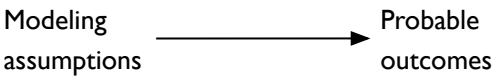
Intermezzo: From Models to Inference[1]

STA 371 · James Scott

Spring 2011

AGAIN AND AGAIN over the first three chapters, we've found ourselves making certain assumptions about the world: that airline no-shows follow a binomial distribution, that basketball scores and stock prices are random walks, that mudslides and hurricanes happen with fixed probabilities known in advance. All of these assumptions took the form of probability models for describing some as-yet-unknown outcome. And all of our subsequent reasoning was deductive: "If Model *A* is true, then what can we conclude using the rules of logic and probability?"



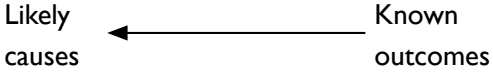
Deduction using probability models can yield surprisingly deep and subtle insights: about the impressiveness of a "miracle comeback" in sports, about the value of insurance, about the importance of diversifying your investments.

But this kind of reasoning has its limits—and not merely those imposed by the scope of our imagination, whence models come. Rather, the main limiting factor is *information*. (Just how likely are hurricanes? Just how variable are stock prices?) John Kenneth Galbraith said it well:

> Nothing is so hard to come by as a new and interesting fact. Nothing is so easy on the feet as a generalization. I now pick up magazines and leaf through them looking for articles that are rich with facts; I do not care much what they are. Richly evocative and deeply percipient theory I avoid. It leaves me cold unless I am the author of it. [2]

[2] "Writing, Typing, and Economics." *The Atlantic*, March 1978.

From now on, we will pursue a new goal: *to learn how to learn*, or more precisely, to learn what use to make of facts if we wish to learn things about the world. For this purpose, we'll still use deduction, but it will cease to be the primary form of argument. No longer will we consider probability models that drift freely in our imagination, without any data mooring them to the real world. Instead, we must incorporate data into our models by "going in the other direction," reasoning from known outcomes to their likely causes:



This process is called *inference*. It is why we learn statistics.

'Facts are stubborn things'

STATISTICS, as a field, has a funny reputation. On the one hand, we'd like to believe that "the numbers don't lie," that statistics are cold, hard facts—as objective as they come. It sometimes seems, moreover, as though statistical methods are the only way to gain credibility in an increasingly data-driven world.

Consider, for example, the intimidating litany of statistical requirements that authors must meet to get their work published in the *Journal of the American Medical Association*:

> Numerical results should be accompanied by confidence intervals, if applicable, and exact levels of statistical significance. Evaluations of screening and diagnostic tests should include sensitivity, specificity, likelihood ratios, receiver operating characteristic curves, and predictive values. [3]

[3] *JAMA Instructions for Authors*, `jama.ama-assn.org`

Or the following blog entry from Peter Orszag, President Obama's director of the Office of Management and Budget:

> The President has made it very clear that policy decisions should be driven by evidence—accentuating the role of Federal statistics as a resource for policymakers. Robust, unbiased data are the first step toward addressing our long-term economic needs and key policy priorities.[4]

[4] "Using Statistics to Drive Sound Policy." Office of Management and Budget Blog; May 8, 2009

The New York Times captured the zeitgeist nicely, summing up some advice for college students in a single headline.

> For Today's Graduate, Just One Word: Statistics[5]

[5] *New York Times* (Technology section); August 5, 2009

On the other hand, too many statistics can make it seem like we're having the wool pulled over our eyes. We may associate them with cheats, with frauds, with hucksters, with Congressmen. Why else would we merely roll our eyes, cynically yet knowingly, at Winston Churchill's brazen politicking?

> I gather, young man, that you wish to be a Member of Parliament. The first lesson that you must learn is that, when I call for statistics about the rate of infant mortality, what I want is proof that fewer babies died when I was Prime Minister than when anyone else was Prime Minister.[6]

[6] Quoted in *The Life of Politics* (1968), Henry Fairlie, Methuen, pp. 203–204

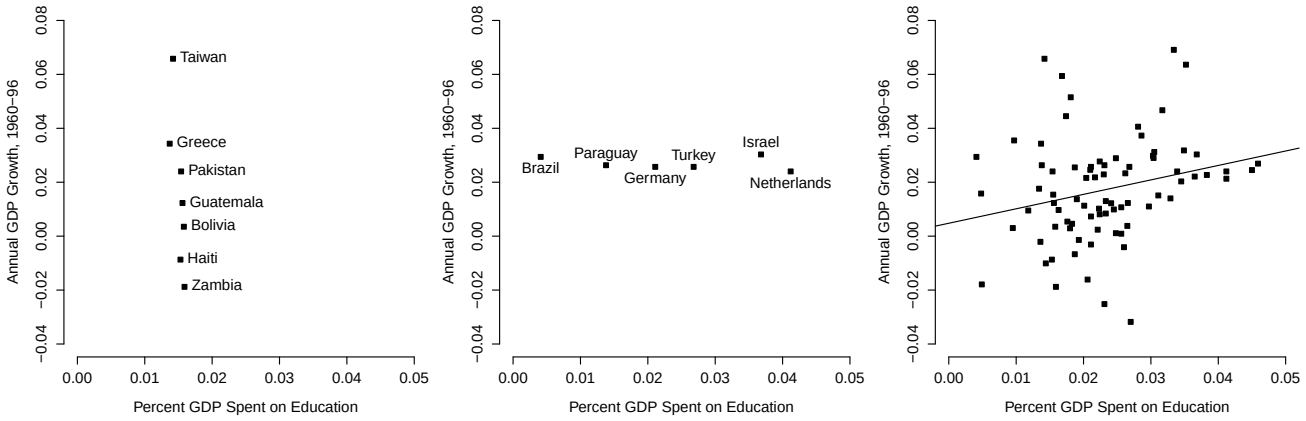
And why else would the famous remark, popularized by Twain and attributed to Disraeli, remain so apposite over a hundred years later?

> Figures often beguile me, particularly when I have the arranging of them myself; in which case the remark attributed to Disraeli would often apply with justice and force: 'There are three kinds of lies: lies, damned lies, and statistics.'[7]

[7] *Chapters from My Autobiography*, North American Review (1907)

Clearly statistics bring not just prestige, but also tremendous incentives for debasement. In that way statistics are just like $20 bills. The real ones are quite valuable, and so are the fake ones, if you can make others believe they're real.

Of course, people can lie with statistics in ways beyond merely reciting statistics that are lies. Selective reporting, whether intentional or not, is also a form of falsehood.



Above, at left, we see six countries that all spend around 1.5% of their GDP on education, and yet had wildly different rates of economic growth for the 37 years spanning 1960 to 1996. In the middle panel, we see another group of six countries with similar growth rates of around 2–3%, but with very different levels of spending on education. Both highly selective samples make it seem as though education and economic growth are barely related, if at all.

The much bigger sample in the third panel tells a more complicated story. It shows a plot of GDP growth versus education spending for a sample of 79 countries worldwide, together with the best-fitting linear trend. This line isn't anywhere near the final word on the subject, as we'll see in the next chapter. Even still, it at least suggests a positive relationship, and does so with a much more comprehensive body of evidence.

Yet how tempting just to cherry pick! Indeed, we're all used to seeing popular news stories that cite data little better than that of the left or middle panel, perhaps with a plausible "just-so" story thrown in for good measure:

> Second, higher levels of education are critical to economic growth. . . . Boston, where there is a high proportion of college graduates, is the perfect example. Well-educated people can react more quickly to technological changes and learn new skills more readily. Even without the climate advantages of a city like San Jose, Calif., Boston evolved into what we now think of as an "information city." By comparison, Detroit, with lower levels of education, languished.[8]

[8] "Economic Scene." *New York Times* (Business section); August 5, 2004

And this from a reporter who presumably has no hidden agenda. Notice how the selective reporting of evidence—one causal hypothesis, two data points—lends an air of such graceful inevitability to what is really quite a superficial analysis of the factors that might explain the divergent fates of Boston and Detroit over the last three decades.

In light of all this, there are at least two good reasons to learn statistics:

(1) To use data honestly and credibly in the service of an argument you believe in.

(2) To know when to be skeptical of someone else's damned lies.

For as John Adams put it:

> Facts are stubborn things; and whatever may be our wishes, our inclinations, or the dictates of our passion, they cannot alter the state of facts and evidence.[9]

[9] John Adams, 'Argument in Defense of the Soldiers in the Boston Massacre Trials' (1770)

Data analysis for making decisions

THE REST OF THIS book is about regression, defined loosely as "fitting equations to data."[10] (For example: the linear trend in growth versus education on the previous page.) Regression is not a single formula, and cannot be reduced to a tidy set of numerical summaries that extract the sum total of all possible truths from a data set. Instead, regression is a *process*, one guided by human intuition and vulnerable to human error—even when all the number-crunching has been done correctly.

[10] E.R. Tufte. *Data Analysis for Politics and Policy* (p. ix), Prentice–Hall, 1974

What do we hope to accomplish when we fit an equation to data? A single analysis can have many goals, but over the next handful of chapters we'll consider three main ones:

1. To *explore* a large body of multivariate data and summarize some of its general features.
2. To *test* our beliefs about cause-and-effect relationships among things in the world.
3. To *predict* the future consequences of some known configuration of forces, using past observations as guidance.

To explore; to test; to predict. These are the goals not merely of regression analysis, but of the scientific method more generally.

Some history

On the time scale of important post-Enlightenment ideas, regression is middle-aged. A man named Tobias Mayer was using something that looked vaguely like regression as early as 1750.[11] But scholars credit two later mathematicians—Legendre, a Frenchmen; and Gauss, a German—with independently inventing the modern tool of *least squares* some time between 1794 and 1805.

[11] Stephen M. Stigler, *The History of Statistics: The Measurement of Uncertainty before 1900*, pp. 16–25. Harvard University Press, 1986

That makes regression newer than the invention of calculus (credited jointly to Leibniz and Newton in the late 1600's), but older than the idea of evolution by natural selection (credited jointly to Darwin and Wallace over a period spanning the 1830's to the 1850's). And as a force shaping the modern world, regression is probably just as important as either calculus or natural selection. It may not be sexy or controversial. But regression analysis does provide the intellectual and empirical support for thousands of decisions—some small, some enormously consequential—that governments and businesses face every day. And when lives, money, or even the long-term fate of nations hang in the balance over a single decision, you want to squeeze every last drop of information from the available data. In these situations, regression is not a magic solution for getting everything right. But it sure can help.

Two broad themes

Regression is the primary methodological theme of the rest of this book. But we'll also explore two main philosophical themes. These may be less tangible, and of less obvious utility, than a fast computer with a good software package for implementing regression. But don't look beyond them, for they are no less important than fancy number-crunching for understanding large, complex data sets. These two themes are:

The difficulty of making causal judgments from empirical evidence. You will, no doubt, have heard it said before that correlation and causation are not the same thing. True enough. But that hasn't stopped humans from learning that smoking causes cancer, or that lightning causes thunder, on the basis of observation. What distinguishes these solid conclusions from the tenuous, or the downright absurd? How do we know that causation doesn't run the other way? (Answer: sometimes we don't, but good study design and thorough data analysis help a lot.)

The tension between sufficient explanations and parsimonious ones. We obviously want our theories to do justice to the complexity of the real world. At the same time, we don't want them to be overdetermined and ad hoc: that is, tuned so perfectly to past experience that they lose all ability to generalize to future cases. This is Occam's Razor: theories should be made as complicated as they need to be to answer a question, and no more so. Yet how do we operationalize this criterion in the situations where we must often turn to regression—where a hundred different things *might* be going on, and we don't know which of them are the most important? All quantitative models of the world must confront this antagonism between the twin virtues of fit and simplicity. Regression models are no exception.