O'REILLY®

Doing Data Science

STRAIGHT TALK FROM THE FRONTLINE

Cathy O'Neil & Rachel Schutt

O'REILLY®

Doing Data Science

Now that people are aware that data can make the difference in an election or a business model, data science as an occupation is gaining ground. But how can you get started working in a wide-ranging, interdisciplinary field that's so clouded in hype? This insightful book, based on Columbia University's Introduction to Data Science class, tells you what you need to know.

In many of these chapter-long lectures, data scientists from companies such as Google, Microsoft, and eBay share new algorithms, methods, and models by presenting case studies and the code they use. If you're familiar with linear algebra, probability, and statistics, and have programming experience, this book is an ideal introduction to data science.

Topics include:

- Statistical inference, exploratory data analysis, and the data science process
- Algorithms
- Spam filters, Naive Bayes, and data wrangling
- Logistic regression
- Financial modeling
- Recommendation engines and causality
- Data visualization
- Social networks and data journalism
- Data engineering, MapReduce, Pregel, and Hadoop

Cathy O'Neil, a senior data scientist at Johnson Research Labs, earned a Ph.D. in math from Harvard, and was a postdoc in the math department at MIT and a professor at Barnard College.

Rachel Schutt, Senior VP of Data Science at News Corp, is an adjunct professor of statistics at Columbia University, and a founding member of CU's Education Committee for the Institute for Data Sciences and Engineering.

DATABASES / DATA

US $39.99 CAN $41.99

ISBN: 978-1-449-35865-5

Twitter: @oreillymedia
facebook.com/oreilly

Doing Data Science

Cathy O'Neil and Rachel Schutt

Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo O'REILLY®

Doing Data Science

by Cathy O'Neil and Rachel Schutt

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://my.safaribooksonline.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Editors: Mike Loukides and Courtney Nash
Production Editor: Kristen Brown
Copyeditor: Kim Cofer
Proofreader: Amanda Kersey

Indexer: WordCo Indexing Services
Cover Designer: Karen Montgomery
Interior Designer: David Futato
Illustrator: Rebecca Demarest

October 2013: First Edition

Revision History for the First Edition:

2013-10-08: First release

2013-12-13: Second release

See *http://oreilly.com/catalog/errata.csp?isbn=9781449358655* for release details.

ISBN: 978-1-449-35865-5

[LSI]

In loving memory of Kelly Feeney.

Table of Contents

CHAPTER 11
Causality

Many of the models and examples in the book so far have been focused on the fundamental problem of prediction. We've discussed examples like in Chapter 8, where your goal was to build a model to predict whether or not a person would be likely to prefer a certain item—a movie or a book, for example. There may be thousands of features that go into the model, and you may use feature selection to narrow those down, but ultimately the model is getting optimized in order to get the highest accuracy. When one is optimizing for accuracy, one doesn't necessarily worry about the *meaning* or *interpretation* of the features, and especially if there are thousands of features, it's well-near impossible to interpret at all.

Additionally, you wouldn't even want to make the statement that certain characteristics *caused* the person to buy the item. So, for example, your model for predicting or recommending a book on Amazon could include a feature "whether or not you've read Wes McKinney's O'Reilly book *Python for Data Analysis*." We wouldn't say that reading his book *caused* you to read *this* book. It just might be a good predictor, which would have been discovered and come out as such in the process of optimizing for accuracy. We wish to emphasize here that it's not simply the familiar correlation-causation trade-off you've perhaps had drilled into your head already, but rather that your *intent* when building such a model or system was not even to understand causality at all, but rather to *predict*. And that if your intent *were* to build a model that helps you get at causality, you would go about that in a different way.

A whole different set of real-world problems that actually use the same statistical methods (logistic regression, linear regression) as part of the

building blocks of the solution are situations where you *do* want to understand causality, when you want to be able to say that a certain type of behavior *causes* a certain outcome. In these cases your mentality or goal is not to optimize for predictive accuracy, but rather to be able to isolate causes.

This chapter will explore the topic of causality, and we have two experts in this area as guest contributors, Ori Stitelman and David Madigan. Madigan's bio will be in the next chapter and requires this chapter as background. We'll start instead with Ori, who is currently a data scientist at Wells Fargo. He got his PhD in biostatistics from UC Berkeley after working at a litigation consulting firm. As part of his job, he needed to create stories from data for experts to testify at trial, and he thus developed what he calls "data intuition" from being exposed to tons of different datasets.

Correlation Doesn't Imply Causation

One of the biggest statistical challenges, from both a theoretical and practical perspective, is establishing a causal relationship between two variables. When does one thing cause another? It's even trickier than it sounds.

Let's say we discover a correlation between ice cream sales and bathing suit sales, which we display by plotting ice cream sales and bathing suit sales over time in Figure 11-1.

This demonstrates a close association between these two variables, but it doesn't establish *causality*. Let's look at this by pretending to know nothing about the situation. All sorts of explanations might work here. Do people find themselves irrestistably drawn toward eating ice cream when they wear bathing suits? Do people change into bathing suits every time they eat ice cream? Or is there some third thing (like hot weather) which we haven't considered that causes both? Causal inference is the field that deals with better understanding the conditions under which association can be interpreted as causality.

Asking Causal Questions

The natural form of a causal question is: What is the effect of x on y?

Some examples are: "What is the effect of *advertising* on *customer behavior*?" or "What is the effect of *drug* on *time until viral failure*?" or in the more general case, "What is the effect of *treatment* on *outcome*?"

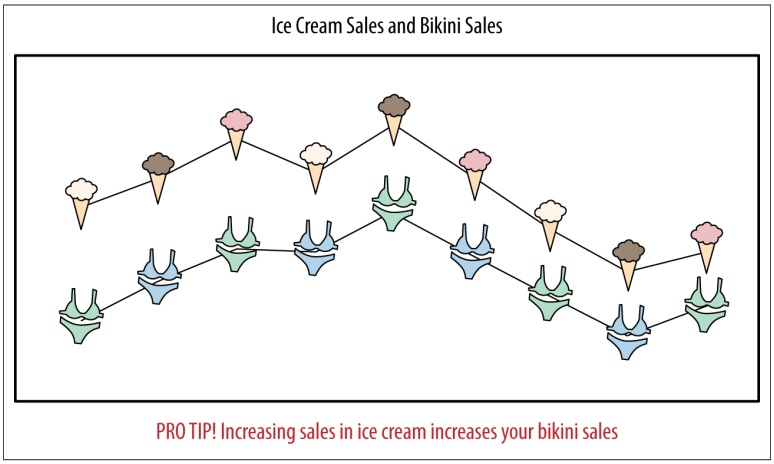

Figure 11-1. Relationship between ice cream sales and bathing suit sales

The terms "treated" and "untreated" come from the biostatistics, medical, and clinical trials realm, where patients are given a medical treatment, examples of which we will encounter in the next chapter. The terminology has been adopted by the statistical and social science literature.

It turns out estimating causal parameters is hard. In fact, the effectiveness of advertising is almost always considered a moot point because it's so hard to measure. People will typically choose metrics of success that are easy to estimate but don't measure what they want, and everyone makes decisions based on them anyway because it's easier. But they have real negative effects. For example, marketers end up being rewarded for selling stuff to people online who would have bought something anyway.

Confounders: A Dating Example

Let's look at an example from the world of online dating involving a lonely guy named Frank. Say Frank is perusing a dating website and comes upon a very desirable woman. He wants to convince her to go out with him on a date, but first he needs to write an email that will get her interested. What should he write in his email to her? Should he tell her she is beautiful? How do we test that with data?

Let's think about a randomized experiment Frank could run. He could select a bunch of beautiful women, and half the time, randomly, tell them they're beautiful. He could then see the difference in response rates between the two groups.

For whatever reason, though, Frank doesn't do this—perhaps he's too much of a romantic—which leaves us to try to work out whether saying a woman is beautiful is a good move for Frank. It's on us to get Frank a date.

If we could, we'd understand the future under two alternative realities: the reality where he sends out the email telling a given woman she's beautiful and the reality where he sends an email but doesn't use the word beautiful. But only one reality is possible. So how can we proceed?

Let's write down our causal question explicitly: what is the effect of Frank telling a woman she's beautiful on him getting a positive response?

In other words, the "treatment" is Frank's telling a woman she's beautiful over email, and the "outcome" is a positive response in an email, or possibly no email at all. An email from Frank that doesn't call the recipient of the email beautiful would be the control for this study.

There are lots of things we're not doing here that we might want to try. For example, we're not thinking about Frank's attributes. Maybe he's a really weird unattractive guy that no woman would want to date no matter what he says, which would make this a tough question to solve. Maybe he can't even spell "beautiful." Conversely, what if he's gorgeous and/or famous and it doesn't matter what he says? Also, most dating sites allow women to contact men just as easily as men contact women, so it's not clear that our definitions of "treated" and "untreated" are well-defined. Some women might ignore their emails but spontaneously email Frank anyway.

OK Cupid's Attempt

As a first pass at understanding the impact of word choice on response rates, the online dating site OK Cupid analyzed over 500,000 first contacts on its site. They looked at keywords and phrases, and how they affected reply rates, shown in Figure 11-2.

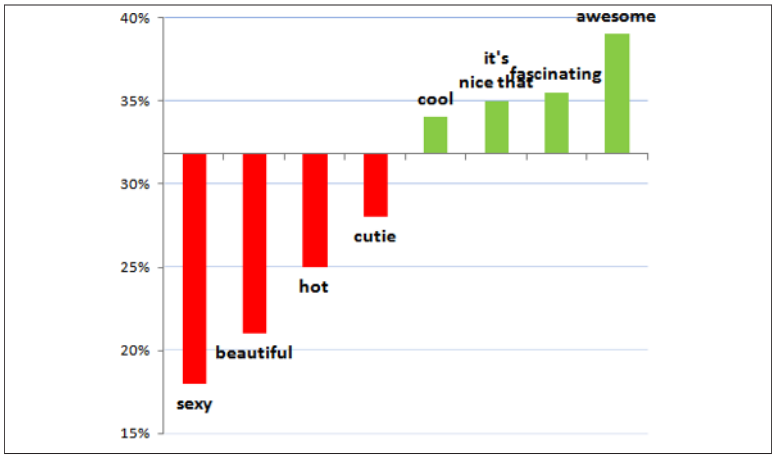

Figure 11-2. OK Cupid's attempt to demonstrate that using the word "beautiful" in an email hurts your chances of getting a response

The y-axis shows the response rate. On average the response rate across *all* emails was ~32%. They then took the subset of emails that included a certain word such as "beautiful" or "awesome," and looked at the response rate for those emails. Writing this in terms of conditional probabilities, we would say they were estimating these: P(response) = 0.32 vs P(response|"beautiful") = 0.22.

One important piece of information missing in this plot is the bucket sizes. How many first contacts contained each of the words? It doesn't really change things, except it would help in making it clear that the horizontal line at 32% is a weighted average across these various buckets of emails.

They interpreted this graph and created a rule called "Avoid Physical Compliments." They discussed this in the blog post "Exactly what to say on a first message" with the following explanation: "You might think that words like gorgeous, beautiful, and sexy are nice things to say to someone, but no one wants to hear them. As we all know, people normally like compliments, but when they're used as pick-up lines, before you've even met in person, they inevitably feel… ew. Besides, when you tell a woman she's beautiful, chances are you're not."

This isn't an experiment, but rather an *observational study*, which we'll discuss more later but for now means we collect data as it naturally occurs in the wild. Is it reasonable to conclude from looking at this plot that adding "awesome" to an email increases the response rate, or that "beautiful" decreases the response rate?

Before you answer that, consider the following three things.

First, it could say more about the *person* who says "beautiful" than the word itself. Maybe they are otherwise ridiculous and overly sappy? Second, people may be describing *themselves* as beautiful, or some third thing like the world we live in.

These are both important issues when we try to understand population-wide data such as in the figure, because they address the question of whether having the word "beautiful" in the body of the email actually implies what we think it does. But note that both of those issues, if present, are consistently true for a given dude like Frank trying to get a date. So if Frank is sappy, he's theoretically equally sappy to all the women he writes to, which makes it a consistent experiment, from his perspective, to decide whether or not to use the word "beautiful" in his emails.

The third and most important issue to consider, because it does *not* stay consistent for a given dude, is that *the specific recipients of emails containing the word "beautiful" might be special*: for example, they might get tons of email, and only respond to a few of them, which would make it less likely for Frank to get any response at all.

In fact, if the woman in question is beautiful (let's pretend that's a well-defined term), that fact affects two separate things at the same time. Both whether Frank uses the word "beautiful" or not in his email, and the outcome, i.e., whether Frank gets a response. For this reason, the fact that the woman is beautiful qualifies as a *confounder*; in other words, a variable that influences or has a causal effect on both the treatment itself as well as the outcome.

Let's be honest about what this plot *actually* shows versus what OK Cupid was implying it showed. It shows the observed response rate for emails that contained the given words. It should *not* be used and cannot correctly be interpreted as a prescription or suggestion for how to construct an email to get a response because after *adjusting* for confounders, which we'll discuss later in the chapter, using the word "beautiful" could be the best thing we could do. We can't say for sure

because we don't have the data, but we'll describe what data we'd need and how we'd analyze it to do this study properly. Their advice might be correct, but the plot they showed does not back up this advice.

The Gold Standard: Randomized Clinical Trials

So what do we do? How do people *ever* determine causality?

The gold standard for establishing causality is the randomized experiment. This a setup whereby we randomly assign some group of people to receive a "treatment" and others to be in the "control" group —that is, they *don't* receive the treatment. We then have some outcome that we want to measure, and the causal effect is simply the difference between the treatment and control group in that measurable outcome. The notion of using experiments to estimate causal effects rests on the statistical assumption that using randomization to select two groups has created "identical" populations from a statistical point of view.

Randomization works really well: because we're flipping coins, all other factors that might be confounders (current or former smoker, say) are more or less removed, because we can guarantee that smokers will be fairly evenly distributed between the two groups if there are enough people in the study.

The truly brilliant thing about randomization is that randomization matches well on the possible confounders we thought of, but will also give us balance on the *50 million things we didn't think of*.

So, although we can algorithmically find a better split for the ones we thought of, that quite possibly wouldn't do as well on the other things. That's why we really do it randomly, because it does quite well on things we think of and things we don't.

But there's bad news for randomized clinical trials as well, as we pointed out earlier. First off, it's only ethically feasible if there's something called clinical equipoise, which means the medical community really doesn't know which treatment is better. If we know treating someone with a drug will be better for them than giving them nothing, we can't randomly not give people the drug.

For example, if we want to tease out the relationship between smoking and heart disease, we can't randomly assign someone to smoke, because it's known to be dangerous. Similarly, the relationship between

cocaine and birthweight is fraught with danger, as is the tricky relationship between diet and mortality.

The other problem is that they are expensive and cumbersome. It takes a long time and lots of people to make a randomized clinical trial work. On the other hand, not doing randomized clinical trials can lead to mistaken assumptions that are extremely expensive as well.

Sometimes randomized studies are just plain unfeasible. Let's go back to our OK Cupid example, where we have a set of observational data and we have a good reason to believe there are confounders that are screwing up our understanding of the effect size. As noted, the gold standard would be to run an experiment, and while the OK Cupid employees *could* potentially run an experiment, it would be unwise for them to do so—randomly sending email to people telling them they are "beautiful" would violate their agreement with their customers.

In conclusion, *when they are possible*, randomized clinical trials are the gold standard for elucidating cause-and-effect relationships. It's just that they aren't always possible.

Average Versus the Individual

Randomized clinical trials measure the effect of a certain drug averaged across all people. Sometimes they might bucket users to figure out the average effect on men or women or people of a certain age, and so on. But in the end, it still has averaged out stuff so that for a given individual we don't know what they effect would be on them. There is a push these days toward personalized medicine with the availability of genetic data, which means we stop looking at averages because we want to make inferences about the one. Even when we were talking about Frank and OK Cupid, there's a difference between conducting this study across all men versus Frank alone.

A/B Tests

In software companies, what we described as random experiments are sometimes referred to as A/B tests. In fact, we found that if we said the word "experiments" to software engineers, it implied to them "trying something new" and not necessarily the underlying statistical design of having users experience different versions of the product in order to measure the impact of that difference using metrics. The concept is

intuitive enough and seems simple. In fact, if we set up the infrastructure properly, running an experiment can come down to writing a short configuration file and changing just one parameter—be it a different color or layout or underlying algorithm—that gives some users a different experience than others. So, there are aspects of running A/B tests in a tech company that make it much easier than in a clinical trial. And there's much less at stake in terms of the idea that we're not dealing with people's lives. Other convenient things are there aren't compliance issues, so with random clinical trials we can't control whether someone takes the drug or not, whereas online, we can control what we show the user. But notice we said *if* we set up the experimental infrastructure properly, and that's a big IF.

It takes a lot of work to set it up well and then to properly analyze the data. When different teams at a company are all working on new features of a product and all want to try out variations, then if you're not careful a single user could end up experiencing multiple changes at once. For example, the UX team might change the color or size of the font, or the layout to see if that increases click-through rate. While at the same time the content ranking team might want to change the algorithm that chooses what to recommend to users, and the ads team might be making changes to their bidding system. Suppose the metric you care about is return rate, and a user starts coming back more, and you had them in three different treatments but you didn't know that because the teams weren't coordinating with each other. Your team might assume the treatment is the reason the user is coming back more, but it might be the combination of all three.

There are various aspects of an experimental infrastructure that you need to consider, which are described in much more detail in Overlapping Experiment Infrastructure: More, Better, Faster Experimentation, a 2010 paper by Google employees Diane Tang, et al. See the following sidebar for an excerpt from this paper.

From "Overlapping Experiment Infrastructure: More, Better, Faster Experimentation"

The design goals for our experiment infrastructure are therefore: more, better, faster.

More
: We need scalability to run more experiments simultaneously. However, we also need flexibility: different experiments need different configurations and different sizes to be able to measure statistically significant effects. Some experiments only need to change a subset of traffic, say Japanese traffic only, and need to be sized appropriately. Other experiments may change all traffic and produce a large change in metrics, and so can be run on less traffic.

Better
: Invalid experiments should not be allowed run on live traffic. Valid but bad experiments (e.g., buggy or unintentionally producing really poor results) should be caught quickly and disabled. Standardized metrics should be easily available for all experiments so that experiment comparisons are fair: two experimenters should use the same filters to remove robot traffic when calculating a metric such as CTR.

Faster
: It should be easy and quick to set up an experiment; easy enough that a non-engineer can do so without writing any code. Metrics should be available quickly so that experiments can be evaluated quickly. Simple iterations should be quick to do. Ideally, the system should not just support experiments, but also controlled ramp-ups, i.e., gradually ramping up a change to all traffic in a systematic and well-understood way.

That experimental infrastructure has a large team working on it and analyzing the results of the experiments on a full-time basis, so this is nontrivial. To make matters more complicated, now that we're in an age of social networks, we can no longer assume that users are independent (which is part of the randomization assumption underlying experiments). So, for example, Rachel might be in the treatment group of an experiment Facebook is running (which is impossible because Rachel isn't actually on Facebook, but just pretend), which lets Rachel

post some special magic kind of post, and Cathy might be in the control group, but she still sees the special magic post, so she actually received a different version of the treatment, so the experimental design must take into account the underlying network structure. This is a nontrivial problem and still an open research area.

Second Best: Observational Studies

While the gold standard is generally understood to be randomized experiments or A/B testing, they might not always be possible, so we sometimes go with second best, namely observational studies.

Let's start with a definition:

> An observational study is an empirical study in which the objective is to elucidate cause-and-effect relationships in which it is not feasible to use controlled experimentation.

Most data science activity revolves around observational data, although A/B tests, as you saw earlier, are exceptions to that rule. Most of the time, the data you have is what you get. You don't get to replay a day on the market where Romney won the presidency, for example.

Designed studies are almost always theoretically better tests, as we know, but there are plenty of examples where it's unethical to run them. Observational studies are done in contexts in which you can't do designed experiments, in order to elucidate cause-and-effect.

In reality, sometimes you don't care about cause-and-effect; you just want to build predictive models. Even so, there are many core issues in common with the two.

Simpson's Paradox

There are all kinds of pitfalls with observational studies.

For example, look at the graph in Figure 11-3, where you're finding a best-fit line to describe whether taking higher doses of the "bad drug" is correlated to higher probability of a heart attack.

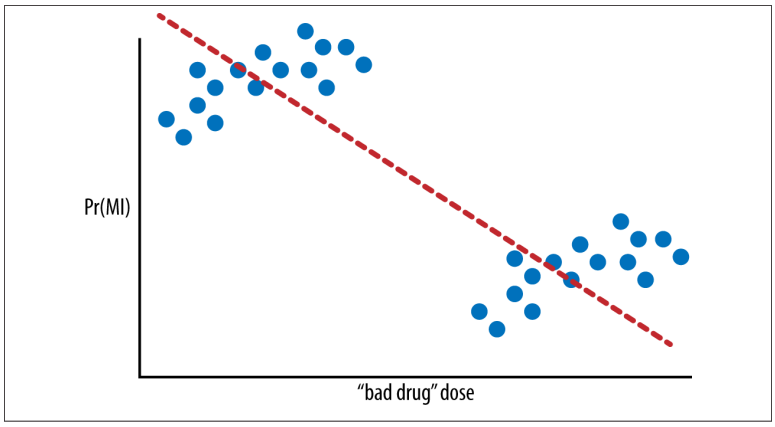

Figure 11-3. Probability of having a heart attack (also known as MI, or myocardial infarction) as a function of the size of the dose of a bad drug

It looks like, from this vantage point, the higher the dose, the fewer heart attacks the patient has. But there are two clusters, and if you know more about those two clusters, you find the opposite conclusion, as you can see in Figure 11-4.

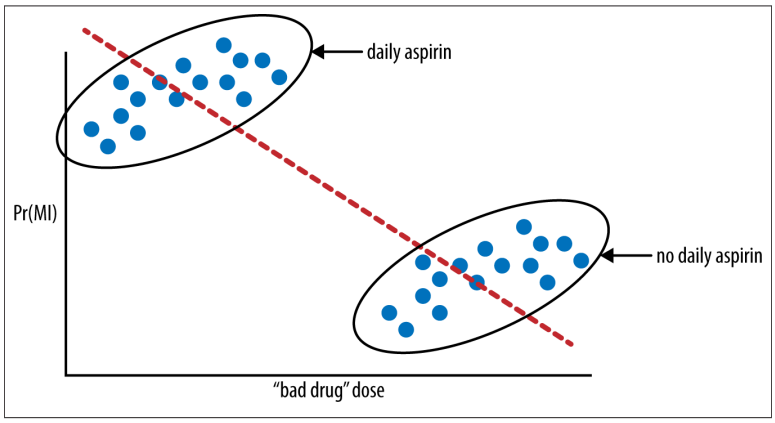

Figure 11-4. Probability of having a heart attack as a function of the size of the dose of a bad drug and whether or not the patient also took aspirin

This picture was rigged, so the issue is obvious. But, of course, when the data is multidimensional, you wouldn't even always draw such a simple picture.

In this example, we'd say aspirin-taking is a confounder. We'll talk more about this in a bit, but for now we're saying that the aspirin-taking or nonaspirin-taking of the people in the study wasn't randomly distributed among the people, and it made a huge difference in the apparent effect of the drug.

Note that, if you think of the original line as a predictive model, it's actually *still* the best model you can obtain knowing nothing more about the aspirin-taking habits or genders of the patients involved. The issue here is really that you're trying to assign causality.

It's a general problem with regression models on observational data. You have no idea what's going on. As Madigan described it, "it's the Wild West out there."

It could be the case that within each group there are males and females, and if you partition by *those*, you see that the more drugs they take, the better again. Because a given person either is male or female, and either takes aspirin or doesn't, this kind of thing really matters.

This illustrates the fundamental problem in observational studies: a trend that appears in different groups of data disappears when these groups are combined, or vice versa. This is sometimes called Simpson's Paradox.

The Rubin Causal Model

The Rubin causal model is a mathematical framework for understanding what information we know and don't know in observational studies.

It's meant to investigate the confusion when someone says something like, "I got lung cancer because I smoked." Is that true? If so, you'd have to be able to support the statement, "If I hadn't smoked, I wouldn't have gotten lung cancer," but nobody knows that for sure.

Define Z_i to be the treatment applied to unit i (0 = control, 1= treatment), $Y_i(1)$ to be the response for unit i if $Z_i = 1$ and $Y_i(0)$ to be the response for unit i if $Z_i = 0$.

Then the *unit level causal effect*, the thing we care about, is $Y_i(1) - Y_i(0)$, but we only see one of $Y_i(0)$ and $Y_i(1)$.

Example: Z_i is 1 if I smoked, 0 if I didn't (I am the unit). $Y_i(1)$ is 1 if I got cancer and I smoked, and 0 if I smoked and didn't get cancer. Similarly $Y_i(0)$ is 1 or 0, depending on whether I got cancer while not smoking. The overall causal effect on me is the difference $Y_i(1) - Y_i(0)$. This is equal to 1 if I really got cancer because I smoked, it's 0 if I got cancer (or didn't) independent of smoking, and it's –1 if I avoided cancer by smoking. But I'll never know my actual value because I only know one term out of the two.

On a population level, we do know how to infer that there are quite a few "1"s among the population, but *we will never be able to assign a given individual that number*.

This is sometimes called the fundamental problem of causal inference.

Visualizing Causality

We can represent the concepts of causal modeling using what is called a *causal graph*.

Denote by W the set of all potential confounders. Note it's a big assumption that we can take account of all of them, and we will soon see how unreasonable this seems to be in epidemiology research in the next chapter.

In our example with Frank, we have singled out one thing as a potential confounder—the woman he's interested in being beautiful—but if we thought about it more we might come up with other confounders, such as whether Frank is himself attractive, or whether he's desperate, both of which affect how he writes to women as well as whether they respond positively to him.

Denote by A the treatment. In our case the treatment is Frank's using the word "beautiful" in an introductory email. We usually assume this to have a binary (0/1) status, so for a given woman Frank writes to, we'd assign her a "1" if Frank uses the word "beautiful." Just keep in mind that if he says it's beautiful weather, we'd be measuring counting that as a "1" even though we're thinking about him calling the woman beautiful.

Denote by Y the binary (0/1) outcome. We'd have to make this well-defined, so, for example, we can make sure Frank asks the women he writes to for their phone number, and we could define a positive outcome, denoted by "1," as Frank getting the number. We'd need to make this as precise as possible, so, for example, we'd say it has to happen in

the OK Cupid platform within a week of Frank's original email. Note we'd be giving a "1" to women who ignore his emails but for some separate reason send him an email with their number. It would also be hard to check that the number isn't fake.

The nodes in a causal graph are labeled by these sets of confounders, treatment, and outcome, and the directed edges, or arrows, indicate causality. In other words, the node the arrow is coming out of in some way directly affects the node the arrow is going into.

In our case we have Figure 11-5.

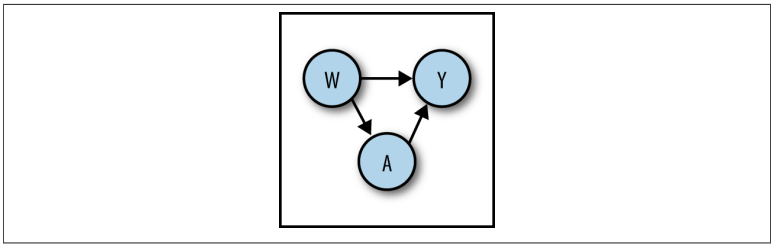

Figure 11-5. Causal graph with one treatment, one confounder, and one outcome

In the case of the OK Cupid example, the causal graph is the simplest possible causal graph: one treatment, one confounder, and one outcome. But they can get much more complicated.

Definition: The Causal Effect

Let's say we have a population of 100 people that take some drug, and we screen them for cancer. Say 30 of them get cancer, which gives them a cancer rate of 0.30. We want to ask the question, did the drug cause the cancer?

To answer that, we'd have to know what would've happened if they hadn't taken the drug. Let's play God and stipulate that, had they not taken the drug, we would have seen 20 get cancer, so a rate of 0.20. We typically measure the increased risk of cancer as the difference of these two numbers, and we call it the *causal effect*. So in this case, we'd say the causal effect is 10%.

The *causal effect* is sometimes defined as the *ratio* of these two numbers instead of the difference.

But we don't have God's knowledge, so instead we choose another population to compare this one to, and we see whether *they* get cancer or not, while *not* taking the drug. Say they have a natural cancer rate of 0.10. Then we would conclude, using them as a proxy, that the increased cancer rate is the difference between 0.30 and 0.10, so 20%. This is of course wrong, but the problem is that the two populations have some underlying differences that we don't account for.

If these were the "same people," down to the chemical makeup of each others' molecules, this proxy calculation would work perfectly. But of course they're not.

So how do we actually select these people? One technique is to use what is called propensity score matching or modeling. Essentially what we're doing here is creating a pseudo-random experiment by creating a synthetic control group by selecting people who were *just as likely* to have been in the treatment group but weren't. How do we do this? See the word in that sentence, "likely"? Time to break out the logistic regression. So there are two stages to doing propensity score modeling. The first stage is to use logistic regression to model the probability of each person's likelihood to have *received the treatment*; we then might pair people up so that one person received the treatment and the other didn't, but they had been *equally likely* (or close to equally likely) to have received it. Then we can proceed as we would if we had a random experiment on our hands.

For example, if we wanted to measure the effect of smoking on the probability of lung cancer, we'd have to find people who shared the same probability of *smoking*. We'd collect as many covariates of people as we could (age, whether or not their parents smoked, whether or not their spouses smoked, weight, diet, exercise, hours a week they work, blood test results), and we'd use as an outcome whether or not they smoked. We'd build a logistic regression that predicted the probability of smoking. We'd then use that model to assign to each person the probability, which would be called their propensity score, and then we'd use that to match. Of course we're banking on the fact that we *figured out* and were able to observe *all* the covariates associated with likelihood of smoking, which we're probably not. And that's the

inherent difficulty in these methods: we'll never know if we actually adjusted for everything we needed to adjust for. *However*, one of the nice aspects of them is we'll see that when we do adjust for confounders, it can make a big difference in the estimated causal effect.

The details of setting of matching can be slightly more complicated than just paired matching—there are more complex schemes to try to create balance in the synthetic treatment and control group. And there are packages in R that can do it all automatically for you, except you must specify the model that you want to use for matching in the first place to generate the propensity scores, and which variable you want to be the outcome corresponding to the causal effect you are estimating.

What kind of data would we need to measure the causal effect in our dating example? One possibility is to have some third party, a mechanical turk, for example, go through the dating profiles of the women that Frank emails and label the ones that are beautiful. That way we could see to what extent being beautiful is a confounder. This approach is called *stratification* and, as we will see in the next chapter, it can introduce problems as well as fix them.

Three Pieces of Advice

Ori took a moment to give three pieces of parting advice for best practices when modeling.

First, when estimating causal parameters, it is crucial to understand the data-generating methods and distributions, which will in turn involve gaining some subject matter knowledge. Knowing exactly how the data was generated will also help you ascertain whether the assumptions you make are reasonable.

Second, the first step in a data analysis should always be to take a step back and figure out *what you want to know*. Write it down carefully, and then find and use the tools you've learned to answer those directly. Later on be sure and come back to decide how close you came to answering your original question or questions. Sounds obvious, but you'd be surprised how often people forget to do this.

Finally, don't ignore the necessary data intuition when you make use of algorithms. Just because your method converges, it doesn't mean the results are meaningful. Make sure you've created a reasonable narrative and ways to check its validity.