1

Observational Studies and Experiments

1.1 Introduction

This book is about regression models and variants like path models, simultaneous-equation models, logits and probits. Regression models can be used for different purposes:

(i) to summarize data,
(ii) to predict the future,
(iii) to predict the results of interventions.

The third—causal inference—is the most interesting and the most slippery. It will be our focus. For background, this section covers some basic principles of study design.

Causal inferences are made from *observational studies*, *natural experiments*, and *randomized controlled experiments*. When using observational (non-experimental) data to make causal inferences, the key problem is *confounding*. Sometimes this problem is handled by subdividing the study population (*stratification*, also called *cross-tabulation*), and sometimes by modeling. These strategies have various strengths and weaknesses, which need to be explored.

In medicine and social science, causal inferences are most solid when based on randomized controlled experiments, where investigators assign subjects at random—by the toss of a coin—to a *treatment group* or to a *control group*. Up to random error, the coin balances the two groups with respect to all relevant factors other than treatment. Differences between the treatment group and the control group are therefore due to treatment. That is why causation is relatively easy to infer from experimental data. However, experiments tend to be expensive, and may be impossible for ethical or practical reasons. Then statisticians turn to observational studies.

In an observational study, it is the subjects who assign themselves to the different groups. The investigators just watch what happens. Studies on the effects of smoking, for instance, are necessarily observational. However, the treatment-control terminology is still used. The investigators compare smokers (the treatment group, also called the *exposed group*) with nonsmokers (the control group) to determine the effect of smoking. The jargon is a little confusing, because the word "control" has two senses:

(i) a control is a subject who did not get the treatment;
(ii) a controlled experiment is a study where the investigators decide who will be in the treatment group.

Smokers come off badly in comparison with nonsmokers. Heart attacks, lung cancer, and many other diseases are more common among smokers. There is a strong *association* between smoking and disease. If cigarettes cause disease, that explains the association: death rates are higher for smokers because cigarettes kill. Generally, association is circumstantial evidence for causation. However, the proof is incomplete. There may be some hidden confounding factor which makes people smoke and also makes them sick. If so, there is no point in quitting: that will not change the hidden factor. Association is not the same as causation.

> Confounding means a difference between the treatment and control groups—other than the treatment—which affects the response being studied.

Typically, a confounder is a third variable which is associated with exposure and influences the risk of disease.

Statisticians like Joseph Berkson and R. A. Fisher did not believe the evidence against cigarettes, and suggested possible confounding variables. Epidemiologists (including Richard Doll and Bradford Hill in England, as well as Wynder, Graham, Hammond, Horn, and Kahn in the United States) ran careful observational studies to show these alternative explanations were

not plausible. Taken together, the studies make a powerful case that smoking causes heart attacks, lung cancer, and other diseases. If you give up smoking, you will live longer.

Epidemiological studies often make comparisons separately for smaller and more homogeneous groups, assuming that within these groups, subjects have been assigned to treatment or control as if by randomization. For example, a crude comparison of death rates among smokers and nonsmokers could be misleading if smokers are disproportionately male, because men are more likely than women to have heart disease and cancer. Gender is therefore a confounder. To control for this confounder—a third use of the word "control"—epidemiologists compared male smokers to male nonsmokers, and females to females.

Age is another confounder. Older people have different smoking habits, and are more at risk for heart disease and cancer. So the comparison between smokers and nonsmokers was made separately by gender and age: for example, male smokers age 55–59 were compared to male nonsmokers in the same age group. This controls for gender and age. Air pollution would be a confounder, if air pollution causes lung cancer and smokers live in more polluted environments. To control for this confounder, epidemiologists made comparisons separately in urban, suburban, and rural areas. In the end, explanations for health effects of smoking in terms of confounders became very, very implausible.

Of course, as we control for more and more variables this way, study groups get smaller and smaller, leaving more and more room for chance effects. This is a problem with cross-tabulation as a method for dealing with confounders, and a reason for using statistical models. Furthermore, most observational studies are less compelling than the ones on smoking. The following (slightly artificial) example illustrates the problem.

Example 1. In cross-national comparisons, there is a striking correlation between the number of telephone lines per capita in a country and the death rate from breast cancer in that country. This is not because talking on the telephone causes cancer. Richer countries have more phones and higher cancer rates. The probable explanation for the excess cancer risk is that women in richer countries have fewer children. Pregnancy—especially early first pregnancy—is protective. Differences in diet and other lifestyle factors across countries may also play some role.

Randomized controlled experiments minimize the problem of confounding. That is why causal inferences from randomized controlled experiments are stronger than those from observational stud-

ies. With observational studies of causation, you always have to worry about confounding. What were the treatment and control groups? How were they different, apart from treatment? What adjustments were made to take care of the differences? Are these adjustments sensible?

The rest of this chapter will discuss examples: the HIP trial of mammography, Snow on cholera, and the causes of poverty.

1.2 The HIP trial

Breast cancer is one of the most common malignancies among women in Canada and the United States. If the cancer is detected early enough—before it spreads—chances of successful treatment are better. "Mammography" means screening women for breast cancer by X-rays. Does mammography speed up detection by enough to matter? The first large-scale randomized controlled experiment was HIP (Health Insurance Plan) in New York, followed by the Two-County study in Sweden. There were about half a dozen other trials as well. Some were negative (screening doesn't help) but most were positive. By the late 1980s, mammography had gained general acceptance.

The HIP study was done in the early 1960s. HIP was a group medical practice which had at the time some 700,000 members. Subjects in the experiment were 62,000 women age 40–64, members of HIP, who were randomized to treatment or control. "Treatment" consisted of invitation to 4 rounds of annual screening—a clinical exam and mammography. The control group continued to receive usual health care. Results from the first 5 years of followup are shown in table 1. In the treatment group, about 2/3 of the women accepted the invitation to be screened, and 1/3 refused. Death rates (per 1000 women) are shown, so groups of different sizes can be compared.

Table 1. HIP data. Group sizes (rounded), deaths in 5 years of followup, and death rates per 1000 women randomized.

	Group size	Breast cancer		All other	
		No.	Rate	No.	Rate
Treatment					
Screened	20,200	23	1.1	428	21
Refused	10,800	16	1.5	409	38
Total	31,000	39	1.3	837	27
Control	31,000	63	2.0	879	28

Which rates show the efficacy of treatment? It seems natural to compare those who accepted screening to those who refused. However, this is an observational comparison, even though it occurs in the middle of an experiment. The investigators decided which subjects would be invited to screening, but it is the subjects themselves who decided whether or not to accept the invitation. Richer and better-educated subjects were more likely to participate than those who were poorer and less well educated. Furthermore, breast cancer (unlike most other diseases) hits the rich harder than the poor. Social status is therefore a confounder—a factor associated with the outcome and with the decision to accept screening.

The tip-off is the death rate from other causes (not breast cancer) in the last column of table 1. There is a big difference between those who accept screening and those who refuse. The refusers have almost double the risk of those who accept. There must be other differences between those who accept screening and those who refuse, in order to account for the doubling in the risk of death from other causes—because screening has no effect on the risk.

One major difference is social status. It is the richer women who come in for screening. Richer women are less vulnerable to other diseases but more vulnerable to breast cancer. So the comparison of those who accept screening with those who refuse is biased, and the bias is against screening.

Comparing the death rate from breast cancer among those who accept screening and those who refuse is *analysis by treatment received*. This analysis is seriously biased, as we have just seen. The experimental comparison is between the whole treatment group—all those invited to be screened, whether or not they accepted screening—and the whole control group. This is the *intention-to-treat analysis*.

Intention-to-treat is the recommended analysis.

HIP, which was a very well-run study, made the intention-to-treat analysis. The investigators compared the breast cancer death rate in the total treatment group to the rate in the control group, and showed that screening works.

The effect of the invitation is small in absolute terms: $63 - 39 = 24$ lives saved (table 1). Since the absolute risk from breast cancer is small, no intervention can have a large effect in absolute terms. On the other hand, in relative terms, the 5-year death rates from breast cancer are in the ratio $39/63 = 62\%$. Followup continued for 18 years, and the savings in lives persisted over that period. The Two-County study—a huge randomized controlled experiment in Sweden—confirmed the results of HIP. So did other studies in Finland, Scotland, and Sweden. That is why mammography became so widely accepted.

1.3 Snow on cholera

A *natural experiment* is an observational study where assignment to treatment or control is as if randomized by nature. In 1855, some twenty years before Koch and Pasteur laid the foundations of modern microbiology, John Snow used a natural experiment to show that cholera is a waterborne infectious disease. At the time, the germ theory of disease was only one of many theories. Miasmas (foul odors, especially from decaying organic material) were often said to cause epidemics. Imbalance in the humors of the body—black bile, yellow bile, blood, phlegm—was an older theory. Poison in the ground was an explanation that came into vogue slightly later.

Snow was a physician in London. By observing the course of the disease, he concluded that cholera was caused by a living organism which entered the body with water or food, multiplied in the body, and made the body expel water containing copies of the organism. The dejecta then contaminated food or reentered the water supply, and the organism proceeded to infect other victims. Snow explained the lag between infection and disease—a matter of hours or days—as the time needed for the infectious agent to multiply in the body of the victim. This multiplication is characteristic of life: inanimate poisons do not reproduce themselves. (Of course, poisons may take some time to do their work: the lag is not compelling evidence.)

Snow developed a series of arguments in support of the germ theory. For instance, cholera spread along the tracks of human commerce. Furthermore, when a ship entered a port where cholera was prevalent, sailors contracted the disease only when they came into contact with residents of the port. These facts were easily explained if cholera was an infectious disease, but were hard to explain by the miasma theory.

There was a cholera epidemic in London in 1848. Snow identified the first or "index" case in this epidemic:

> "a seaman named John Harnold, who had newly arrived by the *Elbe* steamer from Hamburgh, where the disease was prevailing." [p. 3]

He also identified the second case: a man named Blenkinsopp who took Harnold's room after the latter died, and became infected by contact with the bedding. Next, Snow was able to find adjacent apartment buildings, one hard hit by cholera and one not. In each case, the affected building had a water supply contaminated by sewage, the other had relatively pure water. Again, these facts are easy to understand if cholera is an infectious disease—but not if miasmas are the cause.

There was an outbreak of the disease in August and September of 1854. Snow made a "spot map," showing the locations of the victims. These clus-

tered near the Broad Street pump. (Broad Street is in Soho, London; at the time, public pumps were used as a source of drinking water.) By contrast, there were a number of institutions in the area with few or no fatalities. One was a brewery. The workers seemed to have preferred ale to water; if any wanted water, there was a private pump on the premises. Another institution almost free of cholera was a poor-house, which too had its own private pump. (Poor-houses will be discussed again, in section 4.)

People in other areas of London did contract the disease. In most cases, Snow was able to show they drank water from the Broad Street pump. For instance, one lady in Hampstead so much liked the taste that she had water from the Broad Street pump delivered to her house by carter.

So far, we have persuasive anecdotal evidence that cholera is an infectious disease, spread by contact or through the water supply. Snow also used statistical ideas. There were a number of water companies in the London of his time. Some took their water from heavily contaminated stretches of the Thames river. For others, the intake was relatively uncontaminated.

Snow made "ecological" studies, correlating death rates from cholera in various areas of London with the quality of the water. Generally speaking, areas with contaminated water had higher death rates. The Chelsea water company was exceptional. This company started with contaminated water, but had quite modern methods of purification—with settling ponds and careful filtration. Its service area had a low death rate from cholera.

In 1852, the Lambeth water company moved its intake pipe upstream to get purer water. The Southwark and Vauxhall company left its intake pipe where it was, in a heavily contaminated stretch of the Thames. Snow made an ecological analysis comparing the areas serviced by the two companies in the epidemics of 1853–54 and in earlier years. Let him now continue in his own words.

> "Although the facts shown in the above table [the ecological analysis] afford very strong evidence of the powerful influence which the drinking of water containing the sewage of a town exerts over the spread of cholera, when that disease is present, yet the question does not end here; for the intermixing of the water supply of the Southwark and Vauxhall Company with that of the Lambeth Company, over an extensive part of London, admitted of the subject being sifted in such a way as to yield the most incontrovertible proof on one side or the other. In the subdistricts enumerated in the above table as being supplied by both Companies, the mixing of the supply is of the most intimate kind. The pipes of each Company go down all the streets, and into nearly all the courts and alleys. A few houses are supplied by one Company and a few by the other, according to the decision of the owner or occupier at that time when the Water Companies were in active competition.

> In many cases a single house has a supply different from that on either side. Each company supplies both rich and poor, both large houses and small; there is no difference either in the condition or occupation of the persons receiving the water of the different Companies. Now it must be evident that, if the diminution of cholera, in the districts partly supplied with improved water, depended on this supply, the houses receiving it would be the houses enjoying the whole benefit of the diminution of the malady, whilst the houses supplied with the [contaminated] water from Battersea Fields would suffer the same mortality as they would if the improved supply did not exist at all. As there is no difference whatever in the houses or the people receiving the supply of the two Water Companies, or in any of the physical conditions with which they are surrounded, it is obvious that no experiment could have been devised which would more thoroughly test the effect of water supply on the progress of cholera than this, which circumstances placed ready made before the observer.
>
> "The experiment, too, was on the grandest scale. No fewer than three hundred thousand people of both sexes, of every age and occupation, and of every rank and station, from gentlefolks down to the very poor, were divided into groups without their choice, and in most cases, without their knowledge; one group being supplied with water containing the sewage of London, and amongst it, whatever might have come from the cholera patients; the other group having water quite free from such impurity.
>
> "To turn this grand experiment to account, all that was required was to learn the supply of water to each individual house where a fatal attack of cholera might occur." [pp. 74–75]

Snow's data are shown in table 2. The denominator data—the number of houses served by each water company—were available from parliamentary records. For the numerator data, however, a house-to-house canvass was needed to determine the source of the water supply at the address of each cholera fatality. (The "bills of mortality," as death certificates were called at the time, showed the address but not the water source for each victim.) The death rate from the Southwark and Vauxhall water is about 9 times the death rate for the Lambeth water. Snow explains that the data could be analyzed as

Table 2. Death rate from cholera by source of water. Rate per 10,000 houses. London. Epidemic of 1854. Snow's table IX.

	No. of Houses	Cholera Deaths	Rate per 10,000
Southwark & Vauxhall	40,046	1,263	315
Lambeth	26,107	98	37
Rest of London	256,423	1,422	59

if they had resulted from a randomized controlled experiment: there was no difference between the customers of the two water companies, except for the water. The data analysis is simple—a comparison of rates. It is the design of the study and the size of the effect that compel conviction.

1.4 Yule on the causes of poverty

Legendre (1805) and Gauss (1809) developed regression techniques to fit data on orbits of astronomical objects. The relevant variables were known from Newtonian mechanics, and so were the functional forms of the equations connecting them. Measurement could be done with high precision. Much was known about the nature of the errors in the measurements and equations. Furthermore, there was ample opportunity for comparing predictions to reality. A century later, investigators were using regression on social science data where these conditions did not hold, even to a rough approximation—with consequences that need to be explored (chapters 4–9).

Yule (1899) was studying the causes of poverty. At the time, paupers in England were supported either inside grim Victorian institutions called "poor-houses" or outside, depending on the policy of local authorities. Did policy choices affect the number of paupers? To study this question, Yule proposed a regression equation,

$$\Delta\text{Paup} = a + b \times \Delta\text{Out} + c \times \Delta\text{Old} + d \times \Delta\text{Pop} + \text{error}. \tag{1}$$

In this equation,

Δ is percentage change over time,
Paup is the percentage of paupers,
Out is the out-relief ratio N/D,
$N =$ number on welfare outside the poor-house,
$D =$ number inside,
Old is the percentage of the population aged over 65,
Pop is the population.

Data are from the English Censuses of 1871, 1881, 1891. There are two Δ's, one for 1871–81 and one for 1881–91. (Error terms will be discussed later.)

Relief policy was determined separately in each "union" (an administrative district comprising several parishes). At the time, there were about 600 unions, and Yule divided them into four kinds: rural, mixed, urban, metropolitan. There are $4\times2 = 8$ equations, one for each type of union and time period. Yule fitted his equations to the data by least squares. That is, he determined a, b, c, and d by minimizing the sum of squared errors,

$$\sum \left(\Delta\text{Paup} - a - b \times \Delta\text{Out} - c \times \Delta\text{Old} - d \times \Delta\text{Pop}\right)^2.$$

The sum is taken over all unions of a given type in a given time period, which assumes (in effect) that coefficients are constant for those combinations of geography and time.

Table 3. Pauperism, Out-relief ratio, Proportion of Old, Population. Ratio of 1881 data to 1871 data, times 100. Metropolitan Unions, England. Yule (1899, table XIX).

	Paup	Out	Old	Pop
Kensington	27	5	104	136
Paddington	47	12	115	111
Fulham	31	21	85	174
Chelsea	64	21	81	124
St. George's	46	18	113	96
Westminster	52	27	105	91
Marylebone	81	36	100	97
St. John, Hampstead	61	39	103	141
St. Pancras	61	35	101	107
Islington	59	35	101	132
Hackney	33	22	91	150
St. Giles'	76	30	103	85
Strand	64	27	97	81
Holborn	79	33	95	93
City	79	64	113	68
Shoreditch	52	21	108	100
Bethnal Green	46	19	102	106
Whitechapel	35	6	93	93
St. George's East	37	6	98	98
Stepney	34	10	87	101
Mile End	43	15	102	113
Poplar	37	20	102	135
St. Saviour's	52	22	100	111
St. Olave's	57	32	102	110
Lambeth	57	38	99	122
Wandsworth	23	18	91	168
Camberwell	30	14	83	168
Greenwich	55	37	94	131
Lewisham	41	24	100	142
Woolwich	76	20	119	110
Croydon	38	29	101	142
West Ham	38	49	86	203