Regression to the Mean

Addendum to Lecture #4

Econ 103

January 30, 2015

For more information on Regression to the Mean, please read Chapter 17 of Daniel Kahneman's book *Thinking Fast and Slow*.

In in lecture # 4 we looked at a dataset of the heights of fathers and sons (slide 45) and uncovered a puzzle. Very tall fathers have tall sons but, on average, these sons are not as tall as their fathers. Similarly, very short fathers have short sons but, on average, these sons are not as short as their fathers. We made this precise by looking at z-scores (slides 47-48). We started by identifying all the fathers who were at least two standard deviations below the mean height of all fathers in the dataset. Looking at the sons of these fathers, we found that only two of them were more than two standard deviations below the mean height of all the sons in the dataset. Next we identified all the fathers who were at least two standard deviations above the mean height of all fathers in the dataset. Looking at the sons of these fathers, we found that only three of them were at least two standard deviations above the mean height of all the sons in the dataset.

That last paragraph was a bit complicated so let me repeat in a slightly different way. The sons of very tall sons are closer to the mean height of sons than their fathers are to the mean height of fathers. Likewise, the sons of very short fathers are likewise closer to the mean height of sons than their fathers are to the mean height of fathers. This is called Regression to the Mean. What we have uncovered is a feature of a real-world dataset. Many real-world phenomena involving an approximately linear relationship between two variables exhibit the phenomenon of regression to the mean. What's very interesting is that our regression line from class today is smart enough to know about this. Indeed it takes it into account when predicting.

Take a look at the two lines on slide 49. One is the regression line and the other is the the 45-degree line: $y = x$. Which do you think is which? Take a second to think about this before reading further.

The dashed line is in fact the regression line, while the solid line is the 45-degree line. So what we see is that the regression line is less steep than the 45-degree line. Remember from class today that the regression line always goes through the means of the data. This means that if your father is of average height, relative to the other fathers, the regression line will predict that you are of average height relative to the other sons. But because this line is less steep than $y = x$, it will not make corresponding predictions for the sons of very tall and very short fathers: our line takes the phenomenon of regression to the mean into account and, as a consequence, makes better predictions than the line $y = x$ would. On Problem Set # 3 you will prove that this is a general feature of regression. In particular, you will show the following result:

$$\frac{\hat{y} - \bar{y}}{s_y} = r_{xy} \frac{x - \bar{x}}{s_x}$$

Intuitively, this means that the least-squares regression prediction $\hat{y}$ is closer to $\bar{y}$ than the input x is to $\bar{x}$ whenever the correlation between two variables is less than one in absolute value. (In each case, "close" is measured in the standard deviations of the relevant variables.)

So what's going on here? Why do tall fathers have tall sons, but sons who aren't quite as tall as they are? The key point is that many separate factors combine to determine how tall you will grow up to be. To keep things simple let's put them into two bins: genetic and environmental factors. By genetic factors, I mean anything that is fixed and unchanging at birth and by environmental factors I mean anything that can be influenced by the way in which you are raised. For example, if you don't eat sufficiently nutritious food as a child, you could turn our to be very short in spite of having genes for "tallness." I'm dramatically oversimplifying the biology here, but the point I'm trying to make is that you don't inherit all the factors that influence your height from your parents. Whenever this is the case, there will not be a perfect correlation between your height and your parents' height, so $r_{xy} \neq 1$ in the equation from a few paragraphs above. To be really tall, you need to have genes for being tall and get lucky in terms of environmental factors. If your dad is tall is really tall, he has the genes for being tall and so will you. But in order to be really tall he also got lucky in the environmental factors.

You might get lucky too, but it's rare for both a father and his son to get lucky. That's the basic intuition behind this effect.

One of the main reasons you need to know about regression to the mean is because it's involved in a very common fallacy called the Regression Fallacy. Here's a brief illustration. Suppose I used the math pre-test for Econ 103 to break you into three groups: the top 10%, middle 80%, and bottom 10%. Now suppose that I took the top 10% and put them into a special "enrichment program" designed to give high-flyers the chance to further develop their math skills, and put the bottom 10% into a "remedial program" designed to bring them up to the level of the middle 80% of the class. Four weeks later, suppose I gave each of these two groups another math test. What do you think I would find?

Almost certainly, I would find that the student in the enrichment program did, on average, worse on the second test and that the students in the remedial program did better, on average, on the second test. Does this mean that enrichment programs are bad and remedial programs are good? Amazingly, the answer is no: even if the remedial program and enrichment programs were total shams, we would probably still see this effect. How can this be? The answer is regression to the mean. To get a really high score on a math test, you need to be good at math and lucky: maybe you had a really good night's sleep the night before, felt great when you woke up, and found that all the questions on the test just happened to be things that you knew well. Similarly, to get a really low score on a math, test you need to be bad a math and unlucky: maybe you didn't get much sleep the night before, woke up with a cold, and found that the test just happened to focus on all the areas in which you were the weakest. The point is not that one math test is a bad predictor for another. Quite the contrary! In Econ 103, for example, correlations of 0.6 between midterm one and midterm two are common. However, the correlation isn't perfect because many other factors can influence your score on a test. Students who got very extreme scores relative to everyone else are very likely to be students who had extreme values of these other factors.

Getting back to the students in the remedial class: they aren't good at math, but they also got unlucky on the first test. If I give them as second test after administering a sham remedial program, they'll still be bad at math. However, it's improbabe that they'll get lucky twice in a row. For this reason, their scores will, on average, be closer to the mean on the second exam. The same story works in reverse for the students in the enrichment

class. They're good at math, but they also got lucky on the first test. After the sham enrichment program, they'll still be just as good a math as before, but it's unlikely that they'll all get lucky two times in a row. On average, their scores will be closer to the mean on the second exam. The fallacy comes from wrongly concluding that there's a causal mechanism at work when all that's really happening is regression to the mean. This kind of faulty reasoning crops up all over the place, for example when people talk about the supposed "sophomore slump" in sports. Why is it that players who have especially strong first years in professional sports nearly always do *worse* the second year? Is it because they become complacent and "rest on their laurels?" This could be the case, but even if it weren't we'd *still* expect to see such a pattern. Those who had particularly strong first years were good a sports and lucky. Next year they'll still be good at sports, but very few of them will get lucky two years in a row.

Numerical Example

Imagine a test where your grade depends on effort and luck according to the following relationship

$$\text{grade} = 40 + 50 \times \text{effort} + \text{luck}$$

where effort is the fraction of effort that you exerted while studying, a number between zero and one, and luck is random: it is equally likely to take on any integer value from -10 to 10, including zero.

Now suppose I give this test to a large number of students and look at *only those* who scored 98 or higher. What effort did these students expend? While we can't say for certain, we know that it was at least 0.96. This is becuase the highest possible value for luck is 10, so the highest possible "baseline grade" is $40 + 10 = 50$. To end up with 98 on top of this baseline, you need $50 \times$ effort to be at least 48, which corresponds to effort $= 0.96$. Thus, if a student scores 98 or higher, her effort level must have been between 0.96 and 1. What about her value of "luck?" If she exerted full effort, then her luck value could have been as low as 8; if she only exerted 0.96 effort, then her luck must have been 10. In summary, someone who scored 98 or above exerted effort between 0.96 and 1, and got a luck value between 8 and 10.

Now suppose that we give a second test to the students who scored 98 or above without giving them another chance to study. Than whatever effort value they had on the first test carries over to the second. Luck, on the other hand, is random and shows no dependence. Maybe you'll wake up sick on the day of the exam and luck will be -10, or maybe I won't ask any questions about regression, which you didn't study, so you luck is 10. The question is, what fraction of the students who got 98 or higher on the first test will get 98 or higher on the second?

Since we don't know their exact effort levels, we can't answer this question exactly, but we can product an *upper bound*. The best case, that is the one that makes it most likely that they'll score well on the second exam, is when effort is 1. Suppose this is the case. Then what is the probability of getting 98 or higher? Since this corresponds to a luck value of 8, 9 or 10 and each possible value for luck is equally likely, the probability is $3/21 = 1/7$ which is just under 0.15. Thus, among those students who got 98 or higher on the first exam, *fewer than 15%* will do at least as well on the second exam. The rest will do worse.