Regression To The Mean

Blithering Genius

2018 January 21

Contents

There is a lot of confusion over the concept of regression to the mean, so I thought I would try to explain what it is, and what it isn't.

1 The Dice Rolling Example

Let's start with a very simple example. Suppose you have 10 6-sided dice. Assume that these dice are all fair, so that the expected value on any given roll is 3.5. You roll each die once. Then you select the 5 dice that rolled the highest and set them aside. Probably the average of the top 5 dice is higher than 3.5, simply because you selected the highest rollers. Now roll those 5 dice a second time. Probably the average of the second roll is lower than the first: something closer to 3.5. That is regression to the mean. If you select a sample based on the measurement of a random variable X, the value of X within the sample is a biased estimator of X. Future measurements of X will tend to "regress" to the mean of X. No physical process is involved in this. It is simply the removal of sample bias by a second measurement or experiment.

2 The Traffic Fatalities At Intersections Example

Now let's consider a more complex example that involves both random and non-random factors. Suppose a city government wants to reduce traffic fatalities. It decides to try installing red light cameras on the most dangerous intersections. So, the traffic engineers identify the 10 most dangerous intersections in the city, based on traffic accidents in the previous year. They install red light

cameras at those intersections. After this intervention, the number of traffic accidents at those 10 intersections declines by 20%. Did the red light cameras reduce the number of traffic accidents?

Before going on, let's eliminate one possible problem with the experiment. Let's assume that the overall traffic patterns of the city did not change from one year to the next. The total amount of traffic and the routes taken did not change. Given that assumption, can we conclude that the red light cameras reduced the frequency of accidents?

Not necessarily, because of regression to the mean.

3 Explaining What's Going On In The Traffic Fatalities Data

The number of accidents in the previous year is not an unbiased estimator of the number of accidents in the current year for the intersections in the experiment, because the intersections were chosen based on the number of accidents in the previous year. There is sample bias. The number of accidents in a year is not precisely determined by the properties of an intersection. It is somewhat random. Without any intervention, we should expect the number of accidents in the current year to be somewhat lower for the intersections in the sample, just because of regression to the mean. The second measurement removes the sample bias, because the second measurement was not used to select the sample.

To properly assess the effect of the red light cameras, the experiment would have to be different. We could select the 20 intersections with the most accidents and install red-light cameras at 10 of them, selected at random. Then we could compare the number of accidents next year between the 10 with cameras and the 10 without cameras. Or we could install red-light cameras at a random subset of intersections, regardless of whether they have a lot of accidents, and then see if they have fewer accidents the next year. Both methods would eliminate the problem of sample bias.

Hopefully those two examples illustrate what regression to the mean is and what it isn't. It is the removal of sample bias by a second measurement or experiment. It is not a physical mechanism. It is just a statistical effect.

4 How Regression To The Mean Applies To Hereditary Traits

Now let's consider how regression to the mean applies to the heritability of traits such as IQ. For parents chosen at random, the average IQ of the parents is a good estimator of the IQ of their children. I don't know for sure that the expectation of child IQ is precisely the average of the parent's IQ, but let's assume that it is for the sake of this example. For all parents, if you measure their IQs and take the average, that predicts the IQ of their children.

However, if you look at high IQ parents you will discover that, on average, their children have IQs lower than the parental average. Conversely, if you look at low IQ parents, you discover that, on average, their children have IQs higher than the parental average. The same applies to other traits as well, such as height. What is going on?

Regression to the mean.

It is not that some force or biological process pulls the children toward the population average. It is that most traits, including IQ, are determined by multiple factors, some of which are random from our perspective (meaning we can't predict them). There are non-additive effects of genes (due to

interaction between alleles of the same gene or different genes), so even the effect of genes on a trait is not necessarily estimated by the average of parental traits (only additive effects are correctly estimated by parental averages). Traits are also to some extent caused by other factors that are essentially random, such as developmental errors. When we select people based on a trait such as IQ, we are not only selecting for the underlying additive genetic factors that affect the trait, but also for other factors that are essentially random. For a randomly selected person, their measured IQ is an unbiased estimate of the additive genetic component (non-random component) of IQ, assuming that positive and negative deviations from it are equally likely. But, if we select people based on IQ, then we are selecting for the random factors as well, and those random factors are not heritable (or not fully heritable). Thus, parental average IQ is a biased estimator of child IQ when parents are selected based on IQ, and the measurement of child IQ removes that sample bias.

Again, no causal mechanism was involved. The IQ of the children did not regress to the mean because of some physical process that pulled it toward the statistical average of the population. It's just that the parental average IQ was a biased estimate of the child's IQ.

5 Explaining Admixture To The Mean

There is a process that might be confused with regression to the mean, but which isn't regression to the mean. That is the tendency of descendants of exceptional individuals to tend toward the mean over multiple generations. In the first generation, this could be due to regression to the mean, but over multiple generations it is due to mixing. Exceptional individuals almost always mate with less exceptional individuals (because less exceptional individuals, by definition, are more common). Thus, the children of great geniuses tend to be less intelligent not only because of regression to the mean, but also because great geniuses tend to mate with less intelligent people. Over multiple generations, the descendants of exceptional individuals blend into the rest of the population. This is not regression to the mean. It is mixing.

6 Mixing In Genetics

There is another mixing process that could be relevant to questions of race, race-mixing and human genetics in general. There could be co-adapted alleles of different genes that occupy the same region on a chromosome, that tend to be passed together to offspring as a package. Recombination could break up those packages, not only in the first generation, but in subsequent generations.

Let me explain.

DNA is packaged into chromosomes. You inherit one chromosome from each parent, but the chromosome you inherit is not a copy of one of your parent's chromosomes. Instead, it is a combination of both of your parent's copies of that chromosome. During meiosis, crossover mixes the DNA of the parent's chromosomes. In a sense, sex takes place not only between organisms, but also between paired chromosomes. Chromosomes "mate" by mixing their DNA to produce a new chromosome. As an aside, I believe that recombination is a mechanism to prevent chromosomal competition. If chromosomes were replicated as units, then there would be the potential for competition between the two copies of a chromosome. In fact, there can be chromosomal competition between sex chromosomes, which do not recombine. (See here for example.)

Recombination would eventually break up these genetic packages, recombining them with the genes on a different strand. It could be that mixing races, or mixing genetically distant subpopulations (whether we consider them to be races or not) could break up co-adapted gene complexes not only in the first generation, but in subsequent generations. Crossover could cause a kind of multi-generational "regression to the mean" that is not true regression to the mean. If your chromosomes are different, then combining material from them to produce a new chromosome could break up gene complexes.

I raise this issue because it is an interesting possibility to consider, but I don't think it is something to worry about in humans. It is theoretically possible that "out-breeding" (mating with a more genetically distant person) could cause problems by breaking up co-adapted gene complexes. On the other hand, in-breeding can also cause problems, such as genetic diseases caused by recessive genes.

Nature doesn't have any rules about who breeds with whom. Organisms mate, produce offspring, and then selection operates on them. Gene flow is not intrinsically harmful, and it is necessary to bring together beneficial mutations that occur in different populations. Also, the repeated mixing of DNA forces genes to work well with other genes, and makes the genome more "modular". The constant mixing of DNA by sex might be really important in creating complex life, not only because it allows beneficial mutations to become associated, but also because it prevents DNA from becoming ad hoc, tightly coupled "spaghetti code". Just as reuse in different contexts forces programmers to create cleaner code, the reuse of genetic code in different contexts forces evolution to create cleaner genetic code.

7 What Regression To The Mean Is Not

Some people, especially race idealists, seem to believe that regression to the mean implies that selection is impossible or futile. They believe that if people of another race are selected based on IQ or some other trait, such as economic success, their descendants will regress back to their racial mean, and thus in the long run selective immigration is just as dysgenic as open immigration. This view is false. If everyone's descendants regressed back to an ancestral mean, then evolution would be impossible because selection would have no long-term effects. Any divergence caused by selection would simply be erased by regression. If that were true, different races could never have come into existence in the first place, because races were created by evolutionary divergence from some ancestral population. In fact, life itself could not exist.

Hopefully this has helped to clarify regression to the mean. Regression to the mean does not imply that selection is always reversed. There is no causal mechanism that reverses selection and pulls descendants back to an ancestral mean. Regression to the mean is a statistical effect due to the removal of sample bias.