CHAPTER 18: CORRELATIONS ARE HARD TO INTERPRET

INTERPRETATION AND CONCLUSIONS

18

In his essay The Danger of Lying in Bed, Mark Twain made folly of people who bought travel insurance. He pointed out that far more people died in bed than on public transportation, so the REAL danger came from lying down.

SECTION 1

Introduction

In most scientific inquiries, we seek the cause of something. We want to know what causes cancer, what drugs cause us to recover from disease or to feel less pain, what cultural practices cause environmental problems, what business practices lead to (cause) increased profits, what kind of sales pitch increases sales, what kind of resume is the most effective in getting a job, and so on. In these cases, we are testing causal models. Not everything we've discussed so far requires evaluation of a causal model: in DNA and drug testing, we are merely trying to measure properties of an individual (a drug level, a DNA bar code). But these exceptions notwithstanding, the most common kind of evaluation everyone encounters is testing of a causal model. "What can we change in our lives or our world to cause a certain outcome?" is the essence of what we want in a causal model.

Causal models are typically evaluated, at least initially, with data that describe an association or correlation between variables. If smoking causes lung cancer, then cancer rates should be higher (associated) with smokers. If some patterns of investment lead to higher profits, then companies which practice those kinds of investment ought to be associated with greater returns to their investors. If alcohol causes reckless driving, then a higher rate of accidents should be associated with drunk driving. The catch is this. Although a causal relationship between 2 sets of data leads to an association between them (drinking and driver accidents), an association may occur even when there is no causation. How then do we decide if the causal model is supported or refuted? This chapter is about associations among variables -- correlations -- and how and when we can tease out causation.

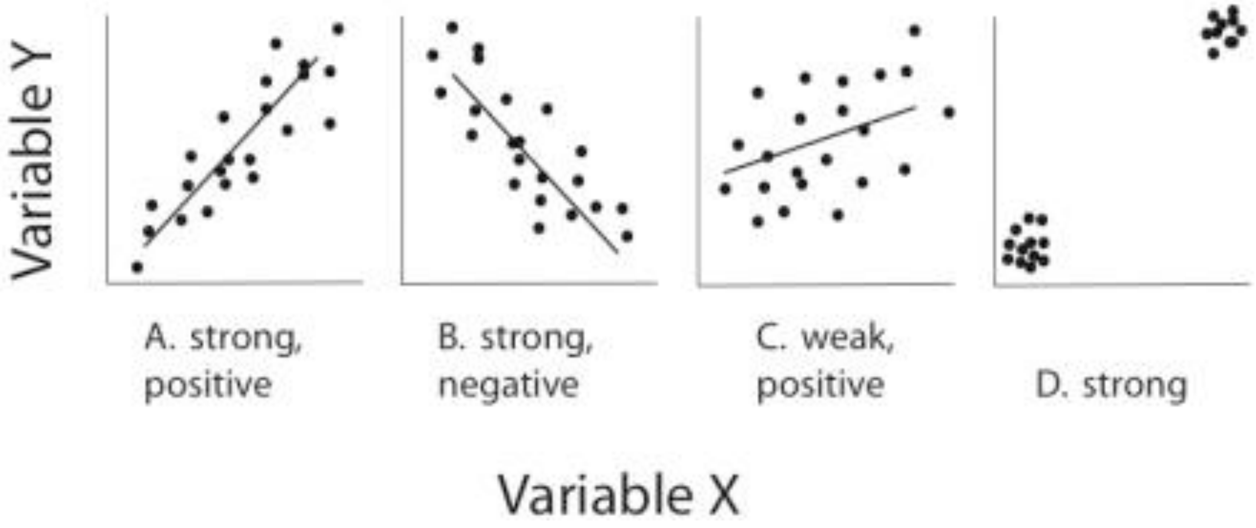
Recall from an earlier chapter that epidemiologists in Britain noted a higher incidence of cancers in young people living near nuclear power plants than in the population at large. These data pointed to a possible environmental hazard of the nuclear power plants -- perhaps the power plants were causing excess cancers. However, the fact that excess cancers were also found in proposed sites that still lacked power plants suggested that the power plants were not the cause of excess cancers. This example is typical of the problems that often arise from a failure to appreciate the limitations of correlations.

SECTION 2

What Are Correlations?

Correlations are associations between variables. The first question to answer in understanding a correlation is therefore "What are variables?" Variables are things we measure that can differ from one observation to the next, such as height, weight, behavior, fat intake, life-span, grade-point average, and income. With these variables we can easily assign a number to represent the value of the variable. Perhaps less obviously, we can also treat sex (gender), country of origin, and political preference as variables, even though we don't know how to assign a number to represent each category. In general, a variable is a measure of something that can take on more than one value. It is somewhat arbitrary how we define a variable, but in general, you must be able to put the different values a variable can take onto a single axis of a graph. If you are wonder whether something you have defined is a variable and it would require two axes, then you are likely dealing with a couple of variables combined.

When an association exists between two variables, it means that the average value of one variable changes as we change the value of the other variable (Fig. 18.1). A correlation is the simplest type of association -- linear. When a correlation is weak (e.g., Model C), it means that the average value of one variable changes only slightly (only occasionally) in response to changes in the other variable. In some cases, the correlation may be positive (Models A, C), or it may be negative (Model B). If the points in such a graph pretty much fall inside a circle or horizontal ellipse such that the "trend-line" through them is horizontal, then a correlation does not exist (the same as a zero or no correlation). When either or both variables cannot be assigned numbers (e.g., political party or country of origin), a correlation may still exist but we no longer apply the terms positive and negative (e.g., Model D, depending on the nature of the variables). Since a correlation is an association among variables, a correlation cannot exist (is not defined) with just one variable; "undefined" is not the same as a zero correlation or no correlation. A graph of points with only one variable would have all points on a perfectly horizontal line or a perfectly vertical line (with no scatter around the line).



Different kinds of correlations:

The horizontal axis represents one variable (X) and the vertical axis represents a different variable (Y), with values of X and Y increasing according to the distance from the origin. Models A, B & C show correlations for continuous variables which can take on a range of values (e.g., height, weight), whereas Model D reveals a correlation for discrete variables (variable X might be gender, variable Y presence or absence of the Y chromosome). Model A reveals a strong positive correlation, Model B a strong negative correlation, and Model C a weak positive correlation. The correlation in Model D would be regarded as positive if values could be assigned to X and Y, but if values cannot be assigned (e.g., gender and presence of Y chromosome), we would not refer to the correlation as being positive or negative.

Correlations are common in Business . Businesses often obtain large quantities of correlational data as they go about their activities (Table 18.1). An insurance company in the course of doing business obtains data about which types of customers are more often involved in accidents. These data are purely observational - the company can't force a 68 year old grandmother to drive a pickup if she doesn't want to. The data consist of driver age, sex, make and model of car, zip code, street address and so forth. In addition, the company knows how many and the type of accidents for each customer. These correlations are clearly quite useful in predicting what customers will have more accidents.

Correlated variables having substantial impact on profits and losses or on the efficiency of government operations:

INSURANCE
accident rate vs. age, and sex of driver, make and year of car hurricane frequency vs. city death rate vs. age and sex
FINANCE
personal loan default rate vs. age, gender, and income of borrower corporation bond default rate vs. Moody's rating of the bond
RETAIL SALES
total sales vs. day of the week customer's name vs. product brand, amount, and dollar value of items sold
TRANSPORTATION AND COMMUNICATION
volume of mail vs. city express mail deliveries vs. zip code profits vs. route
MANUFACTURING
steel mill profits vs. type of order and type of ingots and coke used
GOVERNMENT
parolee recidivism rate vs. age, sex, family status, crime committed

Correlations are used to manipulate us. Most advertisements, sales pitches, and political speeches invoke correlations to influence our behavior. A company tends to display its product in favorable settings to build an imaginary correlation between its product and the desirable surroundings (e.g., beer commercials using attractive members of the opposite sex, 4WD autos being pictured with a backdrop of remote, montane scenery). Negative campaigning usually involves describing some unfavorable outcome that occurred during an opponent's tenure in office to develop a correlation in the viewer's mind between the candidate and bad consequences of their election to office.

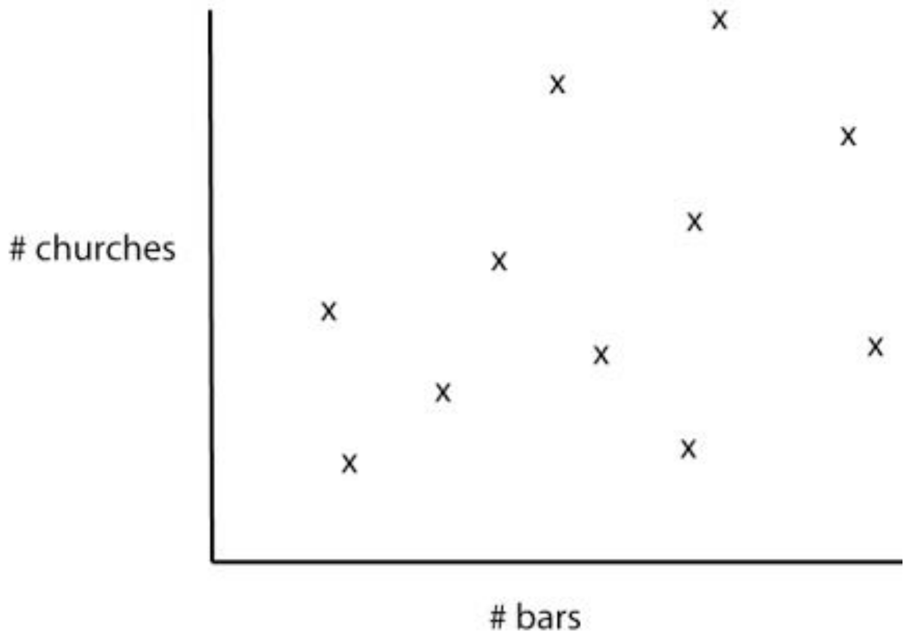
The reason that correlations are used so often in commercials is that they work-- people make the causal extrapolation from correlations. We tend to blame our current president for many social problems, even though the president has little control over many of them. In a well-known but unfortunate psychological experiment of some decades ago, a child was encouraged to develop a close attachment to a white rat, whereupon the experimenters intentionally frightened the child with the rat. Thereafter, the child avoided white objects -- a rather surprising correlate of the rat. Other studies have shown that people respond differently to an item of clothing according to what they are told about an imaginary person who wore it: the response is more favorable if the supposed previous wearer is famous than if the person is infamous. The information thus established a correlation between the clothing and a desirable or undesirable person, and the subjects mentally extrapolated that correlation to some kind of causation of good or bad from wearing the object. And some of our responses to correlations are very powerful. The experience of getting overly drunk on one kind of alcoholic beverage is often enough to cause a person to avoid that beverage years into the future but not to avoid other kinds of alcoholic beverages.

Negative uses. A more negative context for the application of correlation to influence behavior is the practice known as character assassination. A person can be denigrated in one aspect of their life by identifying an unfavorable characteristic in some other (and perhaps trivial) aspect of their life We automatically extrapolate the negative correlation to them as a whole.

The Problem With Correlations: Hidden Variables

The problem that underlies evaluation of correlations is extremely common in science. We observe an association, or correlation, between two or more variables. In the nuclear power plant example, there is a correlation between residential proximity to a nuclear plant and cancer, because people near power plants are more likely to get cancer than those who live away from power plants. And we try to infer the causation from that correlation (does the plant actually cause cancer?). Time and again, science has learned the hard way that we cannot infer causation from correlation: **correlation does not imply causation.**

What does this mean? Say that you observe a correlation between smoking and lung cancer. To infer that smoking CAUSES lung cancer, you would argue that people should stop smoking to lower their lung cancer rates. If smoking does not cause lung cancer, however, then stopping smoking would actually have no effect on lung cancer rates (we are very confident, however, that smoking causes lung cancer). How can a correlation not reflect causation? Consider a plot of the number of churches in a town (city) and the number of bars in a town:



This is drawn so that there tend to be more churches than bars in a town, but as the number of churches increases, so does the number of bars on average. Although these data were made up for illustration, the correlation is almost certainly true. To argue causation from these data, we would either have to say that churches cause people to drink more (whether intentionally or unintentionally), or argue that lots of drinkers in a town causes more churches to be built (e.g., churches move in where there are sinners). Furthermore, causation would suggest either that banning bars would reduce the number of churches in the town, or that the way to cut down on the number of bars was to close down churches (depending on which way the causation went). In reality, the correlation is due to a hidden variable -- population size. That is, larger towns have more demand for churches and for bars, as well as other social institutions.

To reiterate the theme of this chapter, the major difficulty with all correlations is that there are many models consistent with any correlation: the correlation between two variables may be caused by a third, fourth, or dozens of variables other than the two being compared. Thus we are left with countless alternative models in addition to the obvious ones. For example, we initially think that the correlation between cancer and residence near a power plant shows that nuclear power plants cause cancer. Then we learn that another factor, site of the power plant, may be important. It appears that the important factor is not the power plant itself, but rather some characteristic of sites chosen for power plants (one obvious possibility is that nuclear power plants are situated in low income areas that have higher cancer rates than suffered by the general population). That is, there are correlations between all sorts of other variables besides just residence and cancer.

There are many issues in society that hinge on correlations (Table 18.2). In some cases, a correlation may identify a causal relationship, such as health defects being caused by environmental toxins. Yet because the correlational data don't reject countless alternatives models, no action is taken to correct the problem. In other cases, a correlation may be assumed to reflect the cause when it does not.

Public policy issues that involve understanding the cause of a correlation:

ISSUE	POSSIBLE CAUSATION
High cancer incidence near industrial sites, toxic waste dumps, nuclear power plants.	If the increased cancer rate is actually caused by the hazard, there would be compelling motivation for taking action. But it is often difficult to rule out the alternative explanation that those living near the hazard have different diets or for other reasons are more susceptible to cancer than the general population.
Racial differences in standardized test scores.	There are two opposing positions in this acrimonious debate: i) a person's race, per se, causes them to have low test scores, or ii) minorities often have low incomes, and it is income rather than race that determines test score. The first explanation states that a person is born with a certain intellectual ability, the second states that they acquire it.

Section 4

Correlations Complicate Studying Diet and Heart Disease

The medical news over the last decade or so has been obsessed with the relationship between diet and heart disease. (Heart disease is chiefly the build-up of deposits inside blood vessels, hardening the arteries and enabling the vessels to rupture and clog.) A report that dietary fiber lowered heart attack risk led to an avalanche of pills and breakfast cereals high in fiber. More recently, a trendy topic has been iron levels in the blood. It is not clear what to make of these reports, but we can be confident that associations between diet and heart disease will continue to be the subject of studies for decades to come. However, let's consider the problems such studies pose.

Your diet consists of literally hundreds of correlated components. For example, people who eat a lot of meat also tend to eat a lot of fat, and people that eat lots of vitamin C tend to also eat much fiber. These, and numerous similar correlations, create huge problems in determining what diet you should eat to avoid heart disease. A study that found an correlation between heart disease and fat, for example, would be hard to interpret because we would not know if it was the fat, per se, or the meat that was the problem. The problem in this example is not as great as it is in other cases, because we can actually conduct experiments with human diets to explore causal relationships. But even in these experiments, it is difficult to control and randomize all relevant factors.

Section 5

Do Electromagnetic Fields Cause Cancer?

Beginning in the 1960's and the 1970's, evidence arose that intense electromagnetic fields (EMFs) could influence behavior and physiology. No study was particularly conclusive. In all cases, the fields were intense and effects seemed reversible, and the concern was neither about cancer nor about effects from fields of low intensity such as those in the typical neighborhood. But in 1979, epidemiologists Nancy Wertheimer & Ed Leeper reported that childhood leukemia rate in Denver was higher for dwellings "near" a transformer than for dwellings away from a transformer. The result was incredible because it suggested that many of us are exposed to a cancer risk in our own dwellings. There have been at least 6 attempts to repeat Wertheimer and Leeper's epidemiological correlations, and the overall trend continues to be born out (with some inconsistencies); studies appear maybe a couple of times a year now. Overall, it appears that there is a slightly elevated risk of leukemia associated with living near high current transformers and the wires that emanate from them (the risk factor is 1-2). The baseline rate for childhood leukemia is about 1/20,000, so the EMF risk raises it to 1/10,000.

Once public awareness had been elevated by these original studies, there was a plethora of anecdotal and post-hoc observations that highlighted incidents in which EMFs might be causing harm. The news was filled with a cluster of miscarriages in women working at CRT's (cathode ray tubes – the computer monitors in the days before flat screens), the news carried stories of people with cell phones who got brain cancer, and so on. A study of a NY telephone company made an attempt to determine if there was a correlation between cancer and occupations which had varying exposure to EMFs, in the hope of showing that more cancers were found with higher doses (Table 1).

Cancer incidence vs. exposure to electromagnetic fields:

OCCUPATION	RELATIVE EXPOSURE	CANCER
cable splicers	highest	2X overall cancer rate
central office	next highest	3X prostrate cancer; 2X oral; some male breast cancer
other	lowest	nothing of particular note

The occupations were ranked according to exposure and the cancer incidence showed some hint of a dose-response. However, this was the only study (of many) showing a possible dose response effect, and even in this case, the results present a heterogeneous array of cancers.

Reasons for Being Skeptical

In determining whether electromagnetic fields might cause cancer, it is reasonable to compare EMFs to a form of radiation that does cause cancer – ionizing radiation. To compare ionizing radiation with household EMFs, one needs to consider the energy and intensity of EMFs. Electromagnetic fields from alternating current are low in energy. The energy of electromagnetic radiation increases with the frequency of radiation. Alternating current cycles at 60 times/second, so its frequency is 60 cycles/second. Visible light has a frequency of 1014 cycles/second, UV light has a frequency of 1015-1016 cycles/second, and X-rays have a frequency of 1016-1020 cycles/second (g and cosmic rays have even higher frequencies). So the electromagnetic fields (EMFs) of alternating current have only a trivial level of energy compared to the mildest form of ionizing radiation that can cause cancer -- UV.

In addition to the energy of EM radiation, one needs to consider the intensity. Intensity is the amount of radiation per unit time. For example, a light bulb emits more intensely when it is bright than when it is dim, even though the energy level of individual photons is the same. So even though EMF from alternating current might be too low in energy to produce mutations, high intensity fields might have some biological effects. Here again, however, there would seem to be little reason for concern. Field intensity falls rapidly with distance, so even though the field intensity of various household appliances is high at the source (e.g., the motor in a hair dryer), the field is quite small only a few inches away. And intensities experienced in the household are small relative to the Earth's magnetic field and to the electrical fields generated by our own cells. The only possible cause for concern, therefore is that the man-made fields oscillate, whereas the cells' electrical fields and Earth's magnetic fields do not.

Oscillating magnetic fields do have biological effects – they generate currents in body tissue that are easily measured. However, normal muscle activity also generates currents as well (with no known function). On the whole, this EMF effect on bodies and tissues is not large compared to normal levels. However, there is general ignorance about these effects, so any conclusions are tentative.

Where Things Stand Now – no cause for concern

A report by the National Academy of Sciences in 1997 (Possible Health Effects of Exposure to Residential Electric and Magnetic Fields, http://www.nap.edu/openbook/0309054478/html) summarized the then present status of residential EMFs and cancer:

1. There remains a statistically significant correlation between childhood leukemia and the wire code of a house (mostly based on the distance between the house and high current power lines). The highest-code houses have about a 1.5 risk factor (50% increase). There is no significant correlation for other childhood cancers or for any adult cancer.

2. There is no correlation between EMFs measured inside the households and childhood leukemia (measured after leukemia was diagnosed). The cause of the correlation in (1) remains unknown.

3. In vitro effects (cell culture) reveal abnormalities only at EMF doses 1,000-100,000 times greater than typical residential exposures. These effects on cells do not include genetic damage.

4. Exposure of lab animals to EMFs has not shown any consistent pattern with cancer, even at high EMF doses. Some behavior responses are seen at high doses, and there is an intriguing result that animals exposed to both a known carcinogen and intense EMF show increased breast cancer levels.

As it stands, there is no reason to be concerned about residential EMF levels. As is true in all scientific matters, our current conclusions may change as new evidence comes in. But there is already compelling evidence that any cancer-causing effect of EMFs is not very large.

SECTION 6

Why Do We Bother With Correlations At All?

Given the problems with interpreting correlational data, one might reasonably ask: why do we bother with them at all if it is a causal relationship that we seek? Why not just gather data that could provide a more definite answer, or otherwise just ignore correlations? The reason is pragmatism. Correlational data are usually relatively easy and inexpensive to obtain, at least in comparison to experimental data. Also, many cause-effect relationships are so subtle that we often first learn of them through correlations detected in observational data. That is, they are often useful.