes the exposure data a_{ui} takes the following form:

$$a_{ui}|\pi_u, \lambda_i \sim \text{Poisson}(\pi_u^\intercal \lambda_i)$$

where $\pi_u \sim \text{Gamma}(c_1, c_2)$ and $\lambda_i \sim \text{Gamma}(c_3, c_4)$.

Many forms of deconfounded data factorization exist in biostatistics and econometrics; however they have not been popularized to the same extent in other fields with interventional queries, such as recommendation systems shown by Wang. Deconfounding factorization approaches allow data scientists to treat the reconstructed observational data as approximately equivalent to RCT data if the factorization model's conditions are met, thus making them a powerful development if the field can be expanded to cover more data generating processes.

4.3 Causal Representation Learning

Causality also has applications to representation learning as described by Schölkopf et al. (2021). Representation learning is set of techniques that allows a model to discover the representations needed for downstream tasks on its own from raw data. This approach aims to replace manual feature engineering, as it allows a machine to simultaneously learn the features and use them to perform a specific task. In this paper, the authors show that causal inference can help some fundamental issues with deep learning:

1. Learning disentangled representations
2. Learning transferable mechanisms
3. Learning interventional world models and reasoning

Adding causality to deep learning methods can allow the same changes to be had as between statistical inference and causal inference. Fundamentally, deep learning models are based in statistics with the multilayer perceptron and cannot by design uncover causal relations. However, by injecting causal inference methods, stronger conclusions can be made so that applications of modern machine learning are much more reliable and generalizable.

5 Discussion

Data analysis has been limited in its capabilities due classical statistical inference's strong and often unrealistic assumptions. Although big data and machine learning methods are quickly becoming the standard for making decisions, many practitioners do not realize the consequences of many of these unmet assumptions. Hidden problems such as spurious correlation, paradoxes, and misleading data are additionally even more common to find in our age of high dimensional big data. But thankfully, these problems can be mitigated using the right tools derived from causal inference.

Building on top of traditional statistics, causal inference can account for many of the downfalls to uncover stronger conclusions about learned relations between variables. With robust frameworks like potential outcomes and graphical models, we can estimate causal relations with observation data to work past these limitations. By using relatively small amounts of prior domain knowledge about each dataset to make additional assumptions, practitioners can have the benefits of causal conclusions through methods such as Mendelian randomization and instrumental variables. Although causal inference requires prior domain expertise and additional assumptions, it is unlikely to be the case that scientists will have some data without any background information about it. Even if there are only sparse amounts of background information, there are tools both available and being developed to test for adherence to these additional assumptions.

With an active research community seeking to model even more complex causal relationships, there appears to be a bright future to this pocket of statistics and data science. I hope that causal inference may become a more widely taught subject as the field matures, as doing so may allow us to accelerate the discovery of causal conclusions in scientific discovery and decision making.

References

Ference, Brian A, Wonsuk Yoo, Issa Alesh, Nitin Mahajan, Karolina K Mirowska, Abhishek Mewada, Joel Kahn, Luis Afonso, Kim Allan Williams, and John M Flack. 2012. "Effect of Long-Term Exposure to Lower Low-Density Lipoprotein Cholesterol Beginning Early in Life on the Risk of Coronary Heart Disease: a Mendelian Randomization Analysis." *Journal of the American College of Cardiology* 60 (25): 2631–2639.

Grossman, Jason, and Fiona J Mackenzie. 2005. "The Randomized Controlled Trial: Gold Standard, or Merely Standard?" *Perspectives in Biology and Medicine* 48 (4): 516–534.

Haavelmo, Trygve. 1943. "The statistical implications of a system of simultaneous equations." *Econometrica, Journal of the Econometric Society,* 1–12.

Hyvarinen, Aapo, and Hiroshi Morioka. 2017. "Nonlinear ICA of Temporally Dependent Stationary Sources." In *Artificial Intelligence and Statistics,* 460–469. PMLR.

Hyvärinen, Aapo. 2013. "Independent Component Analysis: Recent Advances." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 371 (1984): 20110534.

Imbens, Guido W. 2020. "Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics." *Journal of Economic Literature* 58 (4): 1129–79.

Khemakhem, Ilyes, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. 2020. "Variational Autoencoders and Nonlinear ICA: A Unifying Framework." In *International Conference on Artificial Intelligence and Statistics,* 2207–2217. PMLR.

Khoury, Muin J., and John P. A. Ioannidis. 2014. "Big Data Meets Public Health." *Science* 346 (6213): 1054–1055. https://doi.org/10.1126/science.aaa2709. eprint: https://www.science.org/doi/pdf/10.1126/science.aaa2709. https://www.science.org/doi/abs/10.1126/science.aaa2709.

Kügelgen, Julius von, Luigi Gresele, and Bernhard Schölkopf. 2021. "Simpson's Paradox in COVID-19 Case Fatality Rates: A Mediation Analysis of Age-Related Causal Effects." *IEEE Transactions on Artificial Intelligence* 2 (1): 18–27.

Luque-Fernandez, Miguel Angel, Michael Schomaker, Bernard Rachet, and Mireille E Schnitzer. 2018. "Targeted maximum likelihood estimation for a binary treatment: A tutorial." *Statistics in medicine* 37 (16): 2530–2546.

Neyman, Jerzy S. 1923. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.(Translated and edited by DM Dabrowska and TP Speed, Statistical Science (1990), 5, 465-480)." *Annals of Agricultural Sciences* 10:1–51.

Pearl, Judea. 1998. "Graphical Models for Probabilistic and Causal Reasoning." *Quantified Representation of Uncertainty and Imprecision,* 367–389.

———. 2009a. "Causal Inference in Statistics: An Overview." *Statistics Surveys* 3:96–146.

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

———. 2011. "Simpson's Paradox: An Anatomy."

Peters, Jonas, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of Causal Inference: Foundations and Learning Algorithms.* The MIT Press.

Rosenbaum, Paul R, and Donald B Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1): 41–55.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66 (5): 688.

Schölkopf, Bernhard, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. 2021. "Toward Causal Representation Learning." *Proceedings of the IEEE* 109 (5): 612–634.

Silber, Jeffrey H, and Paul R Rosenbaum. 1997. "A Spurious Correlation Between Hospital Mortality and complication Rates: the Importance of Severity Adjustment." *Medical Care,* OS77–OS92.

Simon, Herbert A. 1977. *Causal Ordering and Identifiability,* 53–80. Dordrecht: Springer Netherlands. ISBN: 978-94-010-9521-1. https://doi.org/10.1007/978-94-010-9521-1_5. https://doi.org/10.1007/978-94-010-9521-1_5.

Simpson, Edward H. 1951. "The Interpretation of Interaction in Contingency Tables." *Journal of the Royal Statistical Society: Series B (Methodological)* 13 (2): 238–241.

Splawa-Neyman, Jerzy, Dorota M Dabrowska, and TP Speed. 1990. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9." *Statistical Science,* 465–472.

Vigen, Tyler. 2015. *Spurious Correlations.* Hachette UK.

Wang, Yixin, Dawen Liang, Laurent Charlin, and David M Blei. 2020. "Causal Inference for Recommender Systems." In *Fourteenth ACM Conference on Recommender Systems,* 426–431.

Wright, Sewall. 1921. "Correlation and Causation." *Journal of Agricultural Research* (Washington, D.C.) 20 (3). ISSN: 0095-9758.

———. 1923. "The Theory of Path Coefficients a Reply to Niles's Criticism." *Genetics* 8 (3): 239.

Yao, Weiran, Yuewen Sun, Alex Ho, Changyin Sun, and Kun Zhang. 2022. "Learning Temporally Latent Causal Processes from General Temporal Data." In *International Conference on Learning Representations.*

Appendix A Instrumental Variables: Simulated Case Study

Here we will provide an instructional handout on Two Stage Least Squares which is meant to be an in-depth lecture-style notes for undergraduates. The expected background is basic linear regression.

A.1 Problem Setup

Let us return to the example with smoking, depression, and health outcomes. Let Y be a metric for health, X be smoking patterns, Z be the tax rate for tobacco products, and U be a metric for depression. Suppose the true generating function between these variables are the following:

$$U \sim \text{Exponential}(8)$$

$$Z \sim \text{Exponential}(3)$$

$$X \sim \text{Exponential}(10) - 5Z + U$$

$$Y = -3U - 5X$$

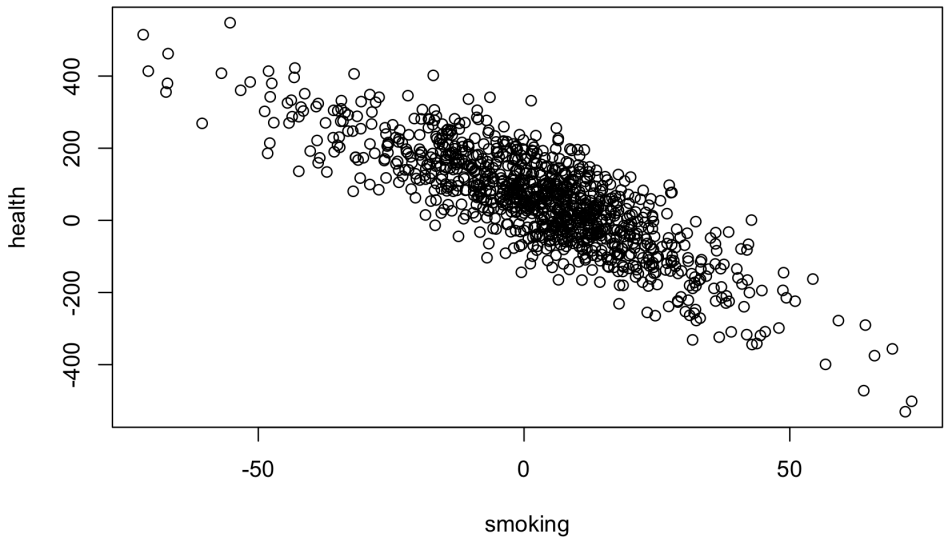
also represented in R as the following:

```
depression = rexp(n, rate = 1/10) + rnorm(n, sd = 2)

tobacco_tax = rexp(n, rate = 1/3) + rnorm(n, sd = 0.5)

smoking = rexp(n, rate = 1/8) + depression + -5*tobacco_tax + rnorm(n, sd = 5)

health = -3*depression + -5*smoking + 100 + rnorm(n, sd = 75)
```

which shows us that depression is a common cause for both health and smoking, and that tobacco tax rates are independent of depression. However, our scientist does not take into account other factors and wants to see the association between smoking and health outcomes.

A.2 Standard Linear Regression

Plotting the data

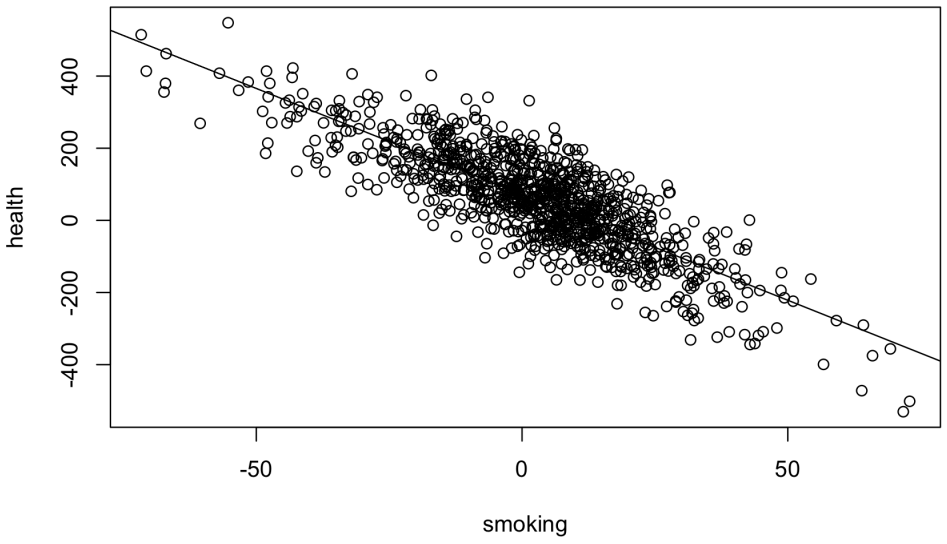


the scientist sees that there is a linear relationship between X and Y, therefore the proposed model he comes up with is

$$Y = \beta X + \epsilon$$

Running a standard linear regression with the data to find $\hat{\beta}$ gives us $\hat{\beta} = (X^T X)^{-1} Y = -5.8746$ with statistical significance:

```
Call:
lm(formula = health ~ smoking)

Residuals:
     Min       1Q   Median       3Q      Max
-219.263  -54.330    0.311   53.274  267.409

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  72.2381     2.5787   28.01   <2e-16 ***
smoking      -5.8746     0.1288  -45.62   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 80.63 on 998 degrees of freedom
Multiple R-squared:  0.6759, Adjusted R-squared:  0.6755
F-statistic:  2081 on 1 and 998 DF,  p-value: < 2.2e-16
```



Our scientist would like to stop here and decide that smoking is a strong predictor of negative health outcomes; however a senior scientist notes that there may be confounders influencing the estimates. The senior scientist then recommends that our scientist uses tobacco tax rates as an instrumental variable (IV) in a two stage least squares estimation.

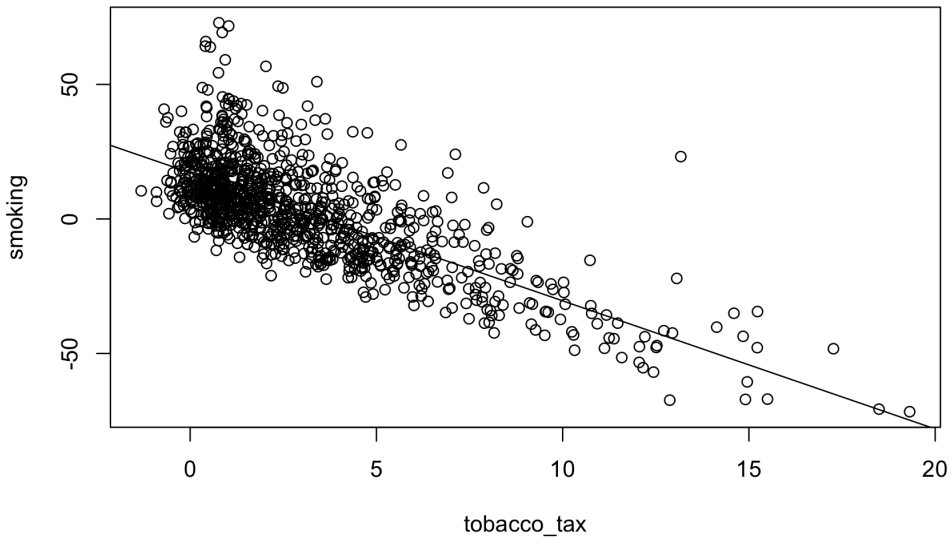
A.3 Two Stage Least Squares

Tobacco tax can likely only be correlated with health through its effect on smoking. If there is a correlation between tobacco tax and health, this may be viewed as evidence that smoking causes changes in health. By using this instrumental variable we can come to directed causal conclusions which can allow us to make decisions which influence this causal chain.

After finding the corresponding data on tobacco tax rates, our scientist checks that it is a proper instrumental variable by showing that it satisfies the following conditions:

1. The IV must be correlated with the endogenous explanatory variables, conditionally on the other covariates.

2. The IV cannot be correlated with the error term in the explanatory equation, conditionally on the other covariates.



Plotting the data and prior domain knowledge that taxes are highly unlikely to be directly correlated with other latent confounders which may affect health gives evidence that tobacco tax rates are a valid instrumental variable.

With this IV, we can run an informed regression with what is called a 2 Stage Least Squares (2SLS) which can allow us to draw causal conclusions with a proper IV. We again have

$$Y = \beta X + \epsilon$$

but now also with instrumental variable Z and error e. First we find the OLS estimator for X using Z with the assumed data generating function to be

$$X = Z\delta + e$$

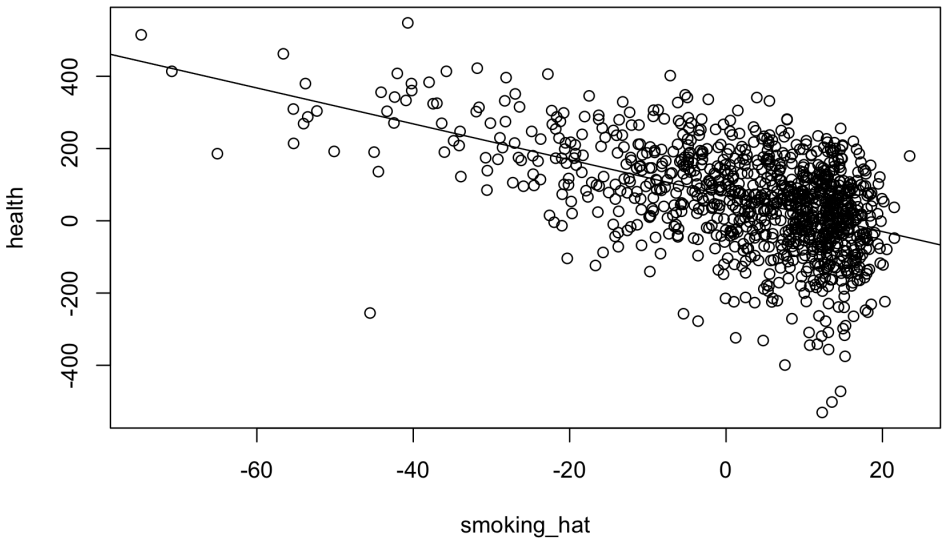
where e is some error and $\hat{\delta} = (Z^T Z)^{-1} Z^T X$ is the standard OLS estimate.

```
Call:
lm(formula = smoking ~ tobacco_tax)

Residuals:
    Min      1Q  Median      3Q     Max
-27.957  -9.278  -2.420   6.996  68.699

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.2185     0.5992   28.73   <2e-16 ***
tobacco_tax  -4.7636     0.1406  -33.88   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 13.52 on 998 degrees of freedom
Multiple R-squared:  0.5349, Adjusted R-squared:  0.5344
F-statistic:  1148 on 1 and 998 DF,  p-value: < 2.2e-16
```

Then, we can use our fit $\hat{\delta}$ to predict $\hat{X} = Z(Z^T Z)^{-1} Z^T X$ and run a second OLS using $\hat{X}$ as a predictor for Y assuming the data generating function $Y = \beta \hat{X} + \epsilon$:

```
smoking_hat = predict(smoking_lm, data.frame(tobacco_tax))

tsls_lm = lm(health ~ smoking_hat)
```




```
Call:
lm(formula = health ~ smoking_hat)

Residuals:
    Min      1Q  Median      3Q     Max
-551.22  -72.01   10.96   82.96  296.81

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  69.5417     3.9364   17.67   <2e-16 ***
smoking_hat  -4.9728     0.2662  -18.68   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 121.9 on 998 degrees of freedom
Multiple R-squared:  0.259,Adjusted R-squared:  0.2583
F-statistic: 348.9 on 1 and 998 DF,  p-value: < 2.2e-16
```

Then, if we have all of our instrumental variable assumptions met, we can measure the causal effect of X on Y because the only way Z can be correlated with Y is through the changes in X. We can also see this mathematically by how $\hat{X} = Z(Z^TZ)^{-1}Z^TX = P_ZX$, which also means that our predictions for health $\hat{Y}$ are purely constructed from the projection of X onto Z: $\hat{Y} = \beta\hat{X} = \beta P_ZX = Z(Z^TZ)^{-1}Z^TX$.

Therefore, overall we have found

Stage 1:

$$X = Z\delta + e$$

$$\hat{\delta} = (Z^TZ)^{-1}Z^TX$$

$$\hat{X} = Z(Z^TZ)^{-1}Z^TX$$

$$\hat{X} = P_ZX$$

Stage 2:

$$Y = \beta\hat{X} + \epsilon$$

$$Y = \beta P_ZX + \epsilon$$

$$\hat{\beta}_{2SLS} = (X^TP_ZX)^{-1}X^TP_ZY$$

We can note that because our ground truth generating function for Y is dependent on both X and U, the estimates for the relationship between X and Y in our simple linear regression is overshooting in magnitude with -5.8746. However, with our 2SLS estimation, we can see the relationship between the $\hat{X}$ and Y is -4.9728 which is much closer to our ground truth coefficient of -5. Thus, we can see 2SLS is a valuable and powerful tool for handling latent confounders if an instrumental variable is selected carefully.