Regression to the Mean

I had one of the most satisfying eureka experiences of my career while teaching flight instructors in the Israeli Air Force about the psychology of effective training. I was telling them about an important principle of skill training: rewards for improved performance work better than punishment of mistakes. This proposition is supported by much evidence from research on pigeons, rats, humans, and other animals.

When I finished my enthusiastic speech, one of the most seasoned instructors in the group raised his hand and made a short speech of his own. He began by conceding that rewarding improved performance might be good for the birds, but he denied that it was optimal for flight cadets. This is what he said: "On many occasions I have praised flight cadets for clean execution of some aerobatic maneuver. The next time they try the same maneuver they usually do worse. On the other hand, I have often screamed into a cadet's earphone for bad execution, and in general he does better t t ask yry abr two repon his next try. So please don't tell us that reward works and punishment does not, because the opposite is the case."

This was a joyous moment of insight, when I saw in a new light a principle of statistics that I had been teaching for years. The instructor was right—but he was also completely wrong! His observation was astute and correct: occasions on which he praised a performance were likely to be followed by a disappointing performance, and punishments were typically followed by an improvement. But the inference he had drawn about the efficacy of reward and punishment was completely off the mark. What he had observed is known as *regression to the mean*, which in that case was due to random fluctuations in the quality of performance. Naturally, he praised only a cadet whose performance was far better than average. But the cadet was probably just lucky on that particular attempt and therefore likely to deteriorate regardless of whether or not he was praised. Similarly, the instructor would shout into a cadet's earphones only when the cadet's performance was unusually bad and therefore likely to improve regardless of what the instructor did. The instructor had attached a causal interpretation to the inevitable fluctuations of a random process.

The challenge called for a response, but a lesson in the algebra of prediction would not be enthusiastically received. Instead, I used chalk to mark a target on the floor. I asked every officer in the room to turn his back to the target and throw two coins at it in immediate succession, without looking. We measured the distances from the target and wrote the two results of each contestant on the blackboard. Then we rewrote the results

in order, from the best to the worst performance on the first try. It was apparent that most (but not all) of those who had done best the first time deteriorated on their second try, and those who had done poorly on the first attempt generally improved. I pointed out to the instructors that what they saw on the board coincided with what we had heard about the performance of aerobatic maneuvers on successive attempts: poor performance was typically followed by improvement and good performance by deterioration, without any help from either praise or punishment.

The discovery I made on that day was that the flight instructors were trapped in an unfortunate contingency: because they punished cadets when performance was poor, they were mostly rewarded by a subsequent improvement, even if punishment was actually ineffective. Furthermore, the instructors were not alone in that predicament. I had stumbled onto a significant fact of the human condition: the feedback to which life exposes us is perverse. Because we tend to be nice to other people when they please us and nasty when they do not, we are statistically punished for being nice and rewarded for being nasty.

Talent and Luck

A few years ago, John Brockman, who edits the online magazine *Edge*, asked a number of scientists to report their "favorite equation." These were my offerings:

success = talent + luck
great success = a little more talent + a lot of luck

The unsurprising idea that luck often contributes to success has surprising consequences when we apply it to the first two days of a high-level golf tournament. To keep things simple, assume that on both days the average score of the competitors was at par 72. We focus on a player who did verye d well on the first day, closing with a score of 66. What can we learn from that excellent score? An immediate inference is that the golfer is more talented than the average participant in the tournament. The formula for success suggests that another inference is equally justified: the golfer who did so well on day 1 probably enjoyed better-than-average luck on that day. If you accept that talent and luck both contribute to success, the conclusion that the successful golfer was lucky is as warranted as the conclusion that he is talented.

By the same token, if you focus on a player who scored 5 over par on

that day, you have reason to infer both that he is rather weak *and* had a bad day. Of course, you know that neither of these inferences is certain. It is entirely possible that the player who scored 77 is actually very talented but had an exceptionally dreadful day. Uncertain though they are, the following inferences from the score on day 1 are plausible and will be correct more often than they are wrong.

> above-average score on day 1 = above-average talent + lucky on day 1

and

> below-average score on day 1 = below-average talent + unlucky on day 1

Now, suppose you know a golfer's score on day 1 and are asked to predict his score on day 2. You expect the golfer to retain the same level of talent on the second day, so your best guesses will be "above average" for the first player and "below average" for the second player. Luck, of course, is a different matter. Since you have no way of predicting the golfers' luck on the second (or any) day, your best guess must be that it will be average, neither good nor bad. This means that in the absence of any other information, your best guess about the players' score on day 2 should not be a repeat of their performance on day 1. This is the most you can say:

- The golfer who did well on day 1 is likely to be successful on day 2 as well, but less than on the first, because the unusual luck he probably enjoyed on day 1 is unlikely to hold.
- The golfer who did poorly on day 1 will probably be below average on day 2, but will improve, because his probable streak of bad luck is not likely to continue.

We also expect the difference between the two golfers to shrink on the second day, although our best guess is that the first player will still do better than the second.

My students were always surprised to hear that the best predicted performance on day 2 is more moderate, closer to the average than the evidence on which it is based (the score on day 1). This is why the pattern is called regression to the mean. The more extreme the original score, the

more regression we expect, because an extremely good score suggests a very lucky day. The regressive prediction is reasonable, but its accuracy is not guaranteed. A few of the golfers who scored 66 on day 1 will do even better on the second day, if their luck improves. Most will do worse, because their luck will no longer be above average.

Now let us go against the time arrow. Arrange the players by their performance on day 2 and look at their performance on day 1. You will find precisely the same pattern of regression to the mean. The golfers who did best on day 2 were probably lucky on that day, and the best guess is that they had been less lucky and had done filess well on day 1. The fact that you observe regression when you predict an early event from a later event should help convince you that regression does not have a causal explanation.

Regression effects are ubiquitous, and so are misguided causal stories to explain them. A well-known example is the "*Sports Illustrated* jinx," the claim that an athlete whose picture appears on the cover of the magazine is doomed to perform poorly the following season. Overconfidence and the pressure of meeting high expectations are often offered as explanations. But there is a simpler account of the jinx: an athlete who gets to be on the cover of *Sports Illustrated* must have performed exceptionally well in the preceding season, probably with the assistance of a nudge from luck—and luck is fickle.

I happened to watch the men's ski jump event in the Winter Olympics while Amos and I were writing an article about intuitive prediction. Each athlete has two jumps in the event, and the results are combined for the final score. I was startled to hear the sportscaster's comments while athletes were preparing for their second jump: "Norway had a great first jump; he will be tense, hoping to protect his lead and will probably do worse" or "Sweden had a bad first jump and now he knows he has nothing to lose and will be relaxed, which should help him do better." The commentator had obviously detected regression to the mean and had invented a causal story for which there was no evidence. The story itself could even be true. Perhaps if we measured the athletes' pulse before each jump we might find that they are indeed more relaxed after a bad first jump. And perhaps not. The point to remember is that the change from the first to the second jump does not need a causal explanation. It is a mathematically inevitable consequence of the fact that luck played a role in the outcome of the first jump. Not a very satisfactory story—we would all prefer a causal account—but that is all there is.

Understanding Regression

Whether undetected or wrongly explained, the phenomenon of regression is strange to the human mind. So strange, indeed, that it was first identified and understood two hundred years after the theory of gravitation and differential calculus. Furthermore, it took one of the best minds of nineteenth-century Britain to make sense of it, and that with great difficulty.

Regression to the mean was discovered and named late in the nineteenth century by Sir Francis Galton, a half cousin of Charles Darwin and a renowned polymath. You can sense the thrill of discovery in an article he published in 1886 under the title "Regression towards Mediocrity in Hereditary Stature," which reports measurements of size in successive generations of seeds and in comparisons of the height of children to the height of their parents. He writes about his studies of seeds:

> They yielded results that seemed very noteworthy, and I used them as the basis of a lecture before the Royal Institution on February 9th, 1877. It appeared from these experiments that the offspring did *not* tend to resemble their parent seeds in size, but to be always more mediocre than they—to be smaller than the parents, if the parents were large; to be larger than the parents, if the parents were very small…The experiments showed further that the mean filial regression towards mediocrity was directly proportional to the parental deviation from it.

Galton obviously expected his learned audience at the Royal Institution—the oldest independent research society in the world—to be as surprised by his "noteworthy observation" as he had been. What is truly noteworthy is that he was surprised by a statistical regularity that is as common as the air we breathe. Regression effects can be found wherever we look, but we do not recognize them for what they are. They hide in plain sight. It took Galton several years to work his way from his discovery of filial regression in size to the broader notion that regression inevitably occurs when the correlation between two measures is less than perfect, and he needed the help of the most brilliant statisticians of his time to reach that conclusion.

One of the hurdles Galton had to overcome was the problem of measuring regression between variables that are measured on different scales, such as weight and piano playing. This is done by using the population as a standard of reference. Imagine that weight and piano playing have been measured for 100 children in all grades of an elementary school, and that they have been ranked from high to low on each measure. If Jane ranks third in piano playing and twenty-seventh in weight, it is appropriate to say that she is a better pianist than she is tall.

Let us make some assumptions that will simplify things:

At any age,

- Piano-playing success depends only on weekly hours of practice.
- Weight depends only on consumption of ice cream.
- Ice cream consumption and weekly hours of practice are unrelated.

Now, using ranks (or the *standard scores* that statisticians prefer), we can write some equations:

weight = age + ice cream consumption
piano playing = age + weekly hours of practice

You can see that there will be regression to the mean when we predict piano playing from weight, or vice versa. If all you know about Tom is that he ranks twelfth in weight (well above average), you can infer (statistically) that he is probably older than average and also that he probably consumes more ice cream than other children. If all you know about Barbara is that she is eighty-fifth in piano (far below the average of the group), you can infer that she is likely to be young and that she is likely to practice less than most other children.

The *correlation coefficient* between two measures, which varies between 0 and 1, is a measure of the relative weight of the factors they share. For example, we all share half our genes with each of our parents, and for traits in which environmental factors have relatively little influence, such as height, the correlation between parent and child is not far from .50. To appreciate the meaning of the correlation measure, the following are some examples of coefficients:

- The correlation between the size of objects measured with precision in English or in metric units is 1. Any factor that influences one measure also influences the other; 100% of determinants are shared.
- The correlation between self-reported height and weight among adult American males is .41. If you included women and children, the correlation would be much higher, because individuals' gender and age influence both their height ann wd their weight, boosting the

relative weight of shared factors.

- The correlation between SAT scores and college GPA is approximately .60. However, the correlation between aptitude tests and success in graduate school is much lower, largely because measured aptitude varies little in this selected group. If everyone has similar aptitude, differences in this measure are unlikely to play a large role in measures of success.
- The correlation between income and education level in the United States is approximately .40.
- The correlation between family income and the last four digits of their phone number is 0.

It took Francis Galton several years to figure out that correlation and regression are not two concepts—they are different perspectives on the same concept. The general rule is straightforward but has surprising consequences: whenever the correlation between two scores is imperfect, there will be regression to the mean. To illustrate Galton's insight, take a proposition that most people find quite interesting:

> Highly intelligent women tend to marry men who are less intelligent than they are.

You can get a good conversation started at a party by asking for an explanation, and your friends will readily oblige. Even people who have had some exposure to statistics will spontaneously interpret the statement in causal terms. Some may think of highly intelligent women wanting to avoid the competition of equally intelligent men, or being forced to compromise in their choice of spouse because intelligent men do not want to compete with intelligent women. More far-fetched explanations will come up at a good party. Now consider this statement:

> The correlation between the intelligence scores of spouses is less than perfect.

This statement is obviously true and not interesting at all. Who would expect the correlation to be perfect? There is nothing to explain. But the statement you found interesting and the statement you found trivial are algebraically equivalent. If the correlation between the intelligence of spouses is less than perfect (and if men and women on average do not differ in intelligence), then it is a mathematical inevitability that highly intelligent women will be married to husbands who are on average less

intelligent than they are (and vice versa, of course). The observed regression to the mean cannot be more interesting or more explainable than the imperfect correlation.

You probably sympathize with Galton's struggle with the concept of regression. Indeed, the statistician David Freedman used to say that if the topic of regression comes up in a criminal or civil trial, the side that must explain regression to the jury will lose the case. Why is it so hard? The main reason for the difficulty is a recurrent theme of this book: our mind is strongly biased toward causal explanations and does not deal well with "mere statistics." When our attention is called to an event, associative memory will look for its cause—more precisely, activation will automatically spread to any cause that is already stored in memory. Causal explanations will be evoked when regression is detected, but they will be wrong because the truth is that regression to the mean has an explanation but does not have a cause. The event that attracts our attention in the golfing tournament is the frequent deterioration of the performance of the golfers who werecte successful on day 1. The best explanation of it is that those golfers were unusually lucky that day, but this explanation lacks the causal force that our minds prefer. Indeed, we pay people quite well to provide interesting explanations of regression effects. A business commentator who correctly announces that "the business did better this year because it had done poorly last year" is likely to have a short tenure on the air.

Our difficulties with the concept of regression originate with both System 1 and System 2. Without special instruction, and in quite a few cases even after some statistical instruction, the relationship between correlation and regression remains obscure. System 2 finds it difficult to understand and learn. This is due in part to the insistent demand for causal interpretations, which is a feature of System 1.

> Depressed children treated with an energy drink improve significantly over a three-month period.

I made up this newspaper headline, but the fact it reports is true: if you treated a group of depressed children for some time with an energy drink, they would show a clinically significant improvement. It is also the case that depressed children who spend some time standing on their head or hug a cat for twenty minutes a day will also show improvement. Most readers of such headlines will automatically infer that the energy drink or the cat hugging caused an improvement, but this conclusion is completely unjustified. Depressed children are an extreme group, they are more

depressed than most other children—and extreme groups regress to the mean over time. The correlation between depression scores on successive occasions of testing is less than perfect, so there will be regression to the mean: depressed children will get somewhat better over time even if they hug no cats and drink no Red Bull. In order to conclude that an energy drink—or any other treatment—is effective, you must compare a group of patients who receive this treatment to a "control group" that receives no treatment (or, better, receives a placebo). The control group is expected to improve by regression alone, and the aim of the experiment is to determine whether the treated patients improve more than regression can explain.

Incorrect causal interpretations of regression effects are not restricted to readers of the popular press. The statistician Howard Wainer has drawn up a long list of eminent researchers who have made the same mistake—confusing mere correlation with causation. Regression effects are a common source of trouble in research, and experienced scientists develop a healthy fear of the trap of unwarranted causal inference.

One of my favorite examples of the errors of intuitive prediction is adapted from Max Bazerman's excellent text *Judgment in Managerial Decision Making*:

> You are the sales forecaster for a department store chain. All stores are similar in size and merchandise selection, but their sales differ because of location, competition, and random factors. You are given the results for 2011 and asked to forecast sales for 2012. You have been instructed to accept the overall forecast of economists that sales will increase overall by 10%. How would you complete the following table?

Store	2011	2012
1	$11,000,000	_____
2	$23,000,000	_____
3	$18,000,000	_____
4	$29,000,000	_____
Total	$61,000,000	$67,100,000

Having read this chapter, you know that the obvious solution of adding

10% to the sales of each store is wrong. You want your forecasts to be regressive, which requires adding more than 10% to the low-performing branches and adding less (or even subtracting) to others. But if you ask other people, you are likely to encounter puzzlement: Why do you bother them with an obvious question? As Galton painfully discovered, the concept of regression is far from obvious.

Speaking of Regression to Mediocrity

"She says experience has taught her that criticism is more effective than praise. What she doesn't understand is that it's all due to regression to the mean."

"Perhaps his second interview was less impressive than the first because he was afraid of disappointing us, but more likely it was his first that was unusually good."

"Our screening procedure is good but not perfect, so we should anticipate regression. We shouldn't be surprised that the very best candidates often fail to meet our expectations."